CONCEPTUAL PHYSICS

Benjamin Crowell
Fullerton College

Fullerton College

Conceptual Physics

Benjamin Crowell

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

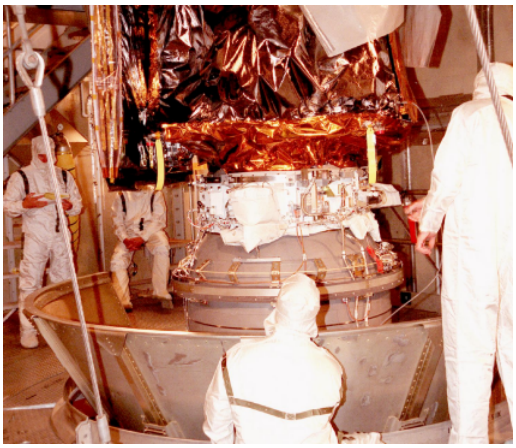
TABLE OF CONTENTS

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

CHAPTER OVERVIEW

1: Introduction and Review



The Mars Climate Orbiter is prepared for its mission. The laws of physics are the same everywhere, even on Mars, so the probe could be designed based on the laws of physics as discovered on earth. There is unfortunately another reason why this spacecraft is relevant to the topics of this chapter: it was destroyed attempting to enter Mars' atmosphere because engineers at Lockheed Martin forgot to convert data on engine thrusts from pounds into the metric unit of force (newtons) before giving the information to NASA. Conversions are important!

Contributors and Attributions

- Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

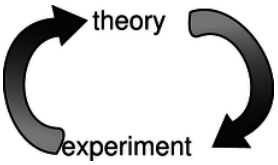
1.1: Introduction and Review

If you drop your shoe and a coin side by side, they hit the ground at the same time. Why doesn't the shoe get there first, since gravity is pulling harder on it? How does the lens of your eye work, and why do your eye's muscles need to squash its lens into different shapes in order to focus on objects nearby or far away? These are the kinds of questions that physics tries to answer about the behavior of light and matter, the two things that the universe is made of.

0.1.1 The scientific method

Until very recently in history, no progress was made in answering questions like these. Worse than that, the *wrong* answers written by thinkers like the ancient Greek physicist Aristotle were accepted without question for thousands of years. Why is it that scientific knowledge has progressed more since the Renaissance than it had in all the preceding millennia since the beginning of recorded history? Undoubtedly the industrial revolution is part of the answer. Building its centerpiece, the steam engine, required improved techniques for precise construction and measurement. (Early on, it was considered a major advance when English machine shops learned to build pistons and cylinders that fit together with a gap narrower than the thickness of a penny.) But even before the industrial revolution, the pace of discovery had picked up, mainly because of the introduction of the modern scientific method. Although it evolved over time, most scientists today would agree on something like the following list of the basic principles of the scientific method:

(1) *Science is a cycle of theory and experiment.* Scientific theories are created to explain the results of experiments that were created under certain conditions. A successful theory will also make new predictions about new experiments under new conditions. Eventually, though, it always seems to happen that a new experiment comes along, showing that under certain conditions the theory is not a good approximation or is not valid at all. The ball is then back in the theorists' court. If an experiment disagrees with the current theory, the theory has to be changed, not the experiment.



a / Science is a cycle of theory and experiment.

b / A satirical drawing of an alchemist's laboratory. H. Cock, after a drawing by Peter Brueghel the Elder (16th century).

(2) *Theories should both predict and explain.* The requirement of predictive power means that a theory is only meaningful if it predicts something that can be checked against experimental measurements that the theorist did not already have at hand. That is, a theory should be testable. Explanatory value means that many phenomena should be accounted for with few basic principles. If you answer every “why” question with “because that's the way it is,” then your theory has no explanatory value. Collecting lots of data without being able to find any basic underlying principles is not science.

(3) *Experiments should be reproducible.* An experiment should be treated with suspicion if it only works for one person, or only in one part of the world. Anyone with the necessary skills and equipment should be able to get the same results from the same experiment. This implies that science transcends national and ethnic boundaries; you can be sure that nobody is doing actual science who claims that their work is “Aryan, not Jewish,” “Marxist, not bourgeois,” or “Christian, not atheistic.” An experiment cannot be reproduced if it is secret, so science is necessarily a public enterprise.

As an example of the cycle of theory and experiment, a vital step toward modern chemistry was the experimental observation that the chemical elements could not be transformed into each other, e.g., lead could not be turned into gold. This led to the theory that chemical reactions consisted of rearrangements of the elements in different combinations, without any change in the identities of the elements themselves. The theory worked for hundreds of years, and was confirmed experimentally over a wide range of

pressures and temperatures and with many combinations of elements. Only in the twentieth century did we learn that one element could be trans-formed into one another under the conditions of extremely high pressure and temperature existing in a nuclear bomb or inside a star. That observation didn't completely invalidate the original theory of the immutability of the elements, but it showed that it was only an approximation, valid at ordinary temperatures and pressures.

self-check:

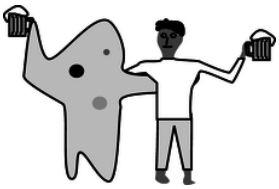
A psychic conducts seances in which the spirits of the dead speak to the participants. He says he has special psychic powers not possessed by other people, which allow him to “channel” the communications with the spirits. What part of the scientific method is being violated here?

(answer in the back of the PDF version of the book)

The scientific method as described here is an idealization, and should not be understood as a set procedure for doing science. Scientists have as many weaknesses and character flaws as any other group, and it is very common for scientists to try to discredit other people's experiments when the results run contrary to their own favored point of view. Successful science also has more to do with luck, intuition, and creativity than most people realize, and the restrictions of the scientific method do not stifle individuality and self-expression any more than the fugue and sonata forms stifled Bach and Haydn. There is a recent tendency among social scientists to go even further and to deny that the scientific method even exists, claiming that science is no more than an arbitrary social system that determines what ideas to accept based on an in-group's criteria. I think that's going too far. If science is an arbitrary social ritual, it would seem difficult to explain its effectiveness in building such useful items as airplanes, CD players, and sewers. If alchemy and astrology were no less scientific in their methods than chemistry and astronomy, what was it that kept them from producing anything useful?

Contributors and Attributions

1.2: Scaling and Order-of-Magnitude Estimates



a / Amoebas this size are seldom encountered.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

1.3: Footnotes

1. Liberia and Myanmar have not legally adopted metric units, but use them in everyday life.
2. Galileo makes a slightly more complicated argument, taking into account the effect of leverage (torque). The result I'm referring to comes out the same regardless of this effect.
3. The customary terms “half-size” and “3/4-size” actually don't describe the sizes in any accurate way. They're really just standard, arbitrary marketing labels.

1.4: Problems

1. Correct use of a calculator: (a) Calculate $\frac{74658}{53222+97554}$ on a calculator. [Self-check: The most common mistake results in 97555.40.] (answer check available at lightandmatter.com)

(b) Which would be more like the price of a TV, and which would be more like the price of a house, $\$3.5 \times 10^5$ *or* $\$3.5^5$*?*

2. Compute the following things. If they don't make sense because of units, say so.
(a) 3 cm + 5 cm
(b) 1.11 m + 22 cm
(c) 120 miles + 2.0 hours
(d) 120 miles / 2.0 hours

3. Your backyard has brick walls on both ends. You measure a distance of 23.4 m from the inside of one wall to the inside of the other. Each wall is 29.4 cm thick. How far is it from the outside of one wall to the outside of the other? Pay attention to significant figures.

4. The speed of light is 3.0×10^8 m/s. Convert this to furlongs per fortnight. A furlong is 220 yards, and a fortnight is 14 days. An inch is 2.54 cm.(answer check available at lightandmatter.com)

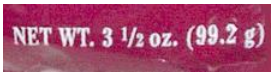
5. Express each of the following quantities in micrograms:
(a) 10 mg, (b) 10^4 g, (c) 10 kg, (d) 100×10^3 g, (e) 1000 ng. (answer check available at lightandmatter.com)

6. (solution in the pdf version of the book) Convert 134 mg to units of kg, writing your answer in scientific notation.

7. In the last century, the average age of the onset of puberty for girls has decreased by several years. Urban folklore has it that this is because of hormones fed to beef cattle, but it is more likely to be because modern girls have more body fat on the average and possibly because of estrogen-mimicking chemicals in the environment from the breakdown of pesticides. A hamburger from a hormone-implanted steer has about 0.2 ng of estrogen (about double the amount of natural beef). A serving of peas contains about 300 ng of estrogen. An adult woman produces about 0.5 mg of estrogen per day (note the different unit!). (a) How many hamburgers would a girl have to eat in one day to consume as much estrogen as an adult woman's daily production? (b) How many servings of peas? (answer check available at lightandmatter.com)

8. (solution in the pdf version of the book) The usual definition of the mean (average) of two numbers a and b is $(a+b)/2$. This is called the arithmetic mean. The geometric mean, however, is defined as $(ab)^{1/2}$ (i.e., the square root of ab). For the sake of definiteness, let's say both numbers have units of mass. (a) Compute the arithmetic mean of two numbers that have units of grams. Then convert the numbers to units of kilograms and recompute their mean. Is the answer consistent? (b) Do the same for the geometric mean. (c) If a and b both have units of grams, what should we call the units of ab? Does your answer make sense when you take the square root? (d) Suppose someone proposes to you a third kind of mean, called the superduper mean, defined as $(ab)^{1/3}$. Is this reasonable?

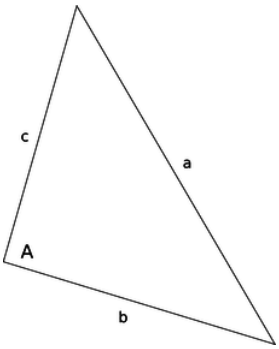
9. In an article on the SARS epidemic, the May 7, 2003 New York Times discusses conflicting estimates of the disease's incubation period (the average time that elapses from infection to the first symptoms). “The study estimated it to be 6.4 days. But other statistical calculations ... showed that the incubation period could be as long as 14.22 days.” What's wrong here?



a / Problem 10.

10. The photo shows the corner of a bag of pretzels. What's wrong here?

11. The distance to the horizon is given by the expression $\sqrt{2rh}$, where r is the radius of the Earth, and h is the observer's height above the Earth's surface. (This can be proved using the Pythagorean theorem.) Show that the units of this expression make sense. (See example 2 on p. 26 for an example of how to do this.) Don't try to prove the result, just check its units.



b / Problem 12

12. (solution in the pdf version of the book) (a) Based on the definitions of the sine, cosine, and tangent, what units must they have? (b) A cute formula from trigonometry lets you find any angle of a triangle if you know the lengths of its sides. Using the notation shown in the figure, and letting $s = (a+b+c)/2$ be half the perimeter, we have

$$\tan A/2 = \sqrt{\frac{(s-b)(s-c)}{s(s-a)}}.$$

Show that the units of this equation make sense. In other words, check that the units of the right-hand side are the same as your answer to part a of the question.

13. A physics homework question asks, "If you start from rest and accelerate at 1.54 $\mathrm{m/s^2}$ for 3.29 s, how far do you travel by the end of that time?" A student answers as follows:

$$1.54 \times 3.29 = 5.07 \text{ m}$$

His Aunt Wanda is good with numbers, but has never taken physics. She doesn't know the formula for the distance traveled under constant acceleration over a given amount of time, but she tells her nephew his answer cannot be right. How does she know?

14. You are looking into a deep well. It is dark, and you cannot see the bottom. You want to find out how deep it is, so you drop a rock in, and you hear a splash 3.0 seconds later. How deep is the well? (answer check available at lightandmatter.com)

15. You take a trip in your spaceship to another star. Setting off, you increase your speed at a constant acceleration. Once you get half-way there, you start decelerating, at the same rate, so that by the time you get there, you have slowed down to zero speed. You see the tourist attractions, and then head home by the same method.
(a) Find a formula for the time, T, required for the round trip, in terms of d, the distance from our sun to the star, and a, the magnitude of the acceleration. Note that the acceleration is not constant over the whole trip, but the trip can be broken up into constant-acceleration parts.
(b) The nearest star to the Earth (other than our own sun) is Proxima Centauri, at a distance of $d = 4 \times 10^{16}$ m . Suppose you use an acceleration of $a = 10 \text{ m/s}^2$, just enough to compensate for the lack of true gravity and make you feel comfortable. How long does the round trip take, in years?
(c) Using the same numbers for d and a, find your maximum speed. Compare this to the speed of light, which is 3.0×10^8 \ \textup{m}/\textup{s}. (Later in this course, you will learn that there are some new things going on in physics when one gets close to the speed of light, and that it is impossible to exceed the speed of light. For now, though, just use the simpler ideas you've learned so far.) (answer check available at lightandmatter.com)

16. You climb half-way up a tree, and drop a rock. Then you climb to the top, and drop another rock. How many times greater is the velocity of the second rock on impact? Explain. (The answer is not two times greater.)

17. (solution in the pdf version of the book) If the acceleration of gravity on Mars is 1/3 that on Earth, how many times longer does it take for a rock to drop the same distance on Mars? Ignore air resistance.

18. A person is parachute jumping. During the time between when she leaps out of the plane and when she opens her chute, her altitude is given by an equation of the form

$$y = b - c\left(t + ke^{-t/k}\right),$$

where e is the base of natural logarithms, and b, c, and k are constants. Because of air resistance, her velocity does not increase at a steady rate as it would for an object falling in vacuum.
(a) What units would b, c, and k have to have for the equation to make sense?
(b) Find the person's velocity, v, as a function of time. [You will need to use the chain rule, and the fact that $d(e^x)/dx = e^x$.] (answer check available at lightandmatter.com)
(c) Use your answer from part (b) to get an interpretation of the constant c. [Hint: e^{-x} approaches zero for large values of x.]
(d) Find the person's acceleration, a, as a function of time.(answer check available at lightandmatter.com)
(e) Use your answer from part (d) to show that if she waits long enough to open her chute, her acceleration will become very small.

19. (solution in the pdf version of the book) In July 1999, Popular Mechanics carried out tests to find which car sold by a major auto maker could cover a quarter mile (402 meters) in the shortest time, starting from rest. Because the distance is so short, this type of test is designed mainly to favor the car with the greatest acceleration, not the greatest maximum speed (which is irrelevant to the average person). The winner was the Dodge Viper, with a time of 12.08 s. The car's top (and presumably final) speed was 118.51 miles per hour (52.98 \ \textup{m}/\textup{s}). (a) If a car, starting from rest and moving with *constant* acceleration, covers a quarter mile in this time interval, what is its acceleration? (b) What would be the final speed of a car that covered a quarter mile with the constant acceleration you found in part a? (c) Based on the discrepancy between your answer in part b and the actual final speed of the Viper, what do you conclude about how its acceleration changed over time?

20. The speed required for a low-earth orbit is 7.9×10^3 m/s (see ch. 10). When a rocket is launched into orbit, it goes up a little at first to get above almost all of the atmosphere, but then tips over horizontally to build up to orbital speed. Suppose the horizontal acceleration is limited to $3g$ to keep from damaging the cargo (or hurting the crew, for a crewed flight). (a) What is the minimum distance the rocket must travel downrange before it reaches orbital speed? How much does it matter whether you take into account the initial eastward velocity due to the rotation of the earth? (b) Rather than a rocket ship, it might be advantageous to use a railgun design, in which the craft would be accelerated to orbital speeds along a railroad track. This has the advantage that it isn't necessary to lift a large mass of fuel, since the energy source is external. Based on your answer to part a, comment on the feasibility of this design for crewed launches from the earth's surface.

21. Consider the following passage from Alice in Wonderland, in which Alice has been falling for a long time down a rabbit hole:

Down, down, down. Would the fall *never* come to an end? “I wonder how many miles I've fallen by this time?” she said aloud. “I must be getting somewhere near the center of the earth. Let me see: that would be four thousand miles down, I think” (for, you see, Alice had learned several things of this sort in her lessons in the schoolroom, and though this was not a *very* good opportunity for showing off her knowledge, as there was no one to listen to her, still it was good practice to say it over)...

Alice doesn't know much physics, but let's try to calculate the amount of time it would take to fall four thousand miles, starting from rest with an acceleration of 10 $\mathrm{m/s^2}$. This is really only a lower limit; if there really was a hole that deep, the fall would actually take a longer time than the one you calculate, both because there is air friction and because gravity gets weaker as you get deeper (at the center of the earth, g is zero, because the earth is pulling you equally in every direction at once). (answer check available at lightandmatter.com)

22. How many cubic inches are there in a cubic foot? The answer is not 12.(answer check available at lightandmatter.com)

23. Assume a dog's brain is twice as great in diameter as a cat's, but each animal's brain cells are the same size and their brains are the same shape. In addition to being a far better companion and much nicer to come home to, how many times more brain cells does a dog have than a cat? The answer is not 2.

24. The population density of Los Angeles is about 4000 people/$\mathrm{km^2}$. That of San Francisco is about 6000 people/$\mathrm{km^2}$. How many times farther away is the average person's nearest neighbor in LA than in San Francisco? The answer is not 1.5. (answer check available at lightandmatter.com)

25. A hunting dog's nose has about 10 square inches of active surface. How is this possible, since the dog's nose is only about 1 in $\times$ 1 in $\times$ 1 in = 1 $\mathrm{in^3}$? After all, 10 is greater than 1, so how can it fit?

26. Estimate the number of blades of grass on a football field.

27. In a computer memory chip, each bit of information (a 0 or a 1) is stored in a single tiny circuit etched onto the surface of a silicon chip. The circuits cover the surface of the chip like lots in a housing development. A typical chip stores 64 Mb (megabytes) of data, where a byte is 8 bits. Estimate (a) the area of each circuit, and (b) its linear size.

28. Suppose someone built a gigantic apartment building, measuring 10 km $\times$ 10 km at the base. Estimate how tall the building would have to be to have space in it for the entire world's population to live.

29. A hamburger chain advertises that it has sold 10 billion Bongo Burgers. Estimate the total mass of feed required to raise the cows used to make the burgers.

30. Estimate the volume of a human body, in cm^3.

31. (solution in the pdf version of the book) How many cm^2 is 1 mm^2?

32. (solution in the pdf version of the book) Compare the light-gathering powers of a 3-cm-diameter telescope and a 30-cm telescope.

33. (solution in the pdf version of the book) One step on the Richter scale corresponds to a factor of 100 in terms of the energy absorbed by something on the surface of the Earth, e.g., a house. For instance, a 9.3-magnitude quake would release 100 times more energy than an 8.3. The energy spreads out from the epicenter as a wave, and for the sake of this problem we'll assume we're dealing with seismic waves that spread out in three dimensions, so that we can visualize them as hemispheres spreading out under the surface of the earth. If a certain 7.6-magnitude earthquake and a certain 5.6-magnitude earthquake produce the same amount of vibration where I live, compare the distances from my house to the two epicenters.

34. In Europe, a piece of paper of the standard size, called A4, is a little narrower and taller than its American counterpart. The ratio of the height to the width is the square root of 2, and this has some useful properties. For instance, if you cut an A4 sheet from left to right, you get two smaller sheets that have the same proportions. You can even buy sheets of this smaller size, and they're called A5. There is a whole series of sizes related in this way, all with the same proportions. (a) Compare an A5 sheet to an A4 in terms of area and linear size. (b) The series of paper sizes starts from an A0 sheet, which has an area of one square meter. Suppose we had a series of boxes defined in a similar way: the B0 box has a volume of one cubic meter, two B1 boxes fit exactly inside an B0 box, and so on. What would be the dimensions of a B0 box? (answer check available at lightandmatter.com)

c / Albert Einstein, and his moustache, problem 35.

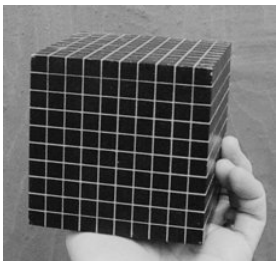
35. Estimate the mass of one of the hairs in Albert Einstein's moustache, in units of kg.

36. According to folklore, every time you take a breath, you are inhaling some of the atoms exhaled in Caesar's last words. Is this true? If so, how many?

37. The Earth's surface is about 70% water. Mars's diameter is about half the Earth's, but it has no surface water. Compare the land areas of the two planets.(answer check available at lightandmatter.com)

38. (solution in the pdf version of the book) The traditional Martini glass is shaped like a cone with the point at the bottom. Suppose you make a Martini by pouring vermouth into the glass to a depth of 3 cm, and then adding gin to bring the depth to 6 cm. What are the proportions of gin and vermouth?

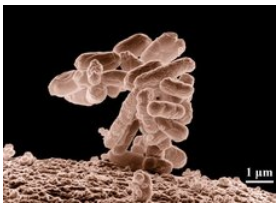
39. The central portion of a CD is taken up by the hole and some surrounding clear plastic, and this area is unavailable for storing data. The radius of the central circle is about 35% of the outer radius of the data-storing area. What percentage of the CD's area is therefore lost? (answer check available at lightandmatter.com)



d / Problem 40.

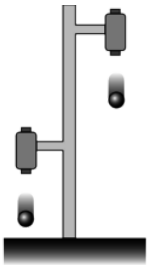
40. The one-liter cube in the photo has been marked off into smaller cubes, with linear dimensions one tenth those of the big one. What is the volume of each of the small cubes?(solution in the pdf version of the book)

41. Estimate the number of man-hours required for building the Great Wall of China. (solution in the pdf version of the book)



e / Problem 42.

42. (a) Using the microscope photo in the figure, estimate the mass of a one cell of the *E. coli* bacterium, which is one of the most common ones in the human intestine. Note the scale at the lower right corner, which is 1 μm. Each of the tubular objects in the column is one cell. (b) The feces in the human intestine are mostly bacteria (some dead, some alive), of which *E. coli* is a large and typical component. Estimate the number of bacteria in your intestines, and compare with the number of human cells in your body, which is believed to be roughly on the order of 10^{13}. (c) Interpreting your result from part b, what does this tell you about the size of a typical human cell compared to the size of a typical bacterial cell?



f / Problem 43.

43. The figure shows a practical, simple experiment for determining g to high precision. Two steel balls are suspended from electromagnets, and are released simultaneously when the electric current is shut off. They fall through unequal heights Δx_1 and Δx_2. A computer records the sounds through a microphone as first one ball and then the other strikes the floor. From this recording, we can accurately determine the quantity T defined as $T = \Delta t_2 - \Delta t_1$, i.e., the time lag between the first and second impacts. Note that since the balls do not make any sound when they are released, we have no way of measuring the individual times Δt_2 and Δt_1.

(a) Find an equation for g in terms of the measured quantities T, Δx_1 and Δx_2.(answer check available at lightandmatter.com)
(b) Check the units of your equation.
(c) Check that your equation gives the correct result in the case where Δx_1 is very close to zero. However, is this case realistic?
(d) What happens when $\Delta x_1 = \Delta x_2$? Discuss this both mathematically and physically.

44. (solution in the pdf version of the book) Estimate the number of jellybeans in figure o on p. 44.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

CHAPTER OVERVIEW

2: Conservation of Mass

It took just a moment for that head to fall, but a hundred years might not produce another like it. -- *Joseph-Louis Lagrange, referring to the execution of Lavoisier on May 8, 1794*

Topic hierarchy

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

2.1: Mass

Change is impossible, claimed the ancient Greek philosopher Parmenides. His work was nonscientific, since he didn't state his ideas in a form that would allow them to be tested experimentally, but modern science nevertheless has a strong Parmenidean flavor. His main argument that change is an illusion was that something can't be turned into nothing, and likewise if you have nothing, you can't turn it into something. To make this into a scientific theory, we have to decide on a way to measure what "something" is, and we can then check by measurements whether the total amount of "something" in the universe really stays constant. How much "something" is there in a rock? Does a sunbeam count as "something?" Does heat count? Motion? Thoughts and feelings.

If you look at the table of contents of this book, you'll see that the first four chapters have the word "conservation" in them. In physics, a conservation law is a statement that the total amount of a certain physical quantity always stays the same. This chapter is about conservation of mass. The metric system is designed around a unit of distance, the meter, a unit of mass, the kilogram, and a time unit, the second. Numerical measurement of distance and time probably date back almost as far into prehistory as counting money, but mass is a more modern concept. Until scientists figured out that mass was conserved, it wasn't obvious that there could be a single, consistent way of measuring an amount of matter, hence jiggers of whiskey and cords of wood. You may wonder why conservation of mass wasn't discovered until relatively modern times, but it wasn't obvious, for example, that gases had mass, and that the apparent loss of mass when wood was burned was exactly matched by the mass of the escaping gases.

Once scientists were on the track of the conservation of mass concept, they began looking for a way to define mass in terms of a definite measuring procedure. If they tried such a procedure, and the result was that it led to nonconservation of mass, then they would throw it out and try a different procedure. For instance, we might be tempted to define mass using kitchen measuring cups, i.e., as a measure of volume. Mass would then be perfectly conserved for a process like mixing marbles with peanut butter, but there would be processes like freezing water that led to a net increase in mass, and others like soaking up water with a sponge that caused a decrease. If, with the benefit of hindsight, it seems like the measuring cup definition was just plain silly, then here's a more subtle example of a wrong definition of mass. Suppose we define it using a bathroom scale, or a more precise device such as a postal scale that works on the same principle of using gravity to compress or twist a spring. The trouble is that gravity is not equally strong all over the surface of the earth, so for instance there would be nonconservation of mass when you brought an object up to the top of a mountain, where gravity is a little weaker.

Although some of the obvious possibilities have problems, there do turn out to be at least two approaches to defining mass that lead to its being a conserved quantity, so we consider these definitions to be "right" in the pragmatic sense that what's correct is what's useful.

One definition that works is to use balances, but compensate for the local strength of gravity. This is the method that is used by scientists who actually specialize in ultraprecise measurements. A standard kilogram, in the form of a platinum-iridium cylinder, is kept in a special shrine in Paris. Copies are made that balance against the standard kilogram in Parisian gravity, and they are then transported to laboratories in other parts of the world, where they are compared with other masses in the local gravity. The quantity defined in this way is called *gravitational mass*.

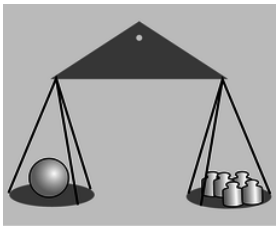


Figure b: A measurement of gravitational mass: the sphere has a gravitational mass of five kilograms.

A second and completely different approach is to measure how hard it is to change an object's state of motion. This tells us its *inertial mass*. For example, I'd be more willing to stand in the way of an oncoming poodle than in the path of a freight train, because my body will have a harder time convincing the freight train to stop. This is a dictionary-style conceptual definition, but in physics we need to back up a conceptual definition with an operational definition, which is one that spells out the operations required in order to measure the quantity being defined. We can operationalize our definition of inertial mass by throwing a standard kilogram at an object at a speed of 1 m/s (one meter per second) and measuring the recoiling object's velocity. Suppose we want to measure the mass of a particular block of cement. We put the block in a toy wagon on the sidewalk, and throw a standard

kilogram at it. Suppose the standard kilogram hits the wagon, and then drops straight down to the sidewalk, having lost all its velocity, and the wagon and the block inside recoil at a velocity of 0.23 m/s. We then repeat the experiment with the block replaced by various numbers of standard kilograms, and find that we can reproduce the recoil velocity of 0.23 m/s with four standard kilograms in the wagon. We have determined the mass of the block to be four kilograms.[1] Although this definition of inertial mass has an appealing conceptual simplicity, it is obviously not very practical, at least in this crude form. Nevertheless, this method of collision is very much like the methods used for measuring the masses of subatomic particles, which, after all, can't be put on little postal scales!

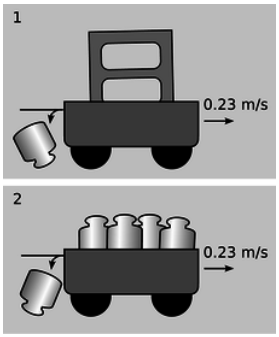


Figure c: A measurement of inertial mass: the wagon recoils with the same velocity in experiments 1 and 2, establishing that the inertial mass of the cement block is four kilograms.

Astronauts spending long periods of time in space need to monitor their loss of bone and muscle mass, and here as well, it's impossible to measure gravitational mass. Since they don't want to have standard kilograms thrown at them, they use a slightly different technique (figures d and e). They strap themselves to a chair which is attached to a large spring, and measure the time it takes for one cycle of vibration.

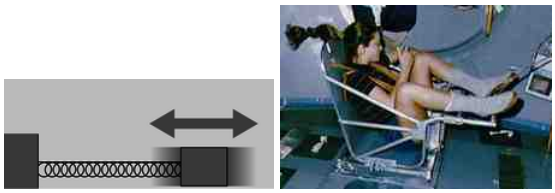


Figure d (left): The time for one cycle of vibration is related to the object's inertial mass.

Figure e (right): Astronaut Tamara Jernigan measures her inertial mass aboard the Space Shuttle.

1.1.1 Problem-solving techniques

How do we use a conservation law, such as conservation of mass, to solve problems? There are two basic techniques.

As an analogy, consider conservation of money, which makes it illegal for you to create dollar bills using your own inkjet printer. (Most people don't intentionally destroy their dollar bills, either!) Suppose the police notice that a particular store doesn't seem to have any customers, but the owner wears lots of gold jewelry and drives a BMW. They suspect that the store is a front for some kind of crime, perhaps counterfeiting. With intensive surveillance, there are two basic approaches they could use in their investigation. One method would be to have undercover agents try to find out how much money goes in the door, and how much money comes back out at the end of the day, perhaps by arranging through some trick to get access to the owner's briefcase in the morning and evening. If the amount of money that comes out every day is greater than the amount that went in, and if they're convinced there is no safe on the premises holding a large reservoir of money, then the owner must be counterfeiting. This inflow-equals-outflow technique is useful if we are sure that there is a region of space within which there is no supply of mass that is being built up or depleted.

Example 1: A stream of water

If you watch water flowing out of the end of a hose, you'll see that the stream of water is fatter near the mouth of the hose, 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.

f / Example 1.

self-check:

Suppose the you point the hose straight up, so that the water is rising rather than falling. What happens as the velocity gets smaller? What happens when the velocity becomes zero?

(answer in the back of the PDF version of the book)

How can we apply a conservation law, such as conservation of mass, in a situation where mass might be stored up somewhere? To use a crime analogy again, a prison could contain a certain number of prisoners, who are not allowed to flow in or out at will. In physics, this is known as a *closed system*. A guard might notice that a certain prisoner's cell is empty, but that doesn't mean he's escaped. He could be sick in the infirmary, or hard at work in the shop earning cigarette money. What prisons actually do is to count all their prisoners every day, and make sure today's total is the same as yesterday's. One way of stating a conservation law is that for a closed system, the total amount of stuff (mass, in this chapter) stays constant.

Example 2: Lavoisier and chemical reactions in a closed system

The French chemist Antoine-Laurent Lavoisier is considered the inventor of the concept of conservation of mass. Before Lavoisier, chemists had never systematically weighed their chemicals to quantify the amount of each substance that was undergoing reactions. They also didn't completely understand that gases were just another state of matter, and hadn't tried performing reactions in sealed chambers to determine whether gases were being consumed from or released into the air. For this they had at least one practical excuse, which is that if you perform a gas-releasing reaction in a sealed chamber with no room for expansion, you get an explosion! Lavoisier invented a balance that was capable of measuring milligram masses, and figured out how to do reactions in an upside-down bowl in a basin of water, so that the gases could expand by pushing out some of the water. In a crucial experiment, Lavoisier heated a red mercury compound, which we would now describe as mercury oxide (HgO), in such a sealed chamber. A gas was produced (Lavoisier later named it “oxygen”), driving out some of the water, and the red compound was transformed into silvery liquid mercury metal. The crucial point was that the total mass of the entire apparatus was exactly the same before and after the reaction. Based on many observations of this type, Lavoisier proposed a general law of nature, that mass is always conserved. (In earlier experiments, in which closed systems were not used, chemists had become convinced that there was a mysterious substance, phlogiston, involved in combustion and oxidation reactions, and that phlogiston's mass could be positive, negative, or zero depending on the situation!)

a / Portrait of Monsieur Lavoisier and His Wife, by Jacques-Louis David, 1788. Lavoisier invented the concept of conservation of mass. The husband is depicted with his scientific apparatus, while in the background on the left is the portfolio belonging to Madame Lavoisier, who is thought to have been a student of David's.

1.1.2 Delta notation

A convenient notation used throughout physics is Δ, the uppercase Greek letter delta, which indicates “change in” or “after minus before.” For example, if b represents how much money you have in the bank, then a deposit of $100 could be represented as $\Delta{}b=$100$. That is, the change in your balance was $100, or the balance after the transaction minus the balance before the transaction equals $100. A withdrawal would be indicated by $\Delta b < 0$. We represent “before” and “after” using the subscripts i (initial) and f (final), e.g., $\Delta b = b_f - b_i$. Often the delta notation allows more precision than English words. For instance, “time” can be used to mean a point in time (“now's the time”), t, or it could mean a period of time (“the whole time, he had spit on his chin”), Δt.

This notation is particularly convenient for discussing conserved quantities. The law of conservation of mass can be stated simply as $\Delta m = 0$, where m is the total mass of any closed system.

self-check:

If x represents the location of an object moving in one dimension, then how would positive and negative signs of Δx be interpreted?

(answer in the back of the PDF version of the book)

Discussion Questions

If an object had a straight-line $x - t$ graph with $\Delta x = 0$ and $\Delta t \neq 0$, what would be true about its velocity? What would this look like on a graph? What about $\Delta t = 0$ and $\Delta x \neq 0$?

Contributors and Attributions

2.2: Equivalence of Gravitational and Inertial Mass

We find experimentally that both gravitational and inertial mass are conserved to a high degree of precision for a great number of processes, including chemical reactions, melting, boiling, soaking up water with a sponge, and rotting of meat and vegetables. Now it's logically possible that both gravitational and inertial mass are conserved, but that there is no particular relationship between them, in which case we would say that they are separately conserved. On the other hand, the two conservation laws may be redundant, like having one law against murder and another law against killing people!

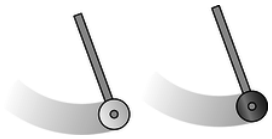


Figure a: The two pendulum bobs are constructed with equal gravitational masses. If their inertial masses are also equal, then each pendulum should take exactly the same amount of time per swing.

Here's an experiment that gets at the issue: stand up now and drop a coin and one of your shoes side by side. I used a 400-gram shoe and a 2-gram penny, and they hit the floor at the same time as far as I could tell by eye. This is an interesting result, but a physicist and an ordinary person will find it interesting for different reasons.

The layperson is surprised, since it would seem logical that heaver objects would always fall faster than light ones. However, it's fairly easy to prove that if air friction is negligible, any two objects made of the same substance must have identical motion when they fall. For instance, a 2-kg copper mass must exhibit the same falling motion as a 1-kg copper mass, because nothing would be changed by physically joining together two 1-kg copper masses to make a single 2-kg copper mass. Suppose, for example, that they are joined with a dab of glue; the glue isn't under any strain, because the two masses are doing the same thing side by side. Since the glue isn't really doing anything, it makes no difference whether the masses fall separately or side by side.[2]

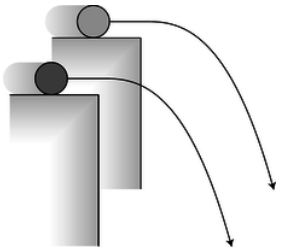


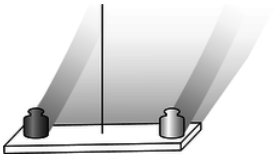
Figure b: If the cylinders have slightly unequal ratios of inertial to gravitational mass, their trajectories will be a little different.

What a physicist finds remarkable about the shoe-and-penny experiment is that it came out the way it did even though the shoe and the penny are made of *different* substances. There is absolutely no theoretical reason why this should be true. We could say that it happens because the greater gravitational mass of the shoe is exactly counteracted by its greater inertial mass, which makes it harder for gravity to get it moving, but that just leaves us wondering why inertial mass and gravitational mass are always in proportion to each other. It's possible that they are only approximately equivalent. Most of the mass of ordinary matter comes from neutrons and protons, and we could imagine, for instance, that neutrons and protons do not have exactly the same ratio of gravitational to inertial mass. This would show up as a different ratio of gravitational to inertial mass for substances containing different proportions of neutrons and protons.

Galileo did the first numerical experiments on this issue in the seventeenth century by rolling balls down inclined planes, although he didn't think about his results in these terms. A fairly easy way to improve on Galileo's accuracy is to use pendulums with bobs made of different materials. Suppose, for example, that we construct an aluminum bob and a brass bob, and use a double-pan balance to verify to good precision that their gravitational masses are equal. If we then measure the time required for each pendulum to perform a hundred cycles, we can check whether the results are the same. If their inertial masses are unequal, then the one with a smaller inertial mass will go through each cycle faster, since gravity has an easier time accelerating and decelerating it. With this type of experiment, one can easily verify that gravitational and inertial mass are proportional to each other to an accuracy of 10^{-3} or 10^{-4}.

In 1889, the Hungarian physicist Roland Eötvös used a slightly different approach to verify the equivalence of gravitational and inertial mass for various substances to an accuracy of about 10^{-8}, and the best such experiment, figure d, improved on even this

phenomenal accuracy, bringing it to the 10^{-12} level.[3] In all the experiments described so far, the two objects move along similar trajectories: straight lines in the penny-and-shoe and inclined plane experiments, and circular arcs in the pendulum version. The Eötvös-style experiment looks for differences in the objects' trajectories. The concept can be understood by imagining the following simplified version. Suppose, as in figure b, we roll a brass cylinder off of a tabletop and measure where it hits the floor, and then do the same with an aluminum cylinder, making sure that both of them go over the edge with precisely the same velocity. An object with zero gravitational mass would fly off straight and hit the wall, while an object with zero inertial mass would make a sudden 90-degree turn and drop straight to the floor. If the aluminum and brass cylinders have ordinary, but slightly unequal, ratios of gravitational to inertial mass, then they will follow trajectories that are just slightly different. In other words, if inertial and gravitational mass are not exactly proportional to each other for all substances, then objects made of different substances will have different trajectories in the presence of gravity.



c / A simplified drawing of an Eötvös-style experiment. If the two masses, made out of two different substances, have slightly different ratios of inertial to gravitational mass, then the apparatus will twist slightly as the earth spins.

A simplified drawing of a practical, high-precision experiment is shown in figure c. Two objects made of different substances are balanced on the ends of a bar, which is suspended at the center from a thin fiber. The whole apparatus moves through space on a complicated, looping trajectory arising from the rotation of the earth superimposed on the earth's orbital motion around the sun. Both the earth's gravity and the sun's gravity act on the two objects. If their inertial masses are not exactly in proportion to their gravitational masses, then they will follow slightly different trajectories through space, which will result in a very slight twisting of the fiber between the daytime, when the sun's gravity is pulling upward, and the night, when the sun's gravity is downward. Figure d shows a more realistic picture of the apparatus.

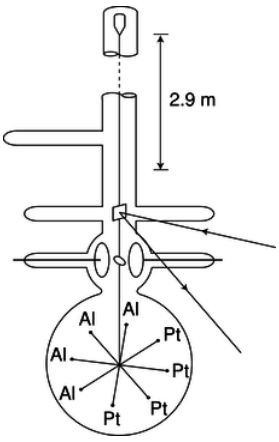


Figure d: A more realistic drawing of Braginskii and Panov's experiment. The whole thing was encased in a tall vacuum tube, which was placed in a sealed basement whose temperature was controlled to within 0.02°\textup{C}. The total mass of the platinum and aluminum test masses, plus the tungsten wire and the balance arms, was only 4.4 g. To detect tiny motions, a laser beam was bounced off of a mirror attached to the wire. There was so little friction that the balance would have taken on the order of several years to calm down completely after being put in place; to stop these vibrations, static electrical forces were applied through the two circular plates to provide very gentle twists on the ellipsoidal mass between them. After Braginskii and Panov.

This type of experiment, in which one expects a null result, is a tough way to make a career as a scientist. If your measurement comes out as expected, but with better accuracy than other people had previously achieved, your result is publishable, but won't be considered earth-shattering. On the other hand, if you build the most sensitive experiment ever, and the result comes out contrary to expectations, you're in a scary situation. You could be right, and earn a place in history, but if the result turns out to be due to a defect in your experiment, then you've made a fool of yourself.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

2.3: 1.3 Galilean Relativity

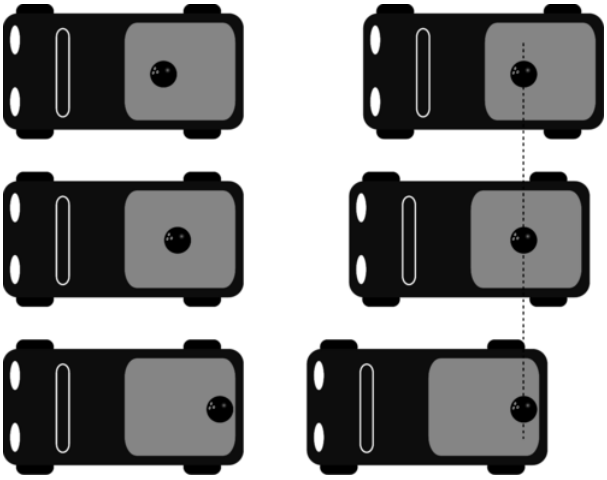
I defined inertial mass conceptually as a measure of how hard it is to *change* an object's state of motion, the implication being that if you don't interfere, the object's motion won't change. Most people, however, believe that objects in motion have a natural tendency to slow down. Suppose I push my refrigerator to the west for a while at 0.1 m/s, and then stop pushing. The average person would say fridge just naturally stopped moving, but let's imagine how someone in China would describe the fridge experiment carried out in my house here in California. Due to the rotation of the earth, California is moving to the east at about 400 m/s. A point in China at the same latitude has the same speed, but since China is on the other side of the planet, China's east is my west. (If you're finding the three-dimensional visualization difficult, just think of China and California as two freight trains that go past each other, each traveling at 400 m/s.) If I insist on thinking of my dirt as being stationary, then China and its dirt are moving at 800 m/s to my west. From China's point of view, however, it's California that is moving 800 m/s in the opposite direction (my east). When I'm pushing the fridge to the west at 0.1 m/s, the observer in China describes its speed as 799.9 m/s. Once I stop pushing, the fridge speeds back up to 800 m/s. From my point of view, the fridge “naturally” slowed down when I stopped pushing, but according to the observer in China, it “naturally” sped up!

What's really happening here is that there's a tendency, due to friction, for the fridge to stop moving *relative to the floor*. In general, only relative motion has physical significance in physics, not absolute motion. It's not even possible to define absolute motion, since there is no special reference point in the universe that everyone can agree is at rest. Of course if we want to measure motion, we do have to pick some arbitrary reference point which we will say is standing still, and we can then define x, y, and z coordinates extending out from that point, which we can define as having $x = 0$, $y = 0$, $z = 0$. Setting up such a system is known as choosing a *frame of reference*. The local dirt is a natural frame of reference for describing a game of basketball, but if the game was taking place on the deck of a moving ocean liner, we would probably pick a frame of reference in which the deck was at rest, and the land was moving.

a / Galileo Galilei (1564-1642).

Galileo was the first scientist to reason along these lines, and we now use the term Galilean relativity to refer to a somewhat modernized version of his principle. Roughly speaking, the principle of Galilean relativity states that the same laws of physics apply in any frame of reference that is moving in a straight line at constant speed. We need to refine this statement, however, since it is not necessarily obvious which frames of reference are going in a straight line at constant speed. A person in a pickup truck pulling away from a stoplight could admit that the car's velocity is changing, or she could insist that the truck is at rest, and the meter on the dashboard is going up because the asphalt picked that moment to start moving faster and faster backward! Frames of reference are not all created equal, however, and the accelerating truck's frame of reference is not as good as the asphalt's. We can tell, because a bowling ball in the back of the truck, as in figure c, appears to behave strangely in the driver's frame of reference: in her rear-view mirror, she sees the ball, initially at rest, start moving faster and faster toward the back of the truck. This goofy behavior is evidence that there is something wrong with her frame of reference. A person on the sidewalk, however, sees the ball as standing still. In the sidewalk's frame of reference, the truck pulls away from the ball, and this makes sense, because the truck is burning gas and using up energy to change its state of motion.

We therefore define an *inertial frame of reference* as one in which we never see objects change their state of motion without any apparent reason. The sidewalk is a pretty good inertial frame, and a car moving relative to the sidewalk at constant speed in a straight line defines a pretty good inertial frame, but a car that is accelerating or turning is not a inertial frame.



c / Left: In a frame of reference that speeds up with the truck, the bowling ball appears to change its state of motion for no reason. Right: In an inertial frame of reference, which the surface of the earth approximately is, the bowling ball stands still, which makes sense because there is nothing that would cause it to change its state of motion.

The principle of Galilean relativity states that inertial frames exist, and that the same laws of physics apply in all inertial frames of reference, regardless of one frame's straight-line, constant-speed motion relative to another.[4]

Another way of putting it is that all inertial frames are created equal. We can say whether one inertial frame is in motion or at rest relative to another, but there is no privileged “rest frame.” There is no experiment that comes out any different in laboratories in different inertial frames, so there is no experiment that could tell us which inertial frame is really, truly at rest.

b / The earth spins. People in Shanghai say they're at rest and people in Los Angeles are moving. Angelenos say the same about the Shanghainese.

Example 3: The speed of sound

▹ The speed of sound in air is only 340 m/s, so unless you live at a near-polar latitude, you're moving at greater than the speed of sound right now due to the Earth's rotation. In that case, why don't we experience exciting phenomena like sonic booms all the time? ▹ It might seem as though you're unprepared to deal with this question right now, since the only law of physics you know is conservation of mass, and conservation of mass doesn't tell you anything obviously useful about the speed of sound or sonic booms. Galilean relativity, however, is a blanket statement about all the laws of physics, so in a situation like this, it may let you predict the results of the laws of physics without actually knowing what all the laws are! If the laws of physics predict a certain value for the speed of sound, then they had better predict the speed of the sound relative to the air, not their speed relative to some special “rest frame.” Since the air is moving along with the rotation of the earth, we don't detect any special phenomena. To get a sonic boom, the source of the sound would have to be moving relative to the air.

Example 4: The Foucault pendulum

d / Foucault demonstrates his pendulum to an audience at a lecture in 1851.

Note that in the example of the bowling ball in the truck, I didn't claim the sidewalk was *exactly* a Galilean frame of reference. This is because the sidewalk is moving in a circle due to the rotation of the Earth, and is therefore changing the direction of its motion continuously on a 24-hour cycle. However, the curve of the motion is so gentle that under ordinary conditions we don't notice that the local dirt's frame of reference isn't quite inertial. The first demonstration of the noninertial nature of the earth-fixed frame of reference was by Foucault using a very massive pendulum (figure d) whose oscillations would persist for many hours without becoming imperceptible. Although Foucault did his demonstration in Paris, it's easier to imagine what would happen at the north pole: the pendulum would keep swinging in the same plane, but the earth would spin underneath it once every 24 hours. To someone standing in the snow, it would appear that the pendulum's plane of motion was twisting. The effect at latitudes less than 90 degrees turns out to be slower, but otherwise similar. The Foucault pendulum was the first definitive experimental proof that the earth really did spin on its axis, although scientists had been convinced of its rotation for a century based on more indirect evidence about the structure of the solar system.

e / Galileo's trial.

Although popular belief has Galileo being prosecuted by the Catholic Church for saying the earth rotated on its axis and also orbited the sun, Foucault's pendulum was still centuries in the future, so Galileo had no hard proof; Galileo's insights into relative versus absolute motion simply made it more plausible that the world could be spinning without producing dramatic effects, but didn't disprove the contrary hypothesis that the sun, moon, and stars went around the earth every 24 hours. Furthermore, the Church was much more liberal and enlightened than most people believe. It didn't (and still doesn't) require a literal interpretation of the Bible, and one of the Church officials involved in the Galileo affair wrote that “the Bible tells us how to go to heaven, not how the heavens go.” In other words, religion and science should be separate. The actual reason Galileo got in trouble is shrouded in mystery, since Italy in the age of the Medicis was a secretive place where unscrupulous people might settle a score with poison or a false accusation of heresy. What is certain is that Galileo's satirical style of scientific writing made many enemies among the powerful Jesuit scholars who were his intellectual opponents --- he compared one to a snake that doesn't know its own back is broken. It's also possible that the Church was far less upset by his astronomical work than by his support for atomism (discussed further in the next section). Some theologians perceived atomism as contradicting transubstantiation, the Church's doctrine that the holy bread and wine were literally transformed into the flesh and blood of Christ by the priest's blessing.

self-check:

What is incorrect about the following supposed counterexamples to the principle of inertia?

(1) When astronauts blast off in a rocket, their huge velocity does cause a physical effect on their bodies --- they get pressed back into their seats, the flesh on their faces gets distorted, and they have a hard time lifting their arms.

(2) When you're driving in a convertible with the top down, the wind in your face is an observable physical effect of your absolute motion.

(answer in the back of the PDF version of the book)

◊ Solved problem: a bug on a wheel — problem 12

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

2.4: A Preview of Some Modern Physics

“Mommy, why do you and Daddy have to go to work?” “To make money, sweetie-pie.” “Why do we need money?” “To buy food.” “Why does food cost money?” When small children ask a chain of “why” questions like this, it usually isn't too long before their parents end up saying something like, “Because that's just the way it is,” or, more honestly, “I don't know the answer.”

The same happens in physics. We may gradually learn to explain things more and more deeply, but there's always the possibility that a certain observed fact, such as conservation of mass, will never be understood on any deeper level. Science, after all, uses limited methods to achieve limited goals, so the ultimate reason for all existence will always be the province of religion. There is, however, an appealing explanation for conservation of mass, which is atomism, the theory that matter is made of tiny, unchanging particles. The atomic hypothesis dates back to ancient Greece, but the first solid evidence to support it didn't come until around the eighteenth century, and individual atoms were never detected until about 1900. The atomic theory implies not only conservation of mass, but a couple of other things as well.

First, it implies that the total mass of one particular element is conserved. For instance, lead and gold are both elements, and if we assume that lead atoms can't be turned into gold atoms, then the total mass of lead and the total mass of gold are separately conserved. It's as though there was not just a law against pickpocketing, but also a law against surreptitiously moving money from one of the victim's pockets to the other. It turns out, however, that although chemical reactions never change one type of atom into another, transmutation can happen in nuclear reactions, such as the ones that created most of the elements in your body out of the primordial hydrogen and helium that condensed out of the aftermath of the Big Bang.

Second, atomism implies that mass is *quantized*, meaning that only certain values of mass are possible and the ones in between can't exist. We can have three atoms of gold or four atoms of gold, but not three an a half. Although quantization of mass is a natural consequence of any theory in which matter is made up of tiny particles, it was discovered in the twentieth century that other quantities, such as energy, are quantized as well, which had previously not been suspected.

self-check:

Is money quantized?

(answer in the back of the PDF version of the book)

If atomism is starting to make conservation of mass seem inevitable to you, then it may disturb you to know that Einstein discovered it isn't really conserved. If you put a 50-gram iron nail in some water, seal the whole thing up, and let it sit on a fantastically precise balance while the nail rusts, you'll find that the system loses about 6×10^{-12} kg of mass by the time the nail has turned completely to rust. This has to do with Einstein's famous equation $E = mc^2$. Rusting releases heat energy, which then escapes out into the room. Einstein's equation states that this amount of heat, E, is equivalent to a certain amount of mass, m. The c in the c^2 is the speed of light, which is a large number, and a large amount of energy is therefore equivalent to a very small amount of mass, so you don't notice nonconservation of mass under ordinary conditions. What is really conserved is not the mass, m, but the mass-plus-energy, $E + mc^2$. The point of this discussion is not to get you to do numerical exercises with $E = mc^2$ (at this point you don't even know what units are used to measure energy), but simply to point out to you the empirical nature of the laws of physics. If a previously accepted theory is contradicted by an experiment, then the theory needs to be changed. This is also a good example of something called the *correspondence principle*, which is a historical observation about how scientific theories change: when a new scientific theory replaces an old one, the old theory is always contained within the new one as an approximation that works within a certain restricted range of situations. Conservation of mass is an extremely good approximation for all chemical reactions, since chemical reactions never release or consume enough energy to change the total mass by a large percentage. Conservation of mass would not have been accepted for 110 years as a fundamental principle of physics if it hadn't been verified over and over again by a huge number of accurate experiments.

Contributors and Attributions

Benjamin Crowell (Fullerton College).

2.5: Footnotes

1. You might think intuitively that the recoil velocity should be exactly one fourth of a meter per second, and you'd be right except that the wagon has some mass as well. Our present approach, however, only requires that we give a way to test for equality of masses. To predict the recoil velocity from scratch, we'd need to use conservation of momentum, which is discussed in a later chapter.
2. The argument only fails for objects light enough to be affected appreciably by air friction: a bunch of feathers falls differently if you wad them up because the pattern of air flow is altered by putting them together.
3. V.B. Braginskii and V.I. Panov, Soviet Physics JETP 34, 463 (1972).
4. The principle of Galilean relativity is extended on page 190.

2.6: Problems

1. Thermometers normally use either mercury or alcohol as their working fluid. If the level of the fluid rises or falls, does this violate conservation of mass?

2. The ratios of the masses of different types of atoms were determined a century before anyone knew any actual atomic masses in units of kg. One finds, for example, that when ordinary table salt, NaCl, is melted, the chlorine atoms bubble off as a gas, leaving liquid sodium metal. Suppose the chlorine escapes, so that its mass cannot be directly determined by weighing. Experiments show that when 1.00000 kg of NaCl is treated in this way, the mass of the remaining sodium metal is 0.39337 kg. Based on this information, determine the ratio of the mass of a chlorine atom to that of a sodium atom (answer check available at lightandmatter.com)

3. An atom of the most common naturally occurring uranium isotope breaks up spontaneously into a thorium atom plus a helium atom. The masses are as follows:

uranium	$3.95292849 \times 10^{25}$ kg
thorium	$3.88638748 \times 10^{25}$ kg
helium	6.646481×10^{27} kg

Each of these experimentally determined masses is uncertain in its last decimal place. Is mass conserved in this process to within the accuracy of the experimental data? How would you interpret this?

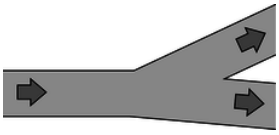
4. If two spherical water droplets of radius b combine to make a single droplet, what is its radius? (Assume that water has constant density.)

5. Make up an experiment that would test whether mass is conserved in an animal's metabolic processes.

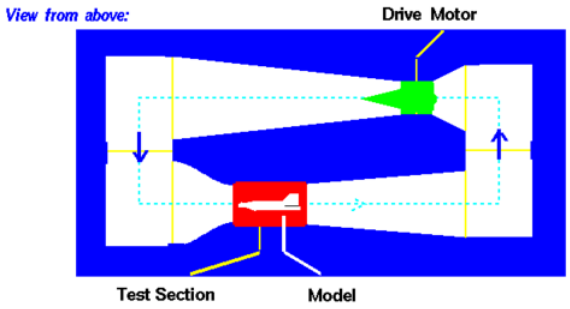
6. The figure shows a hydraulic jack. What is the relationship between the distance traveled by the plunger and the distance traveled by the object being lifted, in terms of the cross-sectional areas?

7. In an example in this chapter, I argued that a stream of water must change its cross-sectional area as it rises or falls. Suppose that the stream of water is confined to a constant-diameter pipe. Which assumption breaks down in this situation?

8. A river with a certain width and depth splits into two parts, each of which has the same width and depth as the original river. What can you say about the speed of the current after the split?



9. The diagram shows a cross-section of a wind tunnel of the kind used, for example, to test designs of airplanes. Under normal conditions of use, the density of the air remains nearly constant throughout the whole wind tunnel. How can the speed of the air be controlled and calculated? (Diagram by NASA, Glenn Research Center.)

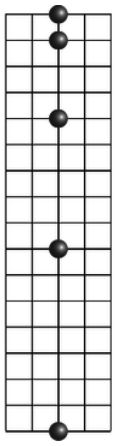


10. A water wave is in a tank that extends horizontally from $x = 0$ to $x = a$, and from $z = 0$ to $z = b$. We assume for simplicity that at a certain moment in time the height y of the water's surface only depends on x, not z, so that we can effectively ignore the z coordinate. Under these assumptions, the total volume of the water in the tank is

$$V = b \int_0^a y(x)\, dx \tag{2.6.1}$$

Since the density of the water is essentially constant, conservation of mass requires that V is always the same. When the water is calm, we have $y = h$, where $h = V/ab$. If two different wave patterns move into each other, we might imagine that they would add in the sense that $y_{total} - h = (y_1 - h) + (y_2 - h)$. Show that this type of addition is consistent with conservation of mass.

11. The figure shows the position of a falling ball at equal time intervals, depicted in a certain frame of reference. On a similar grid, show how the ball's motion would appear in a frame of reference that was moving horizontally at a speed of one box per unit time relative to the first frame.



12. (solution in the pdf version of the book) The figure shows the motion of a point on the rim of a rolling wheel. (The shape is called a cycloid.) Suppose bug A is riding on the rim of the wheel on a bicycle that is rolling, while bug B is on the spinning wheel of a bike that is sitting upside down on the floor. Bug A is moving along a cycloid, while bug B is moving in a circle. Both wheels are doing the same number of revolutions per minute. Which bug has a harder time holding on, or do they find it equally difficult?

Contributors and Attributions

CHAPTER OVERVIEW

3: Conservation of Energy

Do you pronounce it Joule's to rhyme with schools,
Joule's to rhyme with Bowls,
or Joule's to rhyme with Scowls?
Whatever you call it, by Joule's,
or Joule's,
or Joule's, it's good! -- *Advertising slogan of the Joule brewery. The name, and the corresponding unit of energy, are now usually pronounced so as to rhyme with "school."*

Topic hierarchy

Thumbnail: Roller coaster "Blue Fire" at Europa Park. (CC SA 3.0; Coaster J).

3.1: Energy

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

3.2: Numerical Techniques

Engineering majors are a majority of the students in the kind of physics course for which this book is designed, so most likely you fall into that category. Although you surely recognize that physics is an important part of your training, if you've had any exposure to how engineers really work, you're probably skeptical about the flavor of problem-solving taught in most science courses. You realize that not very many practical engineering calculations fall into the narrow range of problems for which an exact solution can be calculated with a piece of paper and a sharp pencil. Real-life problems are usually complicated, and typically they need to be solved by number-crunching on a computer, although we can often gain insight by working simple approximations that have algebraic solutions. Not only is numerical problem-solving more useful in real life, it's also educational; as a beginning physics student, I really only felt like I understood projectile motion after I had worked it both ways, using algebra and then a computer program. (This was back in the days when 64 kilobytes of memory was considered a lot.)

In this section, we'll start by seeing how to apply numerical techniques to some simple problems for which we know the answer in "closed form," i.e., a single algebraic expression without any calculus or infinite sums. After that, we'll solve a problem that would have made you world-famous if you could have done it in the seventeenth century using paper and a quill pen! Before you continue, you should read Appendix 1 on page 908 that introduces you to the Python programming language.

First let's solve the trivial problem of finding how much time it takes an object moving at speed v to travel a straight-line distance dist. This closed-form answer is, of course, dist/v, but the point is to introduce the techniques we can use to solve other problems of this type. The basic idea is to divide the distance up into n equal parts, and add up the times required to traverse all the parts. The following Python function does the job. Note that you shouldn't type in the line numbers on the left, and you don't need to type in the comments, either. I've omitted the prompts >>> and ... in order to save space.

```
import math
def time1(dist,v,n):
  x=0 # Initialize the position.
  dx = dist/n # Divide dist into n equal parts.
  t=0 # Initialize the time.
  for i in range(n):
    x = x+dx # Change x.
    dt=dx/v # time=distance/speed
    t=t+dt # Keep track of elapsed time.
  return t
```

How long does it take to move 1 meter at a constant speed of 1 m/s? If we do this,

```
>>> print(time1(1.0,1.0,10)) # dist, v, n
0.99999999999999989
```

Python produces the expected answer by dividing the distance into ten equal 0.1-meter segments, and adding up the ten 0.1-second times required to traverse each one. Since the object moves at constant speed, it doesn't even matter whether we set n to 10, 1, or a million:

```
>>> print(time1(1.0,1.0,1)) # dist, v, n
1.0
```

Now let's do an example where the answer isn't obvious to people who don't know calculus: how long does it take an object to fall through a height h, starting from rest? We know from example 8 on page 83 that the exact answer, found using calculus, is *[Math Processing Error]*. Let's see if we can reproduce that answer numerically. The main difference between this program and the previous one is that now the velocity isn't constant, so we need to update it as we go along. Conservation of energy gives *[Math Processing Error]* for the velocity *[Math Processing Error]* at height *[Math Processing Error]*, so *[Math Processing Error]*. (We

choose the negative root because the object is moving down, and our coordinate system has the positive *[Math Processing Error]* axis pointing up.)

```
import math
def time2(h,n):
  g=9.8 # gravitational field
  y=h # Initialize the height.
  v=0 # Initialize the velocity.
  dy = -h/n # Divide h into n equal parts.
  t=0 # Initialize the time.
  for i in range(n):
    y = y+dy # Change y. (Note dy<0.)
    v = -math.sqrt(2*g*(h-y)) # from cons. of energy
    dt=dy/v # dy and v are <0, so dt is >0
    t=t+dt # Keep track of elapsed time.
  return t
```

For h=1.0 m, the closed-form result is *[Math Processing Error]*. With the drop split up into only 10 equal height intervals, the numerical technique provides a pretty lousy approximation:

```
>>> print(time2(1.0,10)) # h, n
0.35864270709233342
```

But by increasing n to ten thousand, we get an answer that's as close as we need, given the limited accuracy of the raw data:

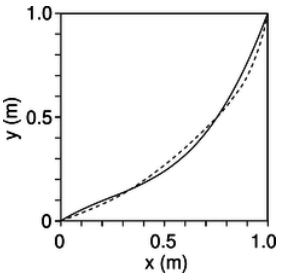
```
>>> print(time2(1.0,10000)) # h, n
0.44846664060793945
```

A subtle point is that we changed y in line 9, and *then* on line 10 we calculated v, which depends on y. Since y is only changing by a ten-thousandth of a meter with each step, you might think this wouldn't make much of a difference, and you'd be almost right, except for one small problem: if we swapped lines 9 and 10, then the very first time through the loop, we'd have v=0, which would produce a division-by-zero error when we calculated dt! Actually what would make the most sense would be to calculate the velocity at height y and the velocity at height y+dy (recalling that dy is negative), average them together, and use that value of y to calculate the best estimate of the velocity between those two points. Since the acceleration is constant in the present example, this modification results in a program that gives an exact result even for n=1:

```
import math
def time3(h,n):
  g=9.8
  y=h
  v=0
  dy = -h/n
  t=0
  for i in range(n):
    y_old = y
    y = y+dy
    v_old = math.sqrt(2*g*(h-y_old))
    v = math.sqrt(2*g*(h-y))
    v_avg = -(v_old+v)/2.
```

```
    dt=dy/v_avg
    t=t+dt
  return t
```

```
>>> print(time3(1.0,1)) # h, n
0.45175395145262565
```



a / Approximations to the brachistochrone curve using a third-order polynomial (solid line), and a seventh-order polynomial (dashed). The latter only improves the time by four milliseconds.

Now we're ready to attack a problem that challenged the best minds of Europe back in the days when there were no computers. In 1696, the mathematician Johann Bernoulli posed the following famous question. Starting from rest, an object slides frictionlessly over a curve joining the point *[Math Processing Error]* to the point *[Math Processing Error]*. Of all the possible shapes that such a curve could have, which one gets the object to its destination in the least possible time, and how much time does it take? The optimal curve is called the *brachistochrone*, from the Greek “short time.” The solution to the brachistochrone problem evaded Bernoulli himself, as well as Leibniz, who had been one of the inventors of calculus. The English physicist Isaac Newton, however, stayed up late one night after a day's work running the royal mint, and, according to legend, produced an algebraic solution at four in the morning. He then published it anonymously, but Bernoulli is said to have remarked that when he read it, he knew instantly from the style that it was Newton --- he could “tell the lion from the mark of his claw.”

Rather than attempting an exact algebraic solution, as Newton did, we'll produce a numerical result for the shape of the curve and the minimum time, in the special case of *[Math Processing Error]*=1.0 m and *[Math Processing Error]*=1.0 m. Intuitively, we want to start with a fairly steep drop, since any speed we can build up at the start will help us throughout the rest of the motion. On the other hand, it's possible to go too far with this idea: if we drop straight down for the whole vertical distance, and then do a right-angle turn to cover the horizontal distance, the resulting time of 0.68 s is quite a bit longer than the optimal result, the reason being that the path is unnecessarily long. There are infinitely many possible curves for which we could calculate the time, but let's look at third-order polynomials,

[Math Processing Error]

where we require *[Math Processing Error]* in order to make the curve pass through the point *[Math Processing Error]*. The Python program, below, is not much different from what we've done before. The function only asks for *[Math Processing Error]* and *[Math Processing Error]*, and calculates *[Math Processing Error]* internally at line 4. Since the motion is two-dimensional, we have to calculate the distance between one point and the next using the Pythagorean theorem, at line 16.

```
import math
def timeb(a,b,c1,c2,n):
  g=9.8
  c3 = (b-c1*a-c2*a**2)/(a**3)
  x=a
  y=b
  dx = -a/n
  t=0
```

```
  for i in range(n):
    y_old = y
    x = x+dx
    y = c1*x+c2*x**2+c3*x**3
    dy = y-y_old
    v_old = math.sqrt(2*g*(b-y_old))
    v = math.sqrt(2*g*(b-y))
    v_avg = (v_old+v)/2.
    ds = math.sqrt(dx**2+dy**2) # Pythagorean thm.
    dt=ds/v_avg
    t=t+dt
  return t
```

As a first guess, we could try a straight diagonal line, *[Math Processing Error]*, which corresponds to setting *[Math Processing Error]*, and all the other coefficients to zero. The result is a fairly long time:

```
>>> a=1.
>>> b=1.
>>> n=10000
>>> c1=1.
>>> c2=0.
>>> print(timeb(a,b,c1,c2,n))
0.63887656499994161
```

What we really need is a curve that's very steep on the right, and flatter on the left, so it would actually make more sense to try *[Math Processing Error]*:

```
>>> c1=0.
>>> c2=0.
>>> print(timeb(a,b,c1,c2,n))
0.59458339947087069
```

This is a significant improvement, and turns out to be only a hundredth of a second off of the shortest possible time! It's possible, although not very educational or entertaining, to find better approximations to the brachistochrone curve by fiddling around with the coefficients of the polynomial by hand. The real point of this discussion was to give an example of a nontrivial problem that can be attacked successfully with numerical techniques. I found the first approximation shown in figure a,

[Math Processing Error]

by using the program listed in appendix 2 on page 911 to search automatically for the optimal curve. The seventh-order approximation shown in the figure came from a straightforward extension of the same program.

Contributors and Attributions

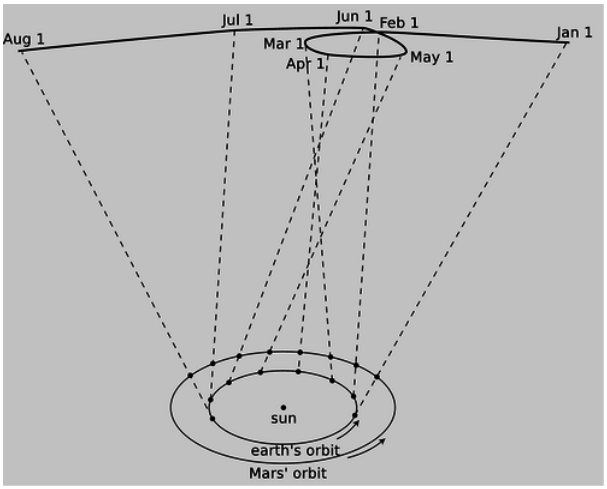
3.3: Gravitational Phenomena

Cruise your radio dial today and try to find any popular song that would have been imaginable without Louis Armstrong. By introducing solo improvisation into jazz, Armstrong took apart the jigsaw puzzle of popular music and fit the pieces back together in a different way. In the same way, Newton reassembled our view of the universe. Consider the titles of some recent physics books written for the general reader: **The God Particle**, **Dreams of a Final Theory**. When the subatomic particle called the neutrino was recently proven for the first time to have mass, specialists in cosmology began discussing seriously what effect this would have on calculations of the evolution of the universe from the Big Bang to its present state. Without the English physicist Isaac Newton, such attempts at universal understanding would not merely have seemed ambitious, they simply would not have occurred to anyone.

This section is about Newton's theory of gravity, which he used to explain the motion of the planets as they orbited the sun. Newton tosses off a general treatment of motion in the first 20 pages of his **Mathematical Principles of Natural Philosophy**, and then spends the next 130 discussing the motion of the planets. Clearly he saw this as the crucial scientific focus of his work. Why? Because in it he showed that the same laws of nature applied to the heavens as to the earth, and that the gravitational interaction that made an apple fall was the same as the as the one that kept the earth's motion from carrying it away from the sun.

2.3.1 Kepler's laws

Newton wouldn't have been able to figure out *why* the planets move the way they do if it hadn't been for the astronomer Tycho Brahe (1546-1601) and his protege Johannes Kepler (1571-1630), who together came up with the first simple and accurate description of *how* the planets actually do move. The difficulty of their task is suggested by the figure below, which shows how the relatively simple orbital motions of the earth and Mars combine so that as seen from earth Mars appears to be staggering in loops like a drunken sailor.

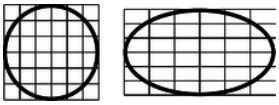


d / As the earth and Mars revolve around the sun at different rates, the combined effect of their motions makes Mars appear to trace a strange, looped path across the background of the distant stars.

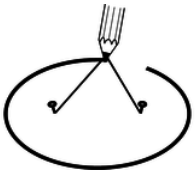
Brahe, the last of the great naked-eye astronomers, collected extensive data on the motions of the planets over a period of many years, taking the giant step from the previous observations' accuracy of about 10 minutes of arc (10/60 of a degree) to an unprecedented 1 minute. The quality of his work is all the more remarkable considering that his observatory consisted of four giant brass protractors mounted upright in his castle in Denmark. Four different observers would simultaneously measure the position of a planet in order to check for mistakes and reduce random errors.

With Brahe's death, it fell to his former assistant Kepler to try to make some sense out of the volumes of data. After 900 pages of calculations and many false starts and dead-end ideas, Kepler finally synthesized the data into the following three laws:

Kepler's elliptical orbit law: The planets orbit the sun in elliptical orbits with the sun at one focus.

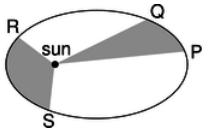


a / An ellipse is circle that has been distorted by shrinking and stretching along perpendicular axes.



b / An ellipse can be constructed by tying a string to two pins and drawing like this with a pencil stretching the string taut. Each pin constitutes one focus of the ellipse.

Kepler's equal-area law: The line connecting a planet to the sun sweeps out equal areas in equal amounts of time.

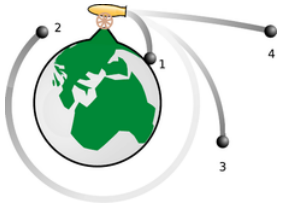


c / If the time interval taken by the planet to move from P to Q is equal to the time interval from R to S, then according to Kepler's equal-area law, the two shaded areas are equal. The planet is moving faster during time interval RS than it was during PQ, because gravitational energy has been transformed into kinetic energy.

Kepler's law of periods: The time required for a planet to orbit the sun, called its period, T, is proportional to the long axis of the ellipse raised to the 3/2 power. The constant of proportionality is the same for all the planets.

Although the planets' orbits are ellipses rather than circles, most are very close to being circular. The earth's orbit, for instance, is only flattened by 1.7% relative to a circle. In the special case of a planet in a circular orbit, the two foci (plural of “focus”) coincide at the center of the circle, and Kepler's elliptical orbit law thus says that the circle is centered on the sun. The equal-area law implies that a planet in a circular orbit moves around the sun with constant speed. For a circular orbit, the law of periods then amounts to a statement that the time for one orbit is proportional to $r^{3/2}$, where r is the radius. If all the planets were moving in their orbits at the same speed, then the time for one orbit would simply depend on the circumference of the circle, so it would only be proportional to r to the first power. The more drastic dependence on $r^{3/2}$ means that the outer planets must be moving more slowly than the inner planets.Our main focus in this section will be to use the law of periods to deduce the general equation for gravitational energy. The equal-area law turns out to be a statement on conservation of angular momentum, which is discussed in chapter 4. We'll demonstrate the elliptical orbit law numerically in chapter 3, and analytically in chapter 4.2.3.2 Circular orbits Kepler's laws say that planets move along elliptical paths (with circles as a special case), which would seem to contradict the proof on page 90 that objects moving under the influence of gravity have parabolic trajectories. Kepler was right. The parabolic path was really only an approximation, based on the assumption that the gravitational field is constant, and that vertical lines are all parallel. In figure e, trajectory 1 is an ellipse, but it gets chopped off when the cannonball hits the earth, and the small piece of it that is above ground is nearly indistinguishable from a parabola. Our goal is to connect the previous calculation of parabolic trajectories, $y = (g/2v^2)x^2$, with Kepler's data for planets orbiting the sun in nearly circular orbits. Let's start by thinking in terms of an orbit that circles the earth, like orbit 2 in figure e. It's more natural now to choose a coordinate system with its origin at the center of the earth, so the parabolic approximation becomes $y = r - (g/2v^2)x^2$, where r is the distance from the center of the earth. For small values of x, i.e., when the cannonball hasn't traveled very far from the muzzle of the gun, the parabola is still a good approximation to the actual circular orbit, defined by the Pythagorean theorem, $r^2 = x^2 + y^2$, or $y = r\sqrt{1 - x^2/r^2}$. For small values of x, we can use the approximation $\sqrt{1+\epsilon} \approx 1 + \epsilon/2$ to find $y \approx r - (1/2r)x^2$. Setting this equal to the equation of the parabola, we have $g/2v^2 = (1/2r)$, or

$$v = \sqrt{gr} \text{[condition for a circular orbit]}.$$



e / A cannon fires cannonballs at different velocities, from the top of an imaginary mountain that rises above the earth's atmosphere. This is almost the same as a figure Newton included in his **Mathematical Principles**.

Example 14: Low-earth orbit

To get a feel for what this all means, let's calculate the velocity required for a satellite in a circular low-earth orbit. Real low-earth-orbit satellites are only a few hundred km up, so for purposes of rough estimation we can take r to be the radius of the earth, and g is not much less than its value on the earth's surface, 10 $\mathrm{m/s^2}$. Taking numerical data from Appendix 5, we have

$$\begin{aligned} v &= \sqrt{gr} \\ &= \sqrt{(10\ \mathrm{m/s^2})(6.4\times 10^3\ \mathrm{km})} \\ &= \sqrt{(10\ \mathrm{m/s^2})(6.4\times 10^6\ \mathrm{m})} \\ &= \sqrt{6.4\times 10^7\ \mathrm{m^2/s^2}} \\ &= 8000\ \mathrm{m/s} \end{aligned}$$

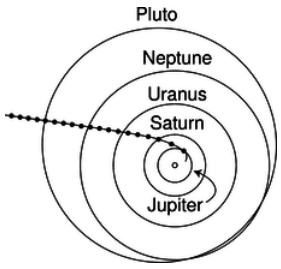
(about twenty times the speed of sound).In one second, the satellite moves 8000 m horizontally. During this time, it drops the same distance any other object would: about 5 m. But a drop of 5 m over a horizontal distance of 8000 m is just enough to keep it at the same altitude above the earth's curved surface.

2.3.3 The sun's gravitational field

We can now use the circular orbit condition $v=\sqrt{gr}$, combined with Kepler's law of periods, $T\propto r^{3/2}$ for circular orbits, to determine how the sun's gravitational field falls off with distance.[8] From there, it will be just a hop, skip, and a jump to get to a universal description of gravitational interactions.The velocity of a planet in a circular orbit is proportional to r/T, so

$$\begin{aligned} r/T &\propto \sqrt{gr} \\ r/r^{3/2} &\propto \sqrt{gr} \\ g &\propto 1/r^2 \end{aligned}$$

If gravity behaves systematically, then we can expect the same to be true for the gravitational field created by any object, not just the sun.There is a subtle point here, which is that so far, r has just meant the radius of a circular orbit, but what we have come up with smells more like an equation that tells us the strength of the gravitational field made by some object (the sun) if we know how far we are from the object. In other words, we could reinterpret r as the distance from the sun.



i / The Pioneer 10 space probe's trajectory from 1974 to 1992, with circles marking its position at one-year intervals. After its 1974 slingshot maneuver around Jupiter, the probe's motion was determined almost exclusively by the sun's gravity.

2.3.4 Gravitational energy in general

We now want to find an equation for the gravitational energy of any two masses that attract each other from a distance r. We assume that r is large enough compared to the distance between the objects so that we don't really have to worry about whether r is measured from center to center or in some other way. This would be a good approximation for describing the solar system, for example, since the sun and planets are small compared to the distances between them --- that's why you see Venus (the “evening star”) with your bare eyes as a dot, not a disk.

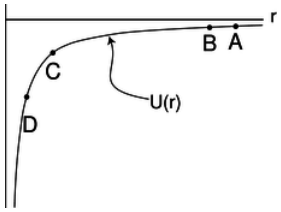
The equation we seek is going to give the gravitational energy, U, as a function of m_1, m_2, and r. We already know from experience with gravity near the earth's surface that U is proportional to the mass of the object that interacts with the earth gravitationally, so it makes sense to assume the relationship is symmetric: U is presumably proportional to the product $m_1 m_2$. We can no longer assume $\Delta U \propto \Delta r$, as in the earth's-surface equation $\Delta U = mg\Delta y$, since we are trying to construct an equation that would be valid for all values of r, and g depends on r. We can, however, consider an infinitesimally small change in distance dr, for which we'll have $dU = m_2 g_1 dr$, where g_1 is the gravitational field created by m_1. (We could just as well have written this as $dU = m_1 g_2 dr$, since we're not assuming either mass is “special” or “active.”) Integrating this equation, we have

$$\begin{aligned}\int dU &= \int m_2 g_1 dr \\ U &= m_2 \int g_1 dr \\ U &\propto m_1 m_2 \int \frac{1}{r^2} dr \\ U &\propto -\frac{m_1 m_2}{r},\end{aligned}$$

where we're free to take the constant of integration to be equal to zero, since gravitational energy is never a well-defined quantity in absolute terms. Writing G for the constant of proportionality, we have the following fundamental description of gravitational interactions:

$$U = -\frac{G m_1 m_2}{r} \begin{bmatrix}\text{gravitational energy of two masses} \\ \text{separated by a distance } r\end{bmatrix}$$

We'll refer to this as Newton's law of gravity, although in reality he stated it in an entirely different form, which turns out to be mathematically equivalent to this one.



f / The gravitational energy $U = -Gm_1 m_2/r$ graphed as a function of r.

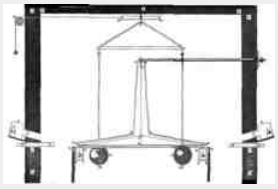
Let's interpret his result. First, don't get hung up on the fact that it's negative, since it's only differences in gravitational energy that have physical significance. The graph in figure f could be shifted up or down without having any physical effect. The slope of this graph relates to the strength of the gravitational field. For instance, suppose figure f is a graph of the gravitational energy of an asteroid interacting with the sun. If the asteroid drops straight toward the sun, from A to B, the decrease in gravitational energy is very small, so it won't speed up very much during that motion. Points C and D, however, are in a region where the graph's slope is much greater. As the asteroid moves from C to D, it loses a lot of gravitational energy, and therefore speeds up considerably. This is due to the stronger gravitational field.

Example 15: Determining G

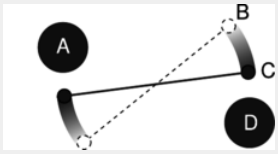
The constant G is not easy to determine, and Newton went to his grave without knowing an accurate value for it. If we knew the mass of the earth, then we could easily determine G from experiments with terrestrial gravity, but the only way to determine the mass of the earth accurately in units of kilograms is by finding G and reasoning the other way around! (If you estimate the average density of the earth, you can make at least a rough estimate of G.) Figures g and h show how G was first measured by Henry Cavendish in the nineteenth century.The rotating arm is released from rest, and the kinetic energy of the two moving balls is measured when they pass position C. Conservation of energy gives

$$-2\frac{GMm}{r_{BA}}-2\frac{GMm}{r_{BD}}=-2\frac{GMm}{r_{CA}}-2\frac{GMm}{r_{CD}}+2K,$$

where M is the mass of one of the large balls, m is the mass of one of the small ones, and the factors of two, which will cancel, occur because every energy is mirrored on the opposite side of the apparatus. (As discussed on page 102, it turns out that we get the right result by measuring all the distances from the center of one sphere to the center of the other.) This can easily be solved for G. The best modern value of G, from later versions of the same experiment, is $6.67\times10^{-11}\ \mathrm{J\cdot m/kg^2}$.



g / Cavendish's original drawing of the apparatus for his experiment, discussed in example 15. The room was sealed to exclude air currents, and the motion was observed through telescopes sticking through holes in the walls.



h / A simplified drawing of the Cavendish experiment, viewed from above. The rod with the two small masses on the ends hangs from a thin fiber, and is free to rotate.

Example 16: Escape velocity

▷ The Pioneer 10 space probe was launched in 1972, and continued sending back signals for 30 years. In the year 2001, not long before contact with the probe was lost, it was about $1.2\times10^{13}\ \mathrm{m}$ from the sun, and was moving almost directly away from the sun at a velocity of $1.21\times10^{4}\ \mathrm{m}$. The mass of the sun is $1.99\times10^{30}\ \mathrm{kg}$. Will Pioneer 10 escape permanently, or will it fall back into the solar system?

▷ We want to know whether there will be a point where the probe will turn around. If so, then it will have zero kinetic energy at the turnaround point:

$$\begin{aligned}K_i+U_i&=U_f\\ \frac{1}{2}mv^2-\frac{GMm}{r_i}&=-\frac{GMm}{r_f}\\ \frac{1}{2}v^2-\frac{GM}{r_i}&=-\frac{GM}{r_f},\end{aligned}$$

where M is the mass of the sun, m is the (irrelevant) mass of the probe, and r_f is the distance from the sun of the hypothetical turnaround point. Plugging in numbers on the left, we get a positive result. There can therefore be no solution, since the right side is negative. There won't be any turnaround point, and Pioneer 10 is never coming back.

The minimum velocity required for this to happen is called *escape velocity*. For speeds above escape velocity, the orbits are open-ended hyperbolas, rather than repeating elliptical orbits. In figure i, Pioneer's hyperbolic trajectory becomes almost indistinguishable from a line at large distances from the sun. The motion slows perceptibly in the first few years after 1974, but later the speed becomes nearly constant, as shown by the nearly constant spacing of the dots.

The gravitational field

We got the energy equation $U = -Gm_1m_2/r$ by integrating $g \propto 1/r^2$ and then inserting a constant of proportionality to make the proportionality into an equation. The opposite of an integral is a derivative, so we can now go backwards and insert a constant of proportionality in $g \propto 1/r^2$ that will be consistent with the energy equation:

$$\begin{aligned} dU &= m_2 g_1 dr \\ g_1 &= \frac{1}{m_2}\frac{dU}{dr} \\ &= \frac{1}{m_2}\frac{d}{dr}\left(-\frac{Gm_1m_2}{r}\right) \\ &= -Gm_1\frac{d}{dr}\left(\frac{1}{r}\right) \\ &= \frac{Gm_1}{r^2} \end{aligned}$$

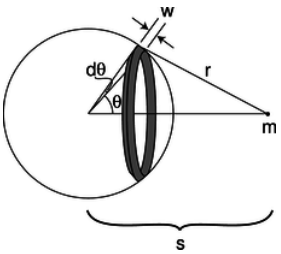
This kind of inverse-square law occurs all the time in nature. For instance, if you go twice as far away from a lightbulb, you receive 1/4 as much light from it, because as the light spreads out, it is like an expanding sphere, and a sphere with twice the radius has four times the surface area. It's like spreading the same amount of peanut butter on four pieces of bread instead of one --- we have to spread it thinner.

Discussion Questions

◊ A bowling ball interacts gravitationally with the earth. Would it make sense for the gravitational energy to be inversely proportional to the distance between their surfaces rather than their centers?

2.3.5 The shell theorem

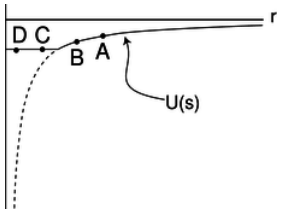
Newton's great insight was that gravity near the earth's surface was the same kind of interaction as the one that kept the planets from flying away from the sun. He told his niece that the idea came to him when he saw an apple fall from a tree, which made him wonder whether the earth might be affecting the apple and the moon in the same way. Up until now, we've generally been dealing with gravitational interactions between objects that are small compared to the distances between them, but that assumption doesn't apply to the apple. A kilogram of dirt a few feet under his garden in England would interact much more strongly with the apple than a kilogram of molten rock deep under Australia, thousands of miles away. Also, we know that the earth has some parts that are more dense, and some parts that are less dense. The solid crust, on which we live, is considerably less dense than the molten rock on which it floats. By all rights, the computation of the total gravitational energy of the apple should be a horrendous mess. Surprisingly, it turns out to be fairly simple in the end. First, we note that although the earth doesn't have the same density throughout, it does have spherical symmetry: if we imagine dividing it up into thin concentric shells, the density of each shell is uniform.



j / A spherical shell of mass M interacts with a pointlike mass m.

Second, it turns out that a uniform spherical shell interacts with external masses as if all its mass were concentrated at its center.

\mythmhdr{The shell theorem} The gravitational energy of a uniform spherical shell of mass M interacting with a pointlike mass m outside it equals $-GMm/s$, where s is the center-to-center distance. If mass m is inside the shell, then the energy is constant, i.e., the shell's interior gravitational field is zero.



k / The gravitational energy of a mass m at a distance s from the center of a hollow spherical shell of mass.

\mythmhdr{Proof} Let b be the radius of the shell, h its thickness, and ρ its density. Its volume is then V=(area)(thickness)= $4\pi b^2 h$, and its mass is $M = \rho V = 4\pi\rho b^2 h$. The strategy is to divide the shell up into rings as shown in figure j, with each ring extending from θ to $\theta + d\theta$. Since the ring is infinitesimally skinny, its entire mass lies at the same distance, r, from mass m. The width of such a ring is found by the definition of radian measure to be $w = bd\theta$, and its mass is $dM = (\rho)$(circumference)(thickness)(width)= $(\rho)(2\pi b\sin\theta)(h)(bd\theta)$=$2\pi\rho b^2 h\sin\theta d\theta$. The gravitational energy of the ring interacting with mass m is therefore

$$\begin{aligned} dU &= -\frac{GmdM}{r} \\ &= -2\pi G\rho b^2 hm\frac{\sin\theta d\theta}{r}. \end{aligned}$$

Integrating both sides, we find the total gravitational energy of the shell:

$$U = -2\pi G\rho b^2 hm\int_0^\pi \frac{\sin\theta d\theta}{r}$$

The integral has a mixture of the variables r and θ, which are related by the law of cosines,

$$r^2 = b^2 + s^2 - 2bs\cos\theta,$$

and to evaluate the integral, we need to get everything in terms of either r and dr or θ and $d\theta$. The relationship between the differentials is found by differentiating the law of cosines,

$$2rdr = 2bs\sin\theta d\theta,$$

and since $\sin\theta d\theta$ occurs in the integral, the easiest path is to substitute for it, and get everything in terms of r and dr:

$$\begin{aligned} U &= -\frac{2\pi G\rho bhm}{s}\int_{s-b}^{s+b} dr \\ &= -\frac{4\pi G\rho b^2 hm}{s} \\ &= -\frac{GMm}{s} \end{aligned}$$

This was all under the assumption that mass m was on the outside of the shell. To complete the proof, we consider the case where it's inside. In this case, the only change is that the limits of integration are different:

$$\begin{aligned} U &= -\frac{2\pi G\rho bhm}{s}\int_{b-s}^{b+s} dr \\ &= -4\pi G\rho bhm \\ &= -\frac{GMm}{b} \end{aligned}$$

The two results are equal at the surface of the sphere, $s = b$, so the constant-energy part joins continuously onto the $1/s$ part, and the effect is to chop off the steepest part of the graph that we would have had if the whole mass M had been concentrated at its center. Dropping a mass m from A to B in figure k releases the same amount of energy as if mass M had been concentrated at its center, but there is no release of gravitational energy at all when moving between two interior points like C and D. In other words, the internal gravitational field is zero. Moving from C to D brings mass m farther away from the nearby side of the shell, but closer to the far side, and the cancellation between these two effects turns out to be perfect. Although the gravitational field has to be zero at the center due to symmetry, it's much more surprising that it cancels out perfectly in the whole interior region; this is a special mathematical characteristic of a $1/r$ interaction like gravity.

Example 17: Newton's apple

Over a period of 27.3 days, the moon travels the circumference of its orbit, so using data from Appendix 5, we can calculate its speed, and solve the circular orbit condition to determine the strength of the earth's gravitational field at the moon's distance from the earth, $g = v^2/r = 2.72 \times 10^{-3}\ \mathrm{m/s^2}$, which is 3600 times smaller than the gravitational field at the earth's surface. The center-to-center distance from the moon to the earth is 60 times greater than the radius of the earth. The earth is, to a very good approximation, a sphere made up of concentric shells, each with uniform density, so the shell theorem tells us that its external gravitational field is the same as if all its mass was concentrated at its center. We already know that a gravitational energy that varies as $-1/r$ is equivalent to a gravitational field proportional to $1/r^2$, so it makes sense that a distance that is greater by a factor of 60 corresponds to a gravitational field that is 60×60=3600 times weaker. Note that the calculation didn't require knowledge of the earth's mass or the gravitational constant, which Newton didn't know.

In 1665, shortly after Newton graduated from Cambridge, the Great Plague forced the college to close for two years, and Newton returned to the family farm and worked intensely on scientific problems. During this productive period, he carried out this calculation, but it came out wrong, causing him to doubt his new theory of gravity. The problem was that during the plague years, he was unable to use the university's library, so he had to use a figure for the radius of the moon's orbit that he had memorized, and he forgot that the memorized value was in units of nautical miles rather than statute miles. Once he realized his mistake, he found that the calculation came out just right, and became confident that his theory was right after all. [9]

Example 18: Weighing the earth

▷ Once Cavendish had found $G = 6.67 \times 10^{-11}\ \mathrm{J \cdot m/kg^2}$ (p. 101, example 15), it became possible to determine the mass of the earth. By the shell theorem, the gravitational energy of a mass m at a distance r from the center of the earth is $U = -GMm/r$, where M is the mass of the earth. The gravitational field is related to this by $mgdr = dU$, or $g = (1/m)dU/dr = GM/r^2$. Solving for M, we have

$$\begin{aligned} M &= gr^2/G \\ &= \frac{(9.8\ \mathrm{m/s^2})(6.4 \times 10^6\ \mathrm{m})^2}{6.67 \times 10^{-11}\ \mathrm{J \cdot m/kg^2}} \\ &= 6.0 \times 10^{24}\ \frac{\mathrm{m^2 \cdot kg^2}}{\mathrm{J \cdot s^2}} \\ &= 6.0 \times 10^{24}\ \mathrm{kg} \end{aligned}$$

Example 19: Gravity inside the earth

▷ The earth is somewhat more dense at greater depths, but as an approximation let's assume it has a constant density throughout. How does its internal gravitational field vary with the distance r from the center?

▷ Let's write b for the radius of the earth. The shell theorem tell us that at a given location r, we only need to consider the mass $M_{<r}$ that is deeper than r. Under the assumption of constant density, this mass is related to the total mass of the earth by

$$\frac{M_{<r}}{M} = \frac{r^3}{\mathrm{b}^3},$$

and by the same reasoning as in example 18,

$$g = \frac{GM_{<r}}{r^2},$$

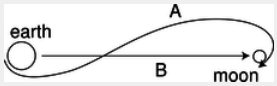
so

$$g = \frac{GMr}{b^3}.$$

In other words, the gravitational field interpolates linearly between zero at $r = 0$ and its ordinary surface value at $r = b$.

The following example applies the numerical techniques of section 2.2.

Example 20: From the earth to the moon



l / The actual trajectory of the Apollo 11 spacecraft, A, and the straight-line trajectory, B, assumed in the example.

The Apollo 11 mission landed the first humans on the moon in 1969. In this example, we'll estimate the time it took to get to the moon, and compare our estimate with the actual time, which was 73.0708 hours from the engine burn that took the ship out of earth orbit to the engine burn that inserted it into lunar orbit. During this time, the ship was coasting with the engines off, except for a small course-correction burn, which we neglect. More importantly, we do the calculation for a straight-line trajectory rather than the real S-shaped one, so the result can only be expected to agree roughly with what really happened. The following data come from the original press kit, which NASA has scanned and posted on the Web:

initial altitude	3.363 X 10^5 m
initial velocity	1.083 X 10^4 m/s

The endpoint of the the straight-line trajectory is a free-fall impact on the lunar surface, which is also unrealistic (luckily for the astronauts).

The ship's energy is

$$E = -\frac{GM_e m}{r} - \frac{GM_m m}{r_m - r} + \frac{1}{2}mv^2,$$

but since everything is proportional to the mass of the ship, m, we can divide it out

$$\frac{E}{m} = -\frac{GM_e}{r} - \frac{GM_m}{r_m - r} + \frac{1}{2}v^2,$$

and the energy variables in the program with names like e, k, and u are actually energies per unit mass. The program is a straightforward modification of the function time3 on page 93.

```
import math
def tmoon(vi,ri,rf,n):
  bigg=6.67e-11        # gravitational constant
  me=5.97e24         # mass of earth
  mm=7.35e22         # mass of moon
  rm=3.84e8        # earth-moon distance
  r=ri
  v=vi
  dr = (rf-ri)/n
  e=-bigg*me/ri-bigg*mm/(rm-ri)+.5*vi**2
  t=0
  for i in range(n):
    u_old = -bigg*me/r-bigg*mm/(rm-r)
    k_old = e - u_old
    v_old = math.sqrt(2.*k_old)
    r = r+dr
    u = -bigg*me/r-bigg*mm/(rm-r)
    k = e - u
    v = math.sqrt(2.*k)
    v_avg = .5*(v_old+v)
    dt=dr/v_avg
    t=t+dt
  return t
```

```
>>> re=6.378e6  # radius of earth
>>> rm=1.74e6   # radius of moon
>>> ri=re+3.363e5       # re+initial altitude
>>> rf=3.8e8-rm # earth-moon distance minus rm
>>> vi=1.083e4 # initial velocity
>>> print(tmoon(vi,ri,rf,1000)/3600.) # convert seconds to hours
59.654047441976552
```

This is pretty decent agreement, considering the wildly inaccurate trajectory assumed. It's interesting to see how much the duration of the trip changes if we increase the initial velocity by only ten percent:

```
>>> vi=1.2e4
>>> print(tmoon(vi,ri,rf,1000)/3600.)
18.177752636111677
```

The most important reason for using the lower speed was that if something had gone wrong, the ship would have been able to whip around the moon and take a “free return” trajectory back to the earth, without having to do any further burns. At a higher speed, the ship would have had so much kinetic energy that in the absence of any further engine burns, it would have escaped from the earth-moon system. The Apollo 13 mission had to take a free return trajectory after an explosion crippled the spacecraft.

Contributors and Attributions

3.4: Atomic Phenomena

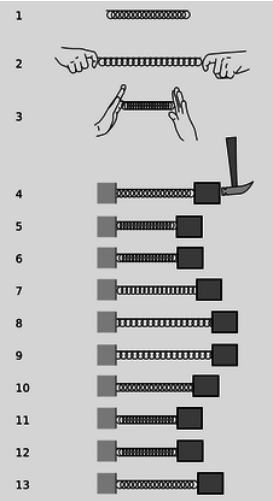
Variety is the spice of life, not of science. So far this chapter has focused on heat energy, kinetic energy, and gravitational energy, but it might seem that in addition to these there is a bewildering array of other forms of energy. Gasoline, chocolate bars, batteries, melting water --- in each case there seems to be a whole new type of energy. The physicist's psyche rebels against the prospect of a long laundry list of types of energy, each of which would require its own equations, concepts, notation, and terminology. The point at which we've arrived in the study of energy is analogous to the period in the 1960's when a half a dozen new subatomic particles were being discovered every year in particle accelerators. It was an embarrassment. Physicists began to speak of the “particle zoo,” and it seemed that the subatomic world was distressingly complex. The particle zoo was simplified by the realization that most of the new particles being whipped up were simply clusters of a previously unsuspected set of fundamental particles (which were whimsically dubbed quarks, a made-up word from a line of poetry by James Joyce, “Three quarks for Master Mark.”) The energy zoo can also be simplified, and it's the purpose of this section to demonstrate the hidden similarities between forms of energy as seemingly different as heat and motion.

Contributors and Attributions

Benjamin Crowell (Fullerton College).

3.5: Oscillations

Let's revisit the example of the stretched spring from the previous section. We know that its energy is a form of electrical energy of interacting atoms, which is nice conceptually but doesn't help us to solve problems, since we don't know how the energy, *[Math Processing Error]*, depends on the length of the spring. All we know is that there's an equilibrium (figure a/1), which is a local minimum of the function *[Math Processing Error]*. An extremely important problem which arises in this connection is how to calculate oscillatory motion around an equilibrium, as in a/4-13. Even if we did special experiments to find out how the spring's energy worked, it might seem like we'd have to go through just as much work to deal with any other kind of oscillation, such as a sapling swinging back and forth in the breeze.



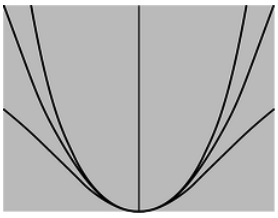
a / The spring has a minimum-energy length, 1, and energy is required in order to compress or stretch it, 2 and 3. A mass attached to the spring will oscillate around the equilibrium, 4-13.

Surprisingly, it's possible to analyze this type of oscillation in a very general and elegant manner, as long as the analysis is limited to *small* oscillations. We'll talk about the mass on the spring for concreteness, but there will be nothing in the discussion at all that is restricted to that particular physical system. First, let's choose a coordinate system in which *[Math Processing Error]* corresponds to the position of the mass where the spring is in equilibrium, and since interaction energies like *[Math Processing Error]* are only well defined up to an additive constant, we'll simply define it to be zero at equilibrium:

[Math Processing Error]

Since *[Math Processing Error]* is an equilibrium, *[Math Processing Error]* must have a local minimum there, and a differentiable function (which we assume *[Math Processing Error]* is) has a zero derivative at a local minimum:

[Math Processing Error]



b / Three functions with the same curvature at *[Math Processing Error]*=0.

There are still infinitely many functions that could satisfy these criteria, including the three shown in figure b, which are *[Math Processing Error]*, *[Math Processing Error]*, and *[Math Processing Error]*. Note, however, how all three functions are virtually identical right near the minimum. That's because they all have the same curvature. More specifically, each function has its second derivative equal to 1 at *[Math Processing Error]*, and the second derivative is a measure of curvature. We write *[Math Processing Error]* for the second derivative of the energy at an equilibrium point,

[Math Processing Error]

Physically, *[Math Processing Error]* is a measure of stiffness. For example, the heavy-duty springs in a car's shock absorbers would have a high value of *[Math Processing Error]*. It is often referred to as the spring constant, but we're only using a spring as an example here. As shown in figure b, any two functions that have *[Math Processing Error]*, *[Math Processing Error]*, and *[Math Processing Error]*, with the same value of *[Math Processing Error]*, are virtually indistinguishable for small values of *[Math Processing Error]*, so if we want to analyze small oscillations, it doesn't even matter which function we assume. For simplicity, we'll just use *[Math Processing Error]* from now on.

Now we're ready to analyze the mass-on-a-spring system, while keeping in mind that it's really only a representative example of a whole class of similar oscillating systems. We expect that the motion is going to repeat itself over and over again, and since we're not going to include frictional heating in our model, that repetition should go on forever without dying out. The most interesting thing to know about the motion would be the period, *[Math Processing Error]*, which is the amount of time required for one complete cycle of the motion. We might expect that the period would depend on the spring constant, *[Math Processing Error]*, the mass, *[Math Processing Error]*, and and the amplitude, *[Math Processing Error]*, defined in figure c.[11]

c / The amplitude would usually be defined as the distance from equilibrium to one extreme of the motion, i.e., half the total travel.

In examples like the brachistochrone and the Apollo 11 mission, it was generally necessary to use numerical techniques to determine the amount of time required for a certain motion. Once again, let's dust off the time3 function from page 93 and modify it for our purposes. For flexibility, we'll define the function *[Math Processing Error]* as a separate Python function. We really want to calculate the time required for the mass to come back to its starting point, but that would be awkward to set up, since our function works by dividing up the distance to be traveled into tiny segments. By symmetry, the time required to go from one end to the other equals the time required to come back to the start, so we'll just calculate the time for half a cycle and then double it when we return the result at the end of the function. The test at lines 16-19 is necessary because otherwise at the very end of the motion we can end up trying to take the square root of a negative number due to rounding errors.

```
import math
def u(k,x):
  return .5*k*x**2

def osc(m,k,a,n):
  x=a
  v=0
  dx = -2.*a/n
  t=0
  e = u(k,x)+.5*m*v**2
  for i in range(n):
    x_old = x
    v_old = v
    x = x+dx
    kinetic = e-u(k,x)
    if kinetic<0. :
      v=0.
      print "warning, K=",kinetic,"<0"
    else :
      v = -math.sqrt(2.*kinetic/m)
    v_avg = (v+v_old)/2.
    dt=dx/v_avg
    t=t+dt
  return 2.*t
```

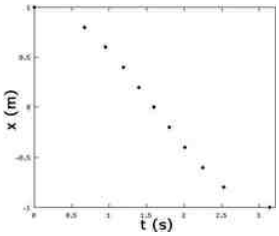
```
>>> print(osc(1.,1.,1.,100000))
warning, K= -1.43707268307e-12 <0
6.2831854132667919
```

The first thing to notice is that with this particular set of inputs (*[Math Processing Error]*=1 kg, *[Math Processing Error]*, and *[Math Processing Error]*), the program has done an excellent job of computing *[Math Processing Error]*. This is Mother Nature giving us a strong hint that the problem has an algebraic solution, not just a numerical one. The next interesting thing happens when we change the amplitude from 1 m to 2 m:

```
>>> print(osc(1.,1.,2.,100000))
warning, K= -5.7482907323e-12 <0
6.2831854132667919
```

Even though the mass had to travel double the distance in each direction, the period is the same to within the numerical accuracy of the calculation!

With these hints, it seems like we should start looking for an algebraic solution. For guidance, here's a graph of *[Math Processing Error]* as a function of *[Math Processing Error]*, as calculated by the osc function with n=10.

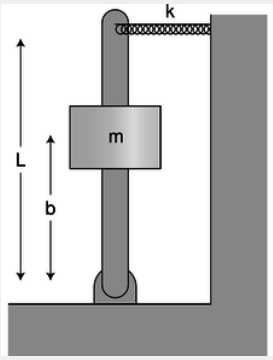


This looks like a cosine function, so let's see if a *[Math Processing Error]* is a solution to the conservation of energy equation --- it's not uncommon to try to “reverse-engineer” the cryptic results of a numerical calculation like this. The symbol *[Math Processing Error]* (Greek omega), called angular frequency, is a standard symbol for the number of radians per second of oscillation. Except for the factor of *[Math Processing Error]*, it is identical to the ordinary frequency *[Math Processing Error]*, which has units of *[Math Processing Error]* or Hz (Hertz). The phase angle *[Math Processing Error]* is to allow for the possibility that *[Math Processing Error]* doesn't coincide with the beginning of the motion. The energy is

[Math Processing Error]

According to conservation of energy, this has to be a constant. Using the identity *[Math Processing Error]*, we can see that it will be a constant if we have *[Math Processing Error]*, or *[Math Processing Error]*, i.e., *[Math Processing Error]*. Note that the period is independent of amplitude.

Example 23: A spring and a lever



d / Example 23. The rod pivots on the hinge at the bottom.

[Math Processing Error] What is the period of small oscillations of the system shown in the figure? Neglect the mass of the lever and the spring. Assume that the spring is so stiff that gravity is not an important effect. The spring is relaxed when the lever is vertical.

[Math Processing Error] This is a little tricky, because the spring constant *[Math Processing Error]*, although it is relevant, is *not* the *[Math Processing Error]* we should be putting into the equation *[Math Processing Error]*. The *[Math Processing Error]* that goes in there has to be the second derivative of *[Math Processing Error]* with respect to the position, *[Math Processing Error]*, of the mass that's moving. The energy *[Math Processing Error]* stored in the spring depends on how far the *tip* of the lever is from the center. This distance equals *[Math Processing Error]*, so the energy in the spring is

[Math Processing Error]

and the *[Math Processing Error]* we have to put in *[Math Processing Error]* is

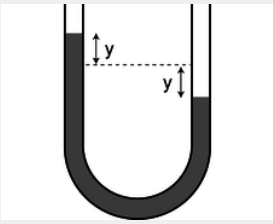
[Math Processing Error]

The result is

[Math Processing Error]

The leverage of the lever makes it as if the spring was stronger, and decreases the period of the oscillations by a factor of *[Math Processing Error]*.

Example 24: Water in a U-shaped tube



e / Water in a U-shaped tube.

[Math Processing Error] What is the period of oscillation of the water in figure e?

[Math Processing Error] In example 13 on p. 89, we found *[Math Processing Error]*, so the “spring constant,” which really isn't a spring constant here at all, is

[Math Processing Error]

This is an interesting example, because *[Math Processing Error]* can be calculated without any approximations, but the kinetic energy requires an approximation, because we don't know the details of the pattern of flow of the water. It could be very complicated. There will be a tendency for the water near the walls to flow more slowly due to friction, and there may also be swirling, turbulent motion. However, if we make the approximation that all the water moves with the same velocity as the surface, *[Math Processing Error]*, then the mass-on-a-spring analysis applies. Letting *[Math Processing Error]* be the total length of the filled part of the tube, the mass is *[Math Processing Error]*, and we have

[Math Processing Error]

Contributors and Attributions

3.6: Footnotes

1. An entertaining account of this form of quackery is given in **Voodoo Science: The Road from Foolishness to Fraud**, Robert Park, Oxford University Press, 2000. Until reading this book, I hadn't realized the degree to which pseudoscience had penetrated otherwise respectable scientific organizations like NASA.
2. Although the definition refers to the Celsius scale of temperature, it's not necessary to give an operational definition of the temperature concept in general (which turns out to be quite a tricky thing to do completely rigorously); we only need to establish two specific temperatures that can be reproduced on thermometers that have been calibrated in a standard way. Heat and temperature are discussed in more detail in section 2.4, and in chapter 5. Conceptually, heat is a measure of energy, whereas temperature relates to how concentrated that energy is.
3. It's not at all obvious that the solution would work out in the earth's frame of reference, although Galilean relativity states that it doesn't matter which frame we use. Chapter 3 discusses the relationship between conservation of energy and Galilean relativity.
4. From Joule's point of view, the point of the experiment was different. At that time, most physicists believed that heat was a quantity that was conserved separately from the rest of the things to which we now refer as energy, i.e., mechanical\ energy. Separate units of measurement had been constructed for heat and mechanical of energy, but Joule was trying to show that one could convert back and forth between them, and that it was actually their sum that was conserved, if they were both expressed in consistent units. His main result was the conversion factor that would allow the two sets of units to be reconciled. By showing that the conversion factor came out the same in different types of experiments, he was supporting his assertion that heat was not separately conserved. From Joule's perspective or from ours, the result is to connect the mysterious, invisible phenomenon of heat with forms of energy that are visible properties of objects, i.e., mechanical energy.
5. If you've had a previous course in physics, you may have seen this presented not as an empirical result but as a theoretical one, derived from Newton's laws, and in that case you might feel you're being cheated here. However, I'm going to reverse that reasoning and derive Newton's laws from the conservation laws in chapter 3. From the modern perspective, conservation laws are more fundamental, because they apply in cases where Newton's laws don't.
6. Système International
7. There is a mathematical loophole in this argument that would allow the object to hover for a while with zero velocity and zero acceleration. This point is discussed on page 910.
8. There is a hidden assumption here, which is that the sun doesn't move. Actually the sun wobbles a little because of the planets' gravitational interactions with it, but the wobble is small due to the sun's large mass, so it's a pretty good approximation to assume the sun is stationary. Chapter 3 provides the tools to analyze this sort of thing completely correctly --- see p. 142.
9. Some historians are suspicious that the story of the apple and the mistake in conversions may have been fabricated by Newton later in life. The conversion incident may have been a way of explaining his long delay in publishing his work, which led to a conflict with Leibniz over priority in the invention of calculus.
10. Subsection 6.1.5 presents some evidence for the Big Bang theory.
11. Many kinds of oscillations are possible, so there is no standard definition of the amplitude. For a pendulum, the natural definition would be in terms of an angle. For a radio transmitter, we'd use some kind of electrical units.

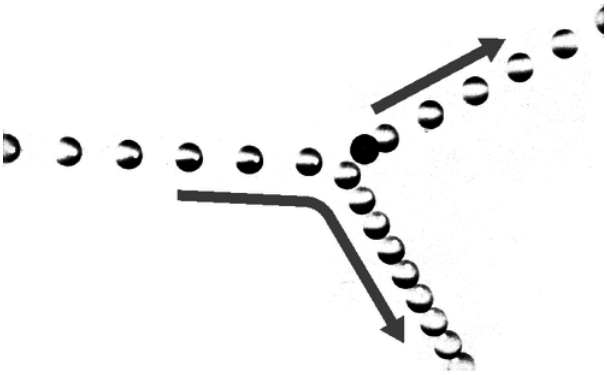
3.7: Problems

1. Experiments show that the power consumed by a boat's engine is approximately proportional to the third power of its speed. (We assume that it is moving at constant speed.)

(a) When a boat is cruising at constant speed, what type of energy transformation do you think is being performed?
(b) If you upgrade to a motor with double the power, by what factor is your boat's maximum cruising speed increased? (solution in the pdf version of the book)

2. Object A has a kinetic energy of 13.4 J. Object B has a mass that is greater by a factor of 3.77, but is moving more slowly by a factor of 2.34. What is object B's kinetic energy? (solution in the pdf version of the book)

3. My 1.25 kW microwave oven takes 126 seconds to bring 250 g of water from room temperature to a boil. What percentage of the power is being wasted? Where might the rest of the energy be going? (solution in the pdf version of the book)



a / Problem 4.

4. The multiflash photograph shows a collision between two pool balls. The ball that was initially at rest shows up as a dark image in its initial position, because its image was exposed several times before it was struck and began moving. By making *measurements* on the figure, determine *numerically* whether or not energy appears to have been conserved in the collision. What systematic effects would limit the accuracy of your test? [From an example in PSSC Physics.]

5. A grasshopper with a mass of 110 mg falls from rest from a height of 310 cm. On the way down, it dissipates 1.1 mJ of heat due to air resistance. At what speed, in m/s, does it hit the ground? (solution in the pdf version of the book)

6. A ball rolls up a ramp, turns around, and comes back down. When does it have the greatest gravitational energy? The greatest kinetic energy? [Based on a problem by Serway and Faughn.]

7. (a) You release a magnet on a tabletop near a big piece of iron, and the magnet leaps across the table to the iron. Does the magnetic energy increase, or decrease? Explain. (b) Suppose instead that you have two repelling magnets. You give them an initial push towards each other, so they decelerate while approaching each other. Does the magnetic energy increase, or decrease? Explain.

8. Estimate the kinetic energy of an Olympic sprinter.

9. (answer check available at lightandmatter.com) You are driving your car, and you hit a brick wall head on, at full speed. The car has a mass of 1500 kg. The kinetic energy released is a measure of how much destruction will be done to the car and to your body. Calculate the energy released if you are traveling at (a) 40 mi/hr, and again (b) if you're going 80 mi/hr. What is counterintuitive about this, and what implication does this have for driving at high speeds?

10. A closed system can be a bad thing --- for an astronaut sealed inside a space suit, getting rid of body heat can be difficult. Suppose a 60-kg astronaut is performing vigorous physical activity, expending 200 W of power. If none of the heat can escape from her space suit, how long will it take before her body temperature rises by 6°{}C (11°{}F), an amount sufficient to kill her? Assume that the amount of heat required to raise her body temperature by 1°{}C is the same as it would be for an equal mass of water. Express your answer in units of minutes. (answer check available at lightandmatter.com)

11. The following table gives the amount of energy required in order to heat, melt, or boil a gram of water.

heat 1 g of ice by 1°C	2.05 J
melt 1 g of ice	333 J
heat 1 g of liquid by 1°C	4.19 J
boil 1 g of water	2500 J
heat 1 g of steam by 1°C	2.01 J

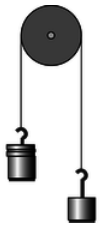
(a) How much energy is required in order to convert 1.00 g of ice at -20 *[Math Processing Error]* into steam at 137 *[Math Processing Error]*? (answer check available at lightandmatter.com)
(b) What is the minimum amount of hot water that could melt 1.00 g of ice? (answer check available at lightandmatter.com)

12. Anya climbs to the top of a tree, while Ivan climbs half-way to the top. They both drop pennies to the ground. Compare the kinetic energies and velocities of the pennies on impact, using ratios.

13. Anya and Ivan lean over a balcony side by side. Anya throws a penny downward with an initial speed of 5 m/s. Ivan throws a penny upward with the same speed. Both pennies end up on the ground below. Compare their kinetic energies and velocities on impact.

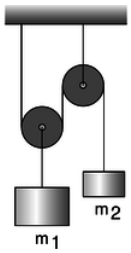
14. (a) A circular hoop of mass *[Math Processing Error]* and radius *[Math Processing Error]* spins like a wheel while its center remains at rest. Let *[Math Processing Error]* (Greek letter omega) be the number of radians it covers per unit time, i.e., *[Math Processing Error]*, where the period, *[Math Processing Error]*, is the time for one revolution. Show that its kinetic energy equals *[Math Processing Error]*.
(b) Show that the answer to part a has the right units. (Note that radians aren't really units, since the definition of a radian is a unitless ratio of two lengths.)
(c) If such a hoop rolls with its center moving at velocity *[Math Processing Error]*, its kinetic energy equals *[Math Processing Error]*, plus the amount of kinetic energy found in part a. Show that a hoop rolls down an inclined plane with half the acceleration that a frictionless sliding block would have.

15. On page 83, I used the chain rule to prove that the acceleration of a free-falling object is given by *[Math Processing Error]*. In this problem, you'll use a different technique to prove the same thing. Assume that the acceleration is a constant, *[Math Processing Error]*, and then integrate to find *[Math Processing Error]* and *[Math Processing Error]*, including appropriate constants of integration. Plug your expressions for *[Math Processing Error]* and *[Math Processing Error]* into the equation for the total energy, and show that *[Math Processing Error]* is the only value that results in constant energy.



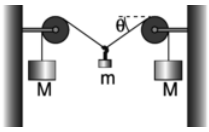
b / Problem 16.

16. The figure shows two unequal masses, *[Math Processing Error]* and *[Math Processing Error]*, connected by a string running over a pulley. Find the acceleration. \hwhint{hwhint:atwood-energy-sn}(answer check available at lightandmatter.com)



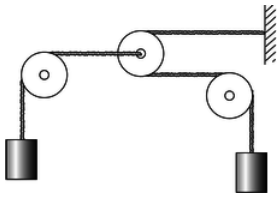
c / Problem 17.

17. What ratio of masses will balance the pulley system shown in the figure? \hwhint{hwhint:pulley}



d / Problem 18.

18. (a) For the apparatus shown in the figure, find the equilibrium angle *[Math Processing Error]* in terms of the two masses. (answer check available at lightandmatter.com)
(b) Interpret your result in the case of *[Math Processing Error]* (*[Math Processing Error]* much greater than *[Math Processing Error]*). Does it make sense physically?
(c) For what combinations of masses would your result give nonsense? Interpret this physically. \hwhint{hwhint:tightropish}



e / Problem 19.

19. In the system shown in the figure, the pulleys on the left and right are fixed, but the pulley in the center can move to the left or right. The two hanging masses are identical, and the pulleys and ropes are all massless. Find the upward acceleration of the mass on the left, in terms of *[Math Processing Error]* only. \hwhint{hwhint:funkyatwood}(answer check available at lightandmatter.com)

20. Two atoms will interact via electrical forces between their protons and electrons. One fairly good approximation to the electrical energy is the Lennard-Jones formula,

[Math Processing Error]

where *[Math Processing Error]* is the center-to-center distance between the atoms and *[Math Processing Error]* is a positive constant. Show that (a) there is an equilibrium point at *[Math Processing Error]*,
(b) the equilibrium is stable, and
(c) the energy required to bring the atoms from their equilibrium separation to infinity is *[Math Processing Error]*. \hwhint{hwhint:lennardjones}

21. The International Space Station orbits at an altitude of about 360 to 400 km. What is the gravitational field of the earth at this altitude?(answer check available at lightandmatter.com)

22. (a) A geosynchronous orbit is one in which the satellite orbits above the equator, and has an orbital period of 24 hours, so that it is always above the same point on the spinning earth. Calculate the altitude of such a satellite.(answer check available at lightandmatter.com)
(b) What is the gravitational field experienced by the satellite? Give your answer as a percentage in relation to the gravitational field at the earth's surface.\hwhint{hwhint:geosynch}(answer check available at lightandmatter.com)

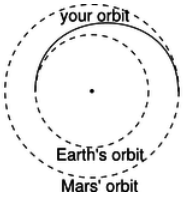
23. Astronomers calculating orbits of planets often work in a nonmetric system of units, in which the unit of time is the year, the unit of mass is the sun's mass, and the unit of distance is the astronomical unit (A.U.), defined as half the long axis of the earth's orbit. In these units, find an exact expression for the gravitational constant, *[Math Processing Error]*.(answer check available at lightandmatter.com)

24. The star Lalande 21185 was found in 1996 to have two planets in roughly circular orbits, with periods of 6 and 30 years. What is the ratio of the two planets' orbital radii?(answer check available at lightandmatter.com)

25. A projectile is moving directly away from a planet of mass *[Math Processing Error]* at exactly escape velocity. (a) Find *[Math Processing Error]*, the distance from the projectile to the center of the planet, as a function of time, *[Math Processing Error]*, and also find *[Math Processing Error]*.(answer check available at lightandmatter.com)
(b) Check the units of your answer.

(c) Does *[Math Processing Error]* show the correct behavior as *[Math Processing Error]* approaches infinity? \hwhint{hwhint:escape2}

26. The purpose of this problem is to estimate the height of the tides. The main reason for the tides is the moon's gravity, and we'll neglect the effect of the sun. Also, real tides are heavily influenced by landforms that channel the flow of water, but we'll think of the earth as if it was completely covered with oceans. Under these assumptions, the ocean surface should be a surface of constant *[Math Processing Error]*. That is, a thimbleful of water, *[Math Processing Error]*, should not be able to gain or lose any gravitational energy by moving from one point on the ocean surface to another. If only the spherical earth's gravity was present, then we'd have *[Math Processing Error]*, and a surface of constant *[Math Processing Error]* would be a surface of constant *[Math Processing Error]*, i.e., the ocean's surface would be spherical. Taking into account the moon's gravity, the main effect is to shift the center of the sphere, but the sphere also becomes slightly distorted into an approximately ellipsoidal shape. (The shift of the center is not physically related to the tides, since the solid part of the earth tends to be centered within the oceans; really, this effect has to do with the motion of the whole earth through space, and the way that it wobbles due to the moon's gravity.) Determine the amount by which the long axis of the ellipsoid exceeds the short axis. \hwhint{hwhint:tides}



f / Problem 27.

27. You are considering going on a space voyage to Mars, in which your route would be half an ellipse, tangent to the Earth's orbit at one end and tangent to Mars' orbit at the other. Your spacecraft's engines will only be used at the beginning and end, not during the voyage. How long would the outward leg of your trip last? (The orbits of Earth and Mars are nearly circular, and Mars's is bigger by a factor of 1.52.) (answer check available at lightandmatter.com)

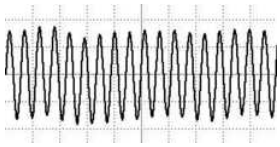
28. When you buy a helium-filled balloon, the seller has to inflate it from a large metal cylinder of the compressed gas. The helium inside the cylinder has energy, as can be demonstrated for example by releasing a little of it into the air: you hear a hissing sound, and that sound energy must have come from somewhere. The total amount of energy in the cylinder is very large, and if the valve is inadvertently damaged or broken off, the cylinder can behave like bomb or a rocket.

Suppose the company that puts the gas in the cylinders prepares cylinder A with half the normal amount of pure helium, and cylinder B with the normal amount. Cylinder B has twice as much energy, and yet the temperatures of both cylinders are the same. Explain, at the atomic level, what form of energy is involved, and why cylinder B has twice as much.

29. Explain in terms of conservation of energy why sweating cools your body, even though the sweat is at the same temperature as your body. Describe the forms of energy involved in this energy transformation. Why don't you get the same cooling effect if you wipe the sweat off with a towel? Hint: The sweat is evaporating.

30. [This problem is now problem 3-73.]

31. All stars, including our sun, show variations in their light output to some degree. Some stars vary their brightness by a factor of two or even more, but our sun has remained relatively steady during the hundred years or so that accurate data have been collected. Nevertheless, it is possible that climate variations such as ice ages are related to long-term irregularities in the sun's light output. If the sun was to increase its light output even slightly, it could melt enough Antarctic ice to flood all the world's coastal cities. The total sunlight that falls on Antarctica amounts to about *[Math Processing Error]* watts. In the absence of natural or human-caused climate change, this heat input to the poles is balanced by the loss of heat via winds, ocean currents, and emission of infrared light, so that there is no net melting or freezing of ice at the poles from year to year. Suppose that the sun changes its light output by some small percentage, but there is no change in the rate of heat loss by the polar caps. Estimate the percentage by which the sun's light output would have to increase in order to melt enough ice to raise the level of the oceans by 10 meters over a period of 10 years. (This would be enough to flood New York, London, and many other cities.) Melting 1 kg of ice requires *[Math Processing Error]* J.

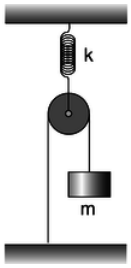


g / Problem 32.

32. The figure shows the oscillation of a microphone in response to the author whistling the musical note “A.” The horizontal axis, representing time, has a scale of 1.0 ms per square. Find the period *[Math Processing Error]*, the frequency *[Math Processing Error]*, and the angular frequency *[Math Processing Error]*.(answer check available at lightandmatter.com)

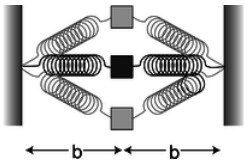
33. (a) A mass *[Math Processing Error]* is hung from a spring whose spring constant is *[Math Processing Error]*. Write down an expression for the total interaction energy of the system, *[Math Processing Error]*, and find its equilibrium position.\hwhint{hwhint:hangfromspring}
(b) Explain how you could use your result from part a to determine an unknown spring constant.

34. A certain mass, when hung from a certain spring, causes the spring to stretch by an amount *[Math Processing Error]* compared to its equilibrium length. If the mass is displaced vertically from this equilibrium, it will oscillate up and down with a period *[Math Processing Error]*. Give a numerical comparison between *[Math Processing Error]* and *[Math Processing Error]*, the time required for the mass to fall from rest through a height *[Math Processing Error]*, when it isn't attached to the spring. (You will need the result of problem 33).(answer check available at lightandmatter.com)



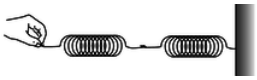
h / Problem 35.

35. Find the period of vertical oscillations of the mass *[Math Processing Error]*. The spring, pulley, and ropes have negligible mass.\hwhint{hwhint:pulleyandspring}(answer check available at lightandmatter.com)



i / Problem 36.

36. The equilibrium length of each spring in the figure is *[Math Processing Error]*, so when the mass *[Math Processing Error]* is at the center, neither spring exerts any force on it. When the mass is displaced to the side, the springs stretch; their spring constants are both *[Math Processing Error]*.
(a) Find the energy, *[Math Processing Error]*, stored in the springs, as a function of *[Math Processing Error]*, the distance of the mass up or down from the center.(answer check available at lightandmatter.com)
(b) Show that the period of small up-down oscillations is infinite.



j / Problem 37.

37. Two springs with spring constants *[Math Processing Error]* and *[Math Processing Error]* are put together end-to-end. Let *[Math Processing Error]* be the amount by which the first spring is stretched relative to its equilibrium length, and similarly for *[Math Processing Error]*. If the combined double spring is stretched by an amount *[Math Processing Error]* relative to *its* equilibrium length, then *[Math Processing Error]*. Find the spring constant, *[Math Processing Error]*, of the combined spring in

terms of *[Math Processing Error]* and *[Math Processing Error]*. \hwhint{hwhint:springsseries}\hwans{hwans:springsseries} (answer check available at lightandmatter.com)

38. A mass *[Math Processing Error]* on a spring oscillates around an equilibrium at *[Math Processing Error]*. Any function *[Math Processing Error]* with an equilibrium at *[Math Processing Error]* can be approximated as *[Math Processing Error]*, and if the energy is symmetric with respect to positive and negative values of *[Math Processing Error]*, then the next level of improvement in such an approximation would be *[Math Processing Error]*. The general idea here is that any smooth function can be approximated locally by a polynomial, and if you want a better approximation, you can use a polynomial with more terms in it. When you ask your calculator to calculate a function like *[Math Processing Error]* or *[Math Processing Error]*, it's using a polynomial approximation with 10 or 12 terms. Physically, a spring with a positive value of *[Math Processing Error]* gets stiffer when stretched strongly than an “ideal” spring with *[Math Processing Error]*. A spring with a negative *[Math Processing Error]* is like a person who cracks under stress --- when you stretch it too much, it becomes more elastic than an ideal spring would. We should not expect any spring to give totally ideal behavior no matter no matter how much it is stretched. For example, there has to be some point at which it breaks.

Do a numerical simulation of the oscillation of a mass on a spring whose energy has a nonvanishing *[Math Processing Error]*. Is the period still independent of amplitude? Is the amplitude-independent equation for the period still approximately valid for small enough amplitudes? Does the addition of a positive *[Math Processing Error]* term tend to increase the period, or decrease it? Include a printout of your program and its output with your homework paper.

39. An idealized pendulum consists of a pointlike mass *[Math Processing Error]* on the end of a massless, rigid rod of length *[Math Processing Error]*. Its amplitude, *[Math Processing Error]*, is the angle the rod makes with the vertical when the pendulum is at the end of its swing. Write a numerical simulation to determine the period of the pendulum for any combination of *[Math Processing Error]*, *[Math Processing Error]*, and *[Math Processing Error]*. Examine the effect of changing each variable while manipulating the others.

40. A ball falls from a height *[Math Processing Error]*. Without air resistance, the time it takes to reach the floor is *[Math Processing Error]*. A numerical version of this calculation was given in program time2 on page 92. Now suppose that air resistance is not negligible. For a smooth sphere of radius *[Math Processing Error]*, moving at speed *[Math Processing Error]* through air of density *[Math Processing Error]*, the amount of energy *[Math Processing Error]* dissipated as heat as the ball falls through a height *[Math Processing Error]* is given (ignoring signs) by *[Math Processing Error]*. Modify the program to incorporate this effect, and find the resulting change in the fall time in the case of a 21 g ball of radius 1.0 cm, falling from a height of 1.0 m. The density of air at sea level is about *[Math Processing Error]*. Turn in a printout of both your program and its output. Answer: 0.34 ms.

41. The factorial of an integer *[Math Processing Error]*, written *[Math Processing Error]*, is defined as the product of all the positive integers less than or equal to *[Math Processing Error]*. For example, *[Math Processing Error]*. Write a Python program to compute the factorial of a number. Test it with a small number whose factorial you can check by hand. Then use it to compute *[Math Processing Error]*. (Python computes integer results with unlimited precision, so you won't get any problems with rounding or overflows.) Turn in a printout of your program and its output, including the test.

42. Estimate the kinetic energy of a buzzing fly's wing.

Contributors and Attributions

CHAPTER OVERVIEW

4: Conservation of Momentum

Forces transfer momentum to the girl.

"I think, therefore I am." I hope that posterity will judge me kindly, not only as to the things which I have explained, but also to those which I have intentionally omitted so as to leave to others the pleasure of discovery. -- *René Descartes*

Contributors and Attributions

- Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

4.1: Momentum In One Dimension

Contributors and Attributions

4.2: Force In One Dimension

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

4.3: Resonance

Resonance is a phenomenon in which an oscillator responds most strongly to a driving force that matches its own natural frequency of vibration. For example, suppose a child is on a playground swing with a natural frequency of 1 Hz. That is, if you pull the child away from equilibrium, release her, and then stop doing anything for a while, she'll oscillate at 1 Hz. If there was no friction, as we assumed in section 2.5, then the sum of her gravitational and kinetic energy would remain constant, and the amplitude would be exactly the same from one oscillation to the next. However, friction is going to convert these forms of energy into heat, so her oscillations would gradually die out. To keep this from happening, you might give her a push once per cycle, i.e., the frequency of your pushes would be 1 Hz, which is the same as the swing's natural frequency. As long as you stay in rhythm, the swing responds quite well. If you start the swing from rest, and then give pushes at 1 Hz, the swing's amplitude rapidly builds up, as in figure a, until after a while it reaches a steady state in which friction removes just as much energy as you put in over the course of one cycle.

self-check:

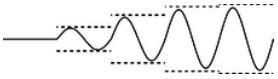


Figure a: An x-versus-t graph for a swing pushed at resonance.

In figure a, compare the amplitude of the cycle immediately following the first push to the amplitude after the second. Compare the energies as well. (answer in the back of the PDF version of the book)

What will happen if you try pushing at 2 Hz? Your first push puts in some momentum, p, but your second push happens after only half a cycle, when the swing is coming right back at you, with momentum $-p$! The momentum transfer from the second push is exactly enough to stop the swing. The result is a very weak, and not very sinusoidal, motion, b.

Figure b: A swing pushed at twice its resonant frequency.

Making the math easy

This is a simple and physically transparent example of resonance: the swing responds most strongly if you match its natural rhythm. However, it has some characteristics that are mathematically ugly and possibly unrealistic. The quick, hard pushes are known as *impulse* forces, c, and they lead to an x-t graph that has nondifferentiable kinks.

Figure c: The F-versus-t graph for an impulsive driving force.

Impulsive forces like this are not only badly behaved mathematically, they are usually undesirable in practical terms. In a car engine, for example, the engineers work very hard to make the force on the pistons change smoothly, to avoid excessive vibration. Throughout the rest of this section, we'll assume a driving force that is sinusoidal, d, i.e., one whose F-t graph is either a sine function or a function that differs from a sine wave in phase, such as a cosine. The force is positive for half of each cycle and negative for the other half, i.e., there is both pushing and pulling. Sinusoidal functions have many nice mathematical characteristics (we can differentiate and integrate them, and the sum of sinusoidal functions that have the same frequency is a sinusoidal function), and they are also used in many practical situations. For instance, my garage door zapper sends out a sinusoidal radio wave, and the receiver is tuned to resonance with it.

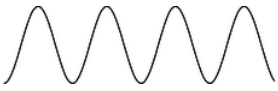


Figure d: A sinusoidal driving force.

A second mathematical issue that I glossed over in the swing example was how friction behaves. In section 3.2.4, about forces between solids, the empirical equation for kinetic friction was independent of velocity. Fluid friction, on the other hand, is velocity-dependent. For a child on a swing, fluid friction is the most important form of friction, and is approximately proportional to v^2. In still other situations, e.g., with a low-density gas or friction between solid surfaces that have been lubricated with a fluid such as oil, we may find that the frictional force has some other dependence on velocity, perhaps being proportional to v, or having some

other complicated velocity dependence that can't even be expressed with a simple equation. It would be extremely complicated to have to treat all of these different possibilities in complete generality, so for the rest of this section, we'll assume friction proportional to velocity

$$F = -bv,$$

simply because the resulting equations happen to be the easiest to solve. Even when the friction doesn't behave in exactly this way, many of our results may still be at least qualitatively correct.

3.3.1 Damped, free motion

Numerical treatment

An oscillator that has friction is referred to as damped. Let's use numerical techniques to find the motion of a damped oscillator that is released away from equilibrium, but experiences no driving force after that. We can expect that the motion will consist of oscillations that gradually die out.

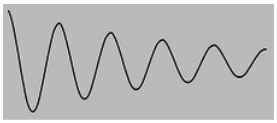


Figure e: A damped sine wave, of the form $x = Ae^{-ct}\sin(\omega_f t + \delta)$.

In section 2.5, we simulated the undamped case using our tried and true Python function based on conservation of energy. Now, however, that approach becomes a little awkward, because it involves splitting up the path to be traveled into n tiny segments, but in the presence of damping, each swing is a little shorter than the last one, and we don't know in advance exactly how far the oscillation will get before turning around. An easier technique here is to use force rather than energy. Newton's second law, $a = F/m$, gives $a = (-kx - bv)/m$, where we've made use of the result of example 40 for the force exerted by the spring. This becomes a little prettier if we rewrite it in the form

$$ma + bv + kx = 0,$$

which gives symmetric treatment to three terms involving x and its first and second derivatives, v and a. Now instead of calculating the time $\Delta t = \Delta x/v$ required to move a predetermined distance Δx, we pick Δt and determine the distance traveled in that time, $\Delta x = v\Delta t$. Also, we can no longer update v based on conservation of energy, since we don't have any easy way to keep track of how much mechanical energy has been changed into heat energy. Instead, we recalculate the velocity using $\Delta v = a\Delta t$.

```
import math
k=39.4784 # chosen to give a period of 1 second
m=1.
b=0.211 # chosen to make the results simple
x=1.
v=0.
t=0.
dt=.01
n=1000
for j in range(n):
  x=x+v*dt
  a=(-k*x-b*v)/m
  if (v>0) and (v+a*dt<0) :
    print("turnaround at t=",t,", x=",x)
  v=v+a*dt
  t=t+dt
```

```
turnaround at t= 0.99 , x= 0.899919262445
turnaround at t= 1.99 , x= 0.809844934046
turnaround at t= 2.99 , x= 0.728777519477
turnaround at t= 3.99 , x= 0.655817260033
turnaround at t= 4.99 , x= 0.590154191135
turnaround at t= 5.99 , x= 0.531059189965
turnaround at t= 6.99 , x= 0.477875914756
turnaround at t= 7.99 , x= 0.430013546991
turnaround at t= 8.99 , x= 0.386940256644
turnaround at t= 9.99 , x= 0.348177318484
```

The spring constant, $k = 4\pi = 39.4784$ N/m, is designed so that if the undamped equation $f = (1/2\pi)\sqrt{k/m}$ was still true, the frequency would be 1 Hz. We start by noting that the addition of a small amount of damping doesn't seem to have changed the period at all, or at least not to within the accuracy of the calculation.[10] You can check for yourself, however, that a large value of b, say 5 $\mathrm{N{\cdot}s/m}$, does change the period significantly.

We release the mass from $x = 1$ m, and after one cycle, it only comes back to about $x = 0.9$ m. I chose $b = 0.211$ $\mathrm{N{\cdot}s/m}$ by fiddling around until I got this result, since a decrease of exactly 10% is easy to discuss. Notice how the amplitude after two cycles is about 0.81 m, i.e., 1 m times 0.9^2: the amplitude has again dropped by exactly 10%. This pattern continues for as long as the simulation runs, e.g., for the last two cycles, we have 0.34818/0.38694=0.89982, or almost exactly 0.9 again. It might have seemed capricious when I chose to use the unrealistic equation $F = -bv$, but this is the payoff. Only with $-bv$ friction do we get this kind of mathematically simple exponential decay.

Because the decay is exponential, it never dies out completely; this is different from the behavior we would have had with Coulomb friction, which does make objects grind completely to a stop at some point. With friction that acts like $F = -bv$, v gets smaller as the oscillations get smaller. The smaller and smaller force then causes them to die out at a rate that is slower and slower.

Contributors and Attributions

4.4: Motion In Three Dimensions

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

4.5: Footnotes

1. Electrical and magnetic interactions *don't* quite behave like this, which is a point we'll take up later in the book.
2. We can now see that the derivation would have been equally valid for $U_i \neq U_f$. The two observers agree on the distance between the particles, so they also agree on the interaction energies, even though they disagree on the kinetic energies.
3. Recall that uppercase P is power, while lowercase p is momentum.
4. This is with the benefit of hindsight. At the time, the word “force” already had certain connotations, and people thought they understood what it meant and how to measure it, e.g., by using a spring scale. From their point of view, $F = dp/dt$ was not a definition but a testable --- and controversial! --- statement.
5. This pathological solution was first noted on page 83, and discussed in more detail on page 910.
6. The converse isn't true, because kinetic energy doesn't depend on the direction of motion, but momentum does. We can change a particle's momentum without changing its energy, as when a pool ball bounces off a bumper, reversing the sign of p.
7. The part of the definition about “by a force” is meant to exclude the transfer of energy by heat conduction, as when a stove heats soup.
8. “Black box” is a traditional engineering term for a device whose inner workings we don't care about.
9. For conceptual simplicity, we ignore the transfer of heat energy to the outside world via the exhaust and radiator. In reality, the sum of these energies plus the useful kinetic energy transferred would equal W.
10. This subroutine isn't as accurate a way of calculating the period as the energy-based one we used in the undamped case, since it only checks whether the mass turned around at some point during the time interval Δt.
11. The relationship is $\omega_{max\,A}/\omega_o = \sqrt{1-1/2Q^2}$, which is similar in form to the equation for the frequency of the free vibration, $\omega_f/\omega_o = \sqrt{1-1/4Q^2}$. A subtle point here is that although the maximum of A and the maximum of A^2 must occur at the same frequency, the maximum energy does not occur, as we might expect, at the same frequency as the maximum of A^2. This is because the interaction energy is proportional to A^2 regardless of frequency, but the kinetic energy is proportional to $A^2\omega^2$. The maximum energy actually occurs are precisely ω_o.
12. For example, the graphs calculated for sinusoidal driving have resonances that are somewhat below the natural frequency, getting lower with increasing damping, until for $Q \leq 1$ the maximum response occurs at $\omega = 0$. In figure m, however, we can see that impulsive driving at $\omega = 2\omega_o$ produces a steady state with more energy than at $\omega = \omega_o$.
13. If you've learned about differential equations, you'll know that any second-order differential equation requires the specification of two boundary conditions in order to specify solution uniquely.
14. Actually, if you know about complex numbers and Euler's theorem, it's not quite so nonsensical.
15. Of course, you could tell in a sealed laboratory which way was down, but that's because there happens to be a big planet nearby, and the planet's gravitational field reaches into the lab, not because space itself has a special down direction. Similarly, if your experiment was sensitive to magnetic fields, it might matter which way the building was oriented, but that's because the earth makes a magnetic field, not because space itself comes equipped with a north direction.
16. The zero here is really a zero *vector*, i.e., a vector whose components are all zero, so we should really represent it with a boldface \vc{0}. There's usually not much danger of confusion, however, so most books, including this one, don't use boldface for the zero vector.
17. There is, however, a different operation, discussed in the next chapter, which multiplies two vectors to give a vector.
18. The symbol ∇ is called a “nabla.” Cool word!

4.E: Problems

1. Derive a formula expressing the kinetic energy of an object in terms of its momentum and mass.(answer check available at lightandmatter.com)

2. Two people in a rowboat wish to move around without causing the boat to move. What should be true about their total momentum? Explain.

3. A bullet leaves the barrel of a gun with a kinetic energy of 90 J. The gun barrel is 50 cm long. The gun has a mass of 4 kg, the bullet 10 g.
(a) Find the bullet's final velocity. (answer check available at lightandmatter.com)
(b) Find the bullet's final momentum. (answer check available at lightandmatter.com)
(c) Find the momentum of the recoiling gun.
(d) Find the kinetic energy of the recoiling gun, and explain why the recoiling gun does not kill the shooter. (answer check available at lightandmatter.com)

4. (solution in the pdf version of the book) The big difference between the equations for momentum and kinetic energy is that one is proportional to v and one to v^2. Both, however, are proportional to m. Suppose someone tells you that there's a third quantity, funkosity, defined as $f = m^2v$, and that funkosity is conserved. How do you know your leg is being pulled?

5. A ball of mass $2m$ collides head-on with an initially stationary ball of mass m. No kinetic energy is transformed into heat or sound. In what direction is the mass-$2m$ ball moving after the collision, and how fast is it going compared to its original velocity? \hwans{hwans:twotoonecollision}

6. A very massive object with velocity v collides head-on with an object at rest whose mass is very small. No kinetic energy is converted into other forms. Prove that the low-mass object recoils with velocity $2v$. [Hint: Use the center-of-mass frame of reference.]

7. A mass m moving at velocity v collides with a stationary target having the same mass m. Find the maximum amount of energy that can be released as heat and sound. (answer check available at lightandmatter.com)

8. A rocket ejects exhaust with an exhaust velocity u. The rate at which the exhaust mass is used (mass per unit time) is b. We assume that the rocket accelerates in a straight line starting from rest, and that no external forces act on it. Let the rocket's initial mass (fuel plus the body and payload) be m_i, and m_f be its final mass, after all the fuel is used up. (a) Find the rocket's final velocity, v, in terms of u, m_i, and m_f. Neglect the effects of special relativity. (b) A typical exhaust velocity for chemical rocket engines is 4000 m/s. Estimate the initial mass of a rocket that could accelerate a one-ton payload to 10% of the speed of light, and show that this design won't work. (For the sake of the estimate, ignore the mass of the fuel tanks. The speed is fairly small compared to c, so it's not an unreasonable approximation to ignore relativity.) (answer check available at lightandmatter.com)

9. An object is observed to be moving at constant speed along a line. Can you conclude that no forces are acting on it? Explain. [Based on a problem by Serway and Faughn.]

10. At low speeds, every car's acceleration is limited by traction, not by the engine's power. Suppose that at low speeds, a certain car is normally capable of an acceleration of $3\ \mathrm{m/s^2}$. If it is towing a trailer with half as much mass as the car itself, what acceleration can it achieve? [Based on a problem from PSSC Physics.]

11. (a) Let T be the maximum tension that an elevator's cable can withstand without breaking, i.e., the maximum force it can exert. If the motor is programmed to give the car an acceleration a ($a > 0$ is upward), what is the maximum mass that the car can have, including passengers, if the cable is not to break?(answer check available at lightandmatter.com)
(b) Interpret the equation you derived in the special cases of $a = 0$ and of a downward acceleration of magnitude g.

12. (solution in the pdf version of the book) When the contents of a refrigerator cool down, the changed molecular speeds imply changes in both momentum and energy. Why, then, does a fridge transfer *power* through its radiator coils, but not *force*?

13. A helicopter of mass m is taking off vertically. The only forces acting on it are the earth's gravitational force and the force, F_{air}, of the air pushing up on the propeller blades.
(a) If the helicopter lifts off at $t = 0$, what is its vertical speed at time t?
(b) Check that the units of your answer to part a make sense.
(c) Discuss how your answer to part a depends on all three 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*.
(d) Plug numbers into your equation from part a, using $m = 2300$ kg, $F_{air} = 27000$ N, and $t = 4.0$ s. (answer check available at lightandmatter.com)

14. A blimp is initially at rest, hovering, when at $t = 0$ the pilot turns on the motor of the propeller. The motor cannot instantly get the propeller going, but the propeller speeds up steadily. The steadily increasing force between the air and the propeller is given by the equation $F = kt$, where k is a constant. If the mass of the blimp is m, find its position as a function of time. (Assume that during the period of time you're dealing with, the blimp is not yet moving fast enough to cause a significant backward force due to air resistance.)(answer check available at lightandmatter.com)

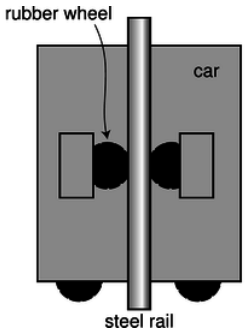
15. (solution in the pdf version of the book) A car is accelerating forward along a straight road. If the force of the road on the car's wheels, pushing it forward, is a constant 3.0 kN, and the car's mass is 1000 kg, then how long will the car take to go from 20 m/s to 50 m/s?

a / Problem 16.

16. A little old lady and a pro football player collide head-on. Compare their forces on each other, and compare their accelerations. Explain.

17. The earth is attracted to an object with a force equal and opposite to the force of the earth on the object. If this is true, why is it that when you drop an object, the earth does not have an acceleration equal and opposite to that of the object?

18. When you stand still, there are two forces acting on you, the force of gravity (your weight) and the normal force of the floor pushing up on your feet. Are these forces equal and opposite? Does Newton's third law relate them to each other? Explain.



b / Problem 19.

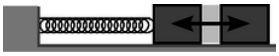
19. Today's tallest buildings are really not that much taller than the tallest buildings of the 1940's. One big problem with making an even taller skyscraper is that every elevator needs its own shaft running the whole height of the building. So many elevators are needed to serve the building's thousands of occupants that the elevator shafts start taking up too much of the space within the building. An alternative is to have elevators that can move both horizontally and vertically: with such a design, many elevator cars can share a few shafts, and they don't get in each other's way too much because they can detour around each other. In this design, it becomes impossible to hang the cars from cables, so they would instead have to ride on rails which they grab onto with wheels. Friction would keep them from slipping. The figure shows such a frictional elevator in its vertical travel mode. (The wheels on the bottom are for when it needs to switch to horizontal motion.)
(a) If the coefficient of static friction between rubber and steel is μ_s, and the maximum mass of the car plus its passengers is M, how much force must there be pressing each wheel against the rail in order to keep the car from slipping? (Assume the car is not accelerating.)(answer check available at lightandmatter.com)

(b) Show that your result has physically reasonable behavior with respect to μ_s. In other words, if there was less friction, would the wheels need to be pressed more firmly or less firmly? Does your equation behave that way?

20. A tugboat of mass m pulls a ship of mass M, accelerating it. Ignore fluid friction acting on their hulls, although there will of course need to be fluid friction acting on the tug's propellers.
(a) If the force acting on the tug's propeller is F, what is the tension, T, in the cable connecting the two ships? \hwhint{hwhint:tugboat}(answer check available at lightandmatter.com)
(b) Interpret your answer in the special cases of $M = 0$ and $M = \infty$.

21. Someone tells you she knows of a certain type of Central American earthworm whose skin, when rubbed on polished diamond, has $\mu_k > \mu_s$. Why is this not just empirically unlikely but logically suspect?

22. A uranium atom deep in the earth spits out an alpha particle. An alpha particle is a fragment of an atom. This alpha particle has initial speed v, and travels a distance d before stopping in the earth.
(a) Find the force, F, from the dirt that stopped the particle, in terms of v, d, and its mass, m. Don't plug in any numbers yet. Assume that the force was constant.(answer check available at lightandmatter.com)
(b) Show that your answer has the right units.
(c) Discuss how your answer to part a depends on all three 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*.
(d) Evaluate your result for $m = 6.7 \times 10^{-27}$ kg, $v = 2.0 \times 10^4$ km/s, and $d = 0.71$ mm.(answer check available at lightandmatter.com)



c / Problem 23, part c.

23. You are given a large sealed box, and are not allowed to open it. Which of the following experiments measure its mass, and which measure its weight? [Hint: Which experiments would give different results on the moon?]
(a) Put it on a frozen lake, throw a rock at it, and see how fast it scoots away after being hit.
(b) Drop it from a third-floor balcony, and measure how loud the sound is when it hits the ground.
(c) As shown in the figure, connect it with a spring to the wall, and watch it vibrate.
(solution in the pdf version of the book)

24. While escaping from the palace of the evil Martian emperor, Sally Spacehound jumps from a tower of height h down to the ground. Ordinarily the fall would be fatal, but she fires her blaster rifle straight down, producing an upward force of magnitude F_B. This force is insufficient to levitate her, but it does cancel out some of the force of gravity. During the time t that she is falling, Sally is unfortunately exposed to fire from the emperor's minions, and can't dodge their shots. Let m be her mass, and g the strength of gravity on Mars.
(a) Find the time t in terms of the other variables.
(b) Check the units of your answer to part a.
(c) For sufficiently large values of F_B, your answer to part a becomes nonsense --- explain what's going on.(answer check available at lightandmatter.com)

25. When I cook rice, some of the dry grains always stick to the measuring cup. To get them out, I turn the measuring cup upside-down and hit the “roof” with my hand so that the grains come off of the “ceiling.” (a) Explain why static friction is irrelevant here. (b) Explain why gravity is negligible. (c) Explain why hitting the cup works, and why its success depends on hitting the cup hard enough.

26. A flexible rope of mass m and length L slides without friction over the edge of a table. Let x be the length of the rope that is hanging over the edge at a given moment in time.
(a) Show that x satisfies the equation of motion $d^2x/dt^2 = gx/L$. [Hint: Use $F = dp/dt$, which allows you to handle the two parts of the rope separately even though mass is moving out of one part and into the other.]
(b) Give a physical explanation for the fact that a larger value of x on the right-hand side of the equation leads to a greater value of the acceleration on the left side.
(c) When we take the second derivative of the function $x(t)$ we are supposed to get essentially the same function back again,

except for a constant out in front. The function e^x has the property that it is unchanged by differentiation, so it is reasonable to look for solutions to this problem that are of the form $x = be^{ct}$, where b and c are constants. Show that this does indeed provide a solution for two specific values of c (and for any value of b).
(d) Show that the sum of any two solutions to the equation of motion is also a solution.
(e) Find the solution for the case where the rope starts at rest at $t = 0$ with some nonzero value of x.

In problems 27-31, analyze the forces using a table in the format shown in section 3.2.6. Analyze the forces in which the italicized object participates.

27. Some people put a spare car key in a little magnetic *box* that they stick under the chassis of their car. Let's say that the box is stuck directly underneath a horizontal surface, and the car is parked. (See instructions above.)

28. Analyze two examples of *objects* at rest relative to the earth that are being kept from falling by forces other than the normal force. Do not use objects in outer space, and do not duplicate problem 27 or 31. (See instructions above.)

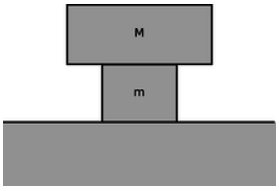
d / Problem 29.

29. A *person* is rowing a boat, with her feet braced. She is doing the part of the stroke that propels the boat, with the ends of the oars in the water (not the part where the oars are out of the water). (See instructions above.)

30. A *farmer* is in a stall with a cow when the cow decides to press him against the wall, pinning him with his feet off the ground. Analyze the forces in which the farmer participates. (See instructions above.)

e / Problem 31.

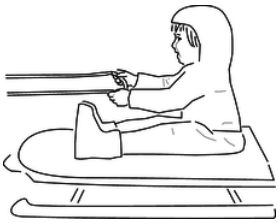
31. A propeller *plane* is cruising east at constant speed and altitude. (See instructions above.)



f / Problem 32

32. The figure shows a stack of two blocks, sitting on top of a table that is bolted to the floor. All three objects are made from identical wood, with their surfaces finished identically using the same sandpaper. We tap the middle block, giving it an initial velocity v to the right. The tap is executed so rapidly that almost no initial velocity is imparted to the top block.
(a) Find the time that will elapse until the slipping between the top and middle blocks stops. Express your answer in terms of v, m, M, g, and the relevant coefficient of friction.(answer check available at lightandmatter.com)
(b) Show that your answer makes sense in terms of units.
(c) Check that your result has the correct behavior when you make m bigger or smaller. Explain. This means that you should discuss the mathematical behavior of the result, and then explain how this corresponds to what would really happen physically.

(d) Similarly, discuss what happens when you make M bigger or smaller.
(e) Similarly, discuss what happens when you make g bigger or smaller.



g / Problem 33.

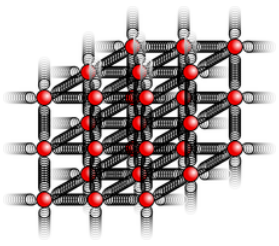
33. Ginny has a plan. She is going to ride her sled while her dog Foo pulls her, and she holds on to his leash. However, Ginny hasn't taken physics, so there may be a problem: she may slide right off the sled when Foo starts pulling.
(a) Analyze all the forces in which Ginny participates, making a table as in subsection 3.2.6.
(b) Analyze all the forces in which the sled participates.
(c) The sled has mass m, and Ginny has mass M. The coefficient of static friction between the sled and the snow is μ_1, and μ_2 is the corresponding quantity for static friction between the sled and her snow pants. Ginny must have a certain minimum mass so that she will not slip off the sled. Find this in terms of the other three variables.(answer check available at lightandmatter.com)
(d) Interpreting your equation from part c, under what conditions will there be no physically realistic solution for M? Discuss what this means physically.

34. (solution in the pdf version of the book) In each case, identify the force that causes the acceleration, and give its Newton's-third-law partner. Describe the effect of the partner force. (a) A swimmer speeds up. (b) A golfer hits the ball off of the tee. (c) An archer fires an arrow. (d) A locomotive slows down.

35. A cop investigating the scene of an accident measures the length L of a car's skid marks in order to find out its speed v at the beginning of the skid. Express v in terms of L and any other relevant variables.(answer check available at lightandmatter.com)

36. An ice skater builds up some speed, and then coasts across the ice passively in a straight line. (a) Analyze the forces, using a table the format shown in subsection 3.2.6.
(b) If his initial speed is v, and the coefficient of kinetic friction is μ_k, find the maximum theoretical distance he can glide before coming to a stop. Ignore air resistance.(answer check available at lightandmatter.com)
(c) Show that your answer to part b has the right units.
(d) Show that your answer to part b depends on the variables in a way that makes sense physically.
(e) Evaluate your answer numerically for $\mu_k = 0.0046$, and a world-record speed of 14.58 m/s. (The coefficient of friction was measured by De Koning et al., using special skates worn by real speed skaters.)(answer check available at lightandmatter.com)
(f) Comment on whether your answer in part e seems realistic. If it doesn't, suggest possible reasons why.

37. (a) Using the solution of problem 37 on page 124, predict how the spring constant of a fiber will depend on its length and cross-sectional area.
(b) The constant of proportionality is called the Young's modulus, E, and typical values of the Young's modulus are about 10^{10} to 10^{11}. What units would the Young's modulus have in the SI system? (solution in the pdf version of the book)



h / Problem 38

38. This problem depends on the results of problems problem 37 on page 124 and problem 37 from this chapter. When atoms form chemical bonds, it makes sense to talk about the spring constant of the bond as a measure of how “stiff” it is. Of course, there aren't really little springs --- this is just a mechanical model. The purpose of this problem is to estimate the spring constant, k, for a single bond in a typical piece of solid matter. Suppose we have a fiber, like a hair or a piece of fishing line, and imagine for simplicity that

it is made of atoms of a single element stacked in a cubical manner, as shown in the figure, with a center-to-center spacing b. A typical value for b would be about 10^{-10} m.
(a) Find an equation for k in terms of b, and in terms of the Young's modulus, E, defined in problem 37 and its solution.
(b) Estimate k using the numerical data given in problem 37.
(c) Suppose you could grab one of the atoms in a diatomic molecule like H_2 or O_2, and let the other atom hang vertically below it. Does the bond stretch by any appreciable fraction due to gravity?

39. This problem has been deleted.

40. Many fish have an organ known as a swim bladder, an air-filled cavity whose main purpose is to control the fish's buoyancy an 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?

41. An oscillator with sufficiently strong damping has its maximum response at $\omega = 0$. Using the result derived on page 912 , find the value of Q at which this behavior sets in. \hwhint{hwhint:maxampatdc}\hwans{hwans:maxampatdc}

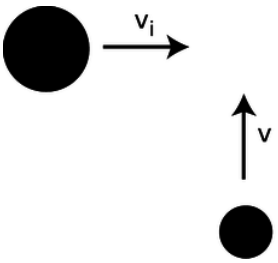
42. An oscillator has Q=6.00, and, for convenience, let's assume $F_m = 1.00$, $\omega_o = 1.00$, and $m = 1.00$. The usual approximations would give

$$\omega_{res} = \omega_o,$$
$$A_{res} = 6.00, \text{and}$$
$$\Delta\omega = 1/6.00.$$

Determine these three quantities numerically using the result derived on page 912 , and compare with the approximations.

43. The apparatus in figure d on page 61 had a natural period of oscillation of 5 hours and 20 minutes. The authors estimated, based on calculations of internal friction in the tungsten wire, that its Q was on the order of 10^6, but they were unable to measure it empirically because it would have taken years for the amplitude to die down by any measurable amount. Although each aluminum or platinum mass was really moving along an arc of a circle, any actual oscillations caused by a violation of the equivalence of gravitational and inertial mass would have been measured in millions of a degree, so it's a good approximation to say that each mass's motion was along a (very short!) straight line segment. We can also treat each mass as if it was oscillating separately from the others. If the principle of equivalence had been violated at the 10^{-12} level, the limit of their experiment's sensitivity, the sun's gravitational force on one of the 0.4-gram masses would have been about 3×10^{-19} N, oscillating with a period of 24 hours due to the rotation of the earth. (We ignore the inertia of the arms, whose total mass was only about 25% of the total mass of the rotating assembly.)
(a) Find the amplitude of the resulting oscillations, and determine the angle to which they would have corresponded, given that the radius of the balance arms was 10 cm.\hwans{hwans:braginskii}
(b) Show that even if their estimate of Q was wildly wrong, it wouldn't have affected this result.

44. (solution in the pdf version of the book) A firework shoots up into the air, and just before it explodes it has a certain momentum and kinetic energy. What can you say about the momenta and kinetic energies of the pieces immediately after the explosion? [Based on a problem from PSSC Physics.]



i / Problem 45

45. (solution in the pdf version of the book) The figure shows a view from above of a collision about to happen between two air hockey pucks sliding without friction. They have the same speed, v_i, before the collision, but the big puck is 2.3 times more massive than the small one. Their sides have sticky stuff on them, so when they collide, they will stick together. At what angle will

they emerge from the collision? In addition to giving a numerical answer, please indicate by drawing on the figure how your angle is defined.

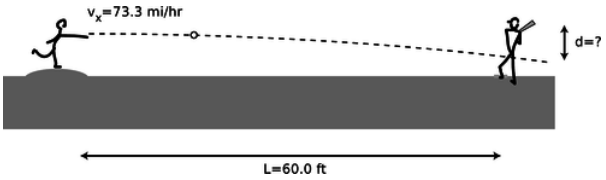
46. A learjet traveling due east at 300 mi/hr collides with a jumbo jet which was heading southwest at 150 mi/hr. The jumbo jet's mass is five times greater than that of the learjet. When they collide, the learjet sticks into the fuselage of the jumbo jet, and they fall to earth together. Their engines stop functioning immediately after the collision. On a map, what will be the direction from the location of the collision to the place where the wreckage hits the ground? (Give an angle.)(answer check available at lightandmatter.com)

47. (a) A ball is thrown straight up with velocity v. Find an equation for the height to which it rises.(answer check available at lightandmatter.com)
(b) Generalize your equation for a ball thrown at an angle θ above horizontal, in which case its initial velocity components are $v_x = v\cos\theta$ and $v_y = v\sin\theta$.(answer check available at lightandmatter.com)

48. At the 2010 Salinas Lettuce Festival Parade, the Lettuce Queen drops her bouquet while riding on a float moving toward the right. Sketch the shape of its trajectory in her frame of reference, and compare with the shape seen by one of her admirers standing on the sidewalk.

49. Two daredevils, Wendy and Bill, go over Niagara Falls. Wendy sits in an inner tube, and lets the 30 km/hr velocity of the river throw her out horizontally over the falls. Bill paddles a kayak, adding an extra 10 km/hr to his velocity. They go over the edge of the falls at the same moment, side by side. Ignore air friction. Explain your reasoning.
(a) Who hits the bottom first?
(b) What is the horizontal component of Wendy's velocity on impact?
(c) What is the horizontal component of Bill's velocity on impact?
(d) Who is going faster on impact?

50. A baseball pitcher throws a pitch clocked at v_x=73.3 mi/h. He throws horizontally. By what amount, d, does the ball drop by the time it reaches home plate, L=60.0 ft away?
(a) First find a symbolic answer in terms of L, v_x, and g.(answer check available at lightandmatter.com)
(b) Plug in and find a numerical answer. Express your answer in units of ft. (Note: 1 ft=12 in, 1 mi=5280 ft, and 1 in=2.54 cm) (answer check available at lightandmatter.com)



j / Problem 50.

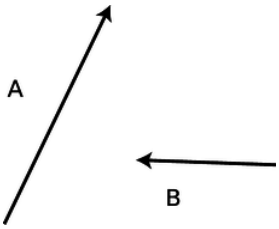
51. A batter hits a baseball at speed v, at an angle θ above horizontal.

(a) Find an equation for the range (horizontal distance to where the ball falls), R, in terms of the relevant variables. Neglect air friction and the height of the ball above the ground when it is hit. \hwans{hwans:baseballrange}
(b) Interpret your equation in the cases of θ=0 and $\theta = 90^\circ$.
(c) Find the angle that gives the maximum range.\hwans{hwans:baseballrange}

52. In this problem you'll extend the analysis in problem 51

to include air friction by writing a computer program. For a game played at sea level, the force due to air friction is approximately $(7\times10^{-4}\ \mathrm{N{\cdot}s^2/m^2})v^2$, in the direction opposite to the motion of the ball.The mass of a baseball is 0.146 kg.
(a) For a ball hit at a speed of 45.0 m/s from a height of 1.0 m, find the optimal angle and the resulting range. \hwans{hwans:baseballrangeair}
(b) How much farther would the ball fly at the Colorado Rockies' stadium, where the thinner air gives 18 percent less air friction? \hwans{hwans:baseballrangeair}

53. If you walk 35 km at an angle 25° counterclockwise from east, and then 22 km at 230° counterclockwise from east, find the distance and direction from your starting point to your destination. (answer check available at lightandmatter.com)



k / Problem 54.

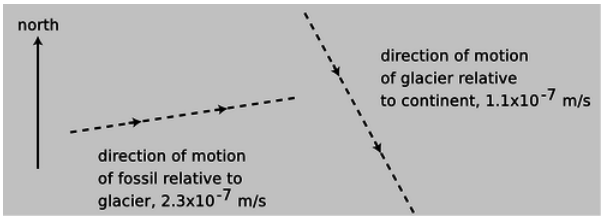
54. The figure shows vectors $\mathbf{A}$ and $\mathbf{B}$. As in figure p on p. 200, graphically calculate the following:

$\mathbf{A}+\mathbf{B}$, $\mathbf{A}-\mathbf{B}$, $\mathbf{B}-\mathbf{A}$, $-2\mathbf{B}$, $\mathbf{A}-2\mathbf{B}$

No numbers are involved.

55. Phnom Penh is 470 km east and 250 km south of Bangkok. Hanoi is 60 km east and 1030 km north of Phnom Penh.
(a) Choose a coordinate system, and translate these data into Δx and Δy values with the proper plus and minus signs.
(b) Find the components of the $\Delta\mathbf{r}$ vector pointing from Bangkok to Hanoi.(answer check available at lightandmatter.com)

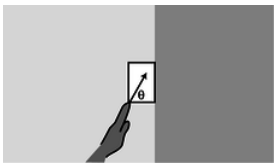
56. Is it possible for a helicopter to have an acceleration due east and a velocity due west? If so, what would be going on? If not, why not?



l / Problem 57.

57. A dinosaur fossil is slowly moving down the slope of a glacier under the influence of wind, rain and gravity. At the same time, the glacier is moving relative to the continent underneath. The dashed lines represent the directions but not the magnitudes of the velocities. Pick a scale, and use graphical addition of vectors to find the magnitude and the direction of the fossil's velocity relative to the continent. You will need a ruler and protractor. (answer check available at lightandmatter.com)

58. A bird is initially flying horizontally east at 21.1 m/s, but one second later it has changed direction so that it is flying horizontally and 7° north of east, at the same speed. What are the magnitude and direction of its acceleration vector during that one second time interval? (Assume its acceleration was roughly constant.) (answer check available at lightandmatter.com)



m / Problem 59

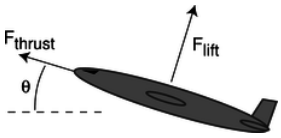
59. Your hand presses a block of mass m against a wall with a force $\mathbf{F}_H$ acting at an angle θ, as shown in the figure. Find the minimum and maximum possible values of $|\mathbf{F}_H|$ that can keep the block stationary, in terms of m, g, θ, and μ_s, the coefficient of static friction between the block and the wall. Check both your answers in the case of $\theta = 90^\circ$, and interpret the case where the maximum force is infinite.(answer check available at lightandmatter.com)

60. A skier of mass m is coasting down a slope inclined at an angle θ compared to horizontal. Assume for simplicity that the treatment of kinetic friction given in chapter 5 is appropriate here, although a soft and wet surface actually behaves a little differently. The coefficient of kinetic friction acting between the skis and the snow is μ_k, and in addition the skier experiences an air friction force of magnitude bv^2, where b is a constant.
(a) Find the maximum speed that the skier will attain, in terms of the variables m, g, θ, μ_k, and b.(answer check available at lightandmatter.com)
(b) For angles below a certain minimum angle θ_{min}, the equation gives a result that is not mathematically meaningful. Find an

equation for θ_{min}, and give a physical explanation of what is happening for $\theta < \theta_{min}$.(answer check available at lightandmatter.com)

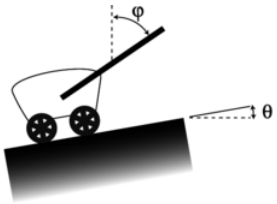
61. A gun is aimed horizontally to the west. The gun is fired, and the bullet leaves the muzzle at $t = 0$. The bullet's position vector as a function of time is $\mathbf{r} = b\hat{\mathbf{x}} + ct\hat{\mathbf{y}} + dt^2\hat{\mathbf{z}}$, where b, c, and d are positive constants.
(a) What units would b, c, and d need to have for the equation to make sense?
(b) Find the bullet's velocity and acceleration as functions of time.
(c) Give physical interpretations of b, c, d, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$.

62. Annie Oakley, riding north on horseback at 30 mi/hr, shoots her rifle, aiming horizontally and to the northeast. The muzzle speed of the rifle is 140 mi/hr. When the bullet hits a defenseless fuzzy animal, what is its speed of impact? Neglect air resistance, and ignore the vertical motion of the bullet.(solution in the pdf version of the book)



n / Problem 63

63. A cargo plane has taken off from a tiny airstrip in the Andes, and is climbing at constant speed, at an angle of $\theta = 17^\circ$ with respect to horizontal. Its engines supply a thrust of $F_{thrust} = 200$ kN, and the lift from its wings is $F_{lift} = 654$ kN. Assume that air resistance (drag) is negligible, so the only forces acting are thrust, lift, and weight. What is its mass, in kg? (solution in the pdf version of the book)



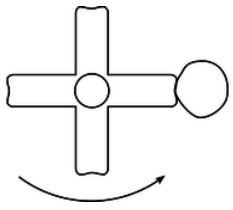
o / Problem 64

64. A wagon is being pulled at constant speed up a slope θ by a rope that makes an angle ϕ with the vertical. (a) Assuming negligible friction, show that the tension in the rope is given by the equation

$$T = \frac{\sin\theta}{\sin(\theta+\phi)}mg,$$

(b) Interpret this equation in the special cases of $\phi = 0$ and $\phi = 180^\circ - \theta$.(solution in the pdf version of the book)

65. The angle of repose is the maximum slope on which an object will not slide. On airless, geologically inert bodies like the moon or an asteroid, the only thing that determines whether dust or rubble will stay on a slope is whether the slope is less steep than the angle of repose.
(a) Find an equation for the angle of repose, deciding for yourself what are the relevant variables.
(b) On an asteroid, where g can be thousands of times lower than on Earth, would rubble be able to lie at a steeper angle of repose? (solution in the pdf version of the book)



p / Problem 66.

66. When you're done using an electric mixer, you can get most of the batter off of the beaters by lifting them out of the batter with the motor running at a high enough speed. Let's imagine, to make things easier to visualize, that we instead have a piece of tape

stuck to one of the beaters.
(a) Explain why static friction has no effect on whether or not the tape flies off.
(b) Analyze the forces in which the tape participates, using a table the format shown in subsection 3.2.6.
(c) Suppose you find that the tape doesn't fly off when the motor is on a low speed, but at a greater speed, the tape won't stay on. Why would the greater speed change things? [Hint: If you don't invoke any law of physics, you haven't explained it.]

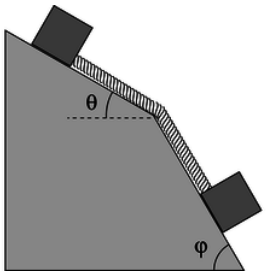
67. Show that the expression $|\mathbf{v}|^2/r$ has the units of acceleration.

68. A plane is flown in a loop-the-loop of radius 1.00 km. The plane starts out flying upside-down, straight and level, then begins curving up along the circular loop, and is right-side up when it reaches the top. (The plane may slow down somewhat on the way up.) How fast must the plane be going at the top if the pilot is to experience no force from the seat or the seatbelt while at the top of the loop? (answer check available at lightandmatter.com)

69. Find the angle between the following two vectors:

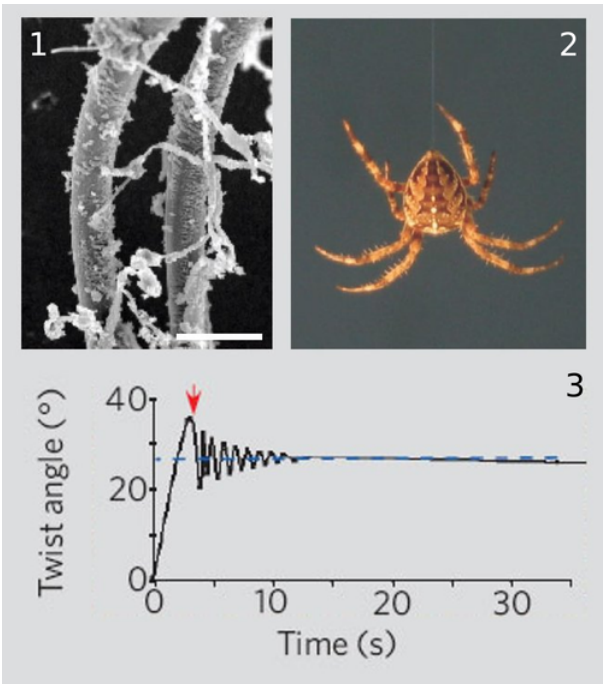
$$\hat{\mathbf{x}}+2\hat{\mathbf{y}}+3\hat{\mathbf{z}}$$
$$4\hat{\mathbf{x}}+5\hat{\mathbf{y}}+6\hat{\mathbf{z}}$$

\hwhint{hwhint:anglebetween}(answer check available at lightandmatter.com)



q / Problem 70.

70. The two blocks shown in the figure have equal mass, m, and the surface is frictionless. (a) What is the tension in the massless rope?\hwhint{hwhint:ropeslopes}(answer check available at lightandmatter.com)
(b) Show that the units of your answer make sense.
(c) Check the physical behavior of your answer in the special cases of $\phi \leq \theta$ and $\theta = 0$, $\phi = 90^\circ$.



r / Problem 71.

71. (a) We observe that the amplitude of a certain free oscillation decreases from A_o to A_o/Z after n oscillations. Find its Q. (answer check available at lightandmatter.com)
(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.
2. It doesn't become brittle with repeated twisting as a copper wire would.
3. 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.
4. Over much longer time scales (hours), the thread eventually resets itself to its original equilbrium 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.(answer check available at lightandmatter.com)

72. A cross-country skier is gliding on a level trail, with negligible friction. Then, when he is at position $x = 0$, the tip of his skis enters a patch of dirt. As he rides onto the dirt, more and more of his weight is being supported by the dirt. The skis have length ℓ, so if he reached $x = \ell$ without stopping, his weight would be completely on the dirt. This problem deals with the motion for $x < \ell$.
(a) Find the acceleration in terms of x, as well as any other relevant constants.

(b) This is a second-order differential equation. You should be able to find the solution simply by thinking about some commonly occuring functions that you know about, and finding two that have the right properties. If these functions are $x = f(t)$ and $x = g(t)$, then the most general solution to the equations of motion will be of the form $x = af + bg$, where a and b are constants to be determined from the initial conditions.
(c) Suppose that the initial velocity v_o at $x = 0$ is such that he stops at $x < \ell$. Find the time until he stops, and show that, counterintuitively, this time is independent of v_o. Explain physically why this is true.(answer check available at lightandmatter.com)

73. A microwave oven works by twisting molecules one way and then the other, counterclockwise and then clockwise about their own centers, millions of times a second. If you put an ice cube or a stick of butter in a microwave, you'll observe that the solid doesn't heat very quickly, although eventually melting begins in one small spot. Once this spot forms, it grows rapidly, while the rest of the solid remains solid; it appears that a microwave oven heats a liquid much more rapidly than a solid. Explain why this should happen, based on the atomic-level description of heat, solids, and liquids. (See, e.g., figure b on page 110.)

Don't repeat the following common mistakes:

In a solid, the atoms are packed more tightly and have less space between them. Not true. Ice floats because it's *less* dense than water.

In a liquid, the atoms are moving much faster. No, the difference in average speed between ice at $-1\,^\circ\mathrm{C}$ and water at $1\,^\circ\mathrm{C}$ is only 0.4%.

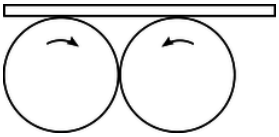
74. Problem 2-16 on page 120 was intended to be solved using conservation of energy. Solve the same problem using Newton's laws.

75. A bead slides down along a piece of wire that is in the shape of a helix. The helix lies on the surface of a vertical cylinder of radius r, and the vertical distance between turns is d.
(a) Ordinarily when an object slides downhill under the influence of kinetic friction, the velocity-independence of kinetic friction implies that the acceleration is constant, and therefore there is no limit to the object's velocity. Explain the physical reason why this argument fails here, so that the bead will in fact have some limiting velocity.
(b) Find the limiting velocity.

(c) Show that your result has the correct behavior in the limit of $r \to \infty$. [Problem by B. Korsunsky.](answer check available at lightandmatter.com)

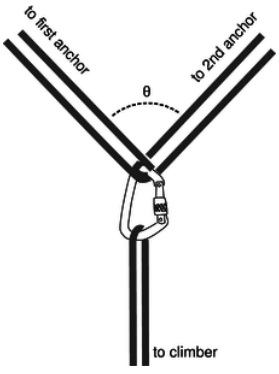
76. A person on a bicycle is to coast down a ramp of height h and then pass through a circular loop of radius r. What is the smallest value of h for which the cyclist will complete the loop without falling? (Ignore the kinetic energy of the spinning wheels.) (answer check available at lightandmatter.com)

77. A car accelerates from rest. At low speeds, its acceleration is limited by static friction, so that if we press too hard on the gas, we will "burn rubber" (or, for many newer cars, a computerized traction-control system will override the gas pedal). At higher speeds, the limit on acceleration comes from the power of the engine, which puts a limit on how fast kinetic energy can be developed.
(a) Show that if a force F is applied to an object moving at speed v, the power required is given by $P = vF$.
(b) Find the speed v at which we cross over from the first regime described above to the second. At speeds higher than this, the engine does not have enough power to burn rubber. Express your result in terms of the car's power P, its mass m, the coefficient of static friction μ_s, and g.(answer check available at lightandmatter.com)
(c) Show that your answer to part b has units that make sense.
(d) Show that the dependence of your answer on each of the four variables makes sense physically.
(e) The 2010 Maserati Gran Turismo Convertible has a maximum power of 3.23×10^5 W (433 horsepower) and a mass (including a 50-kg driver) of 2.03×10^3 kg. (This power is the maximum the engine can supply at its optimum frequency of 7600 r.p.m. Presumably the automatic transmission is designed so a gear is available in which the engine will be running at very nearly this frequency when the car is at moving at v.) Rubber on asphalt has $\mu_s \approx 0.9$. Find v for this car. Answer: 18 m/s, or about 40 miles per hour.
(f) Our analysis has neglected air friction, which can probably be approximated as a force proportional to v^2. The existence of this force is the reason that the car has a maximum speed, which is 176 miles per hour. To get a feeling for how good an approximation it is to ignore air friction, find what fraction of the engine's maximum power is being used to overcome air resistance when the car is moving at the speed v found in part e. Answer: 1%



s / Problem 78.

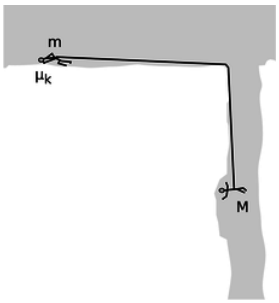
78. Two wheels of radius r rotate in the same vertical plane with angular velocities $+\Omega$ and $-\Omega$ about axes that are parallel and at the same height. The wheels touch one another at a point on their circumferences, so that their rotations mesh like gears in a gear train. A board is laid on top of the wheels, so that two friction forces act upon it, one from each wheel. Characterize the three qualitatively different types of motion that the board can exhibit, depending on the initial conditions.



t / Problem 79.

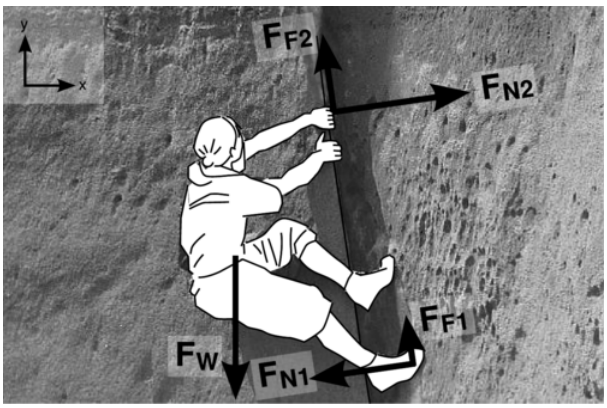
79. For safety, mountain climbers often wear a climbing harness and tie in to other climbers on a rope team or to anchors such as pitons or snow anchors. When using anchors, the climber usually wants to tie in to more than one, both for extra strength and for redundancy in case one fails. The figure shows such an arrangement, with the climber hanging from a pair of anchors forming a "Y" at an angle θ. The usual advice is to make $\theta < 90^\circ$; for large values of θ, the stress placed on the anchors can be many times

greater than the actual load L, so that two anchors are actually *less* safe than one.
(a) Find the force S at each anchor in terms of L and θ. (answer check available at lightandmatter.com)
(b) Verify that your answer makes sense in the case of $\theta = 0$.
(c) Interpret your answer in the case of $\theta = 180^\circ$.
(d) What is the smallest value of θ for which S equals or exceeds L, so that for larger angles a failure of at least one anchor is *more* likely than it would have been with a single anchor?(answer check available at lightandmatter.com)



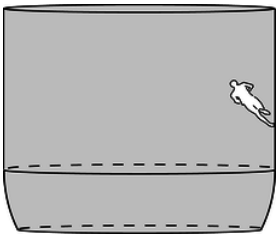
u / Problem 80.

80. Mountain climbers with masses m and M are roped together while crossing a horizontal glacier when a vertical crevasse opens up under the climber with mass M. The climber with mass m drops down on the snow and tries to stop by digging into the snow with the pick of an ice ax. Alas, this story does not have a happy ending, because this doesn't provide enough friction to stop. Both m and M continue accelerating, with M dropping down into the crevasse and m being dragged across the snow, slowed only by the kinetic friction with coefficient μ_k acting between the ax and the snow. There is no significant friction between the rope and the lip of the crevasse.
(a) Find the acceleration a.(answer check available at lightandmatter.com)
(b) Check the units of your result.
(c) Check the dependence of your equation on the variables. That means that for each variable, you should determine what its effect on a should be physically, and then what your answer from part a says its effect would be mathematically.



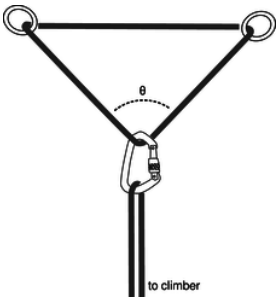
v / Problem 81.

81. Complete example 71 on p. 205 by expressing the remaining nine x and y components of the forces in terms of the five magnitudes and the small, positive angle $\theta \approx 9^\circ$ by which the crack overhangs. (answer check available at lightandmatter.com)



w / Problem 82.

82. In a well known stunt from circuses and carnivals, a motorcyclist rides around inside a big bowl, gradually speeding up and rising higher. Eventually the cyclist can get up to where the walls of the bowl are vertical. Let's estimate the conditions under which a running human could do the same thing.
(a) If the runner can run at speed v, and her shoes have a coefficient of static friction μ_s, what is the maximum radius of the circle? (answer check available at lightandmatter.com)
(b) Show that the units of your answer make sense.
(c) Check that its dependence on the variables makes sense.
(d) Evaluate your result numerically for $v = 10\ \mathrm{m/s}$ (the speed of an olympic sprinter) and $\mu_s = 5$. (This is roughly the highest coefficient of static friction ever achieved for surfaces that are not sticky. The surface has an array of microscopic fibers like a hair brush, and is inspired by the hairs on the feet of a gecko. These assumptions are not necessarily realistic, since the person would have to run at an angle, which would be physically awkward.)(answer check available at lightandmatter.com)



x / Problem 83.

83. Problem 79 discussed a possible correct way of setting up a redundant anchor for mountaineering. The figure for this problem shows an incorrect way of doing it, by arranging the rope in a triangle (which we'll take to be isoceles). One of the bad things about the triangular arrangement is that it requires more force from the anchors, making them more likely to fail. (a) Using the same notation as in problem 79, find S in terms of L and θ. (answer check available at lightandmatter.com)
(b) Verify that your answer makes sense in the case of $\theta = 0$, and compare with the correct setup.

84. At a picnic, someone hands you a can of beer. The ground is uneven, and you don't want to spill your drink. You reason that it will be more stable if you drink some of it first in order to lower its center of mass. How much should you drink in order to make the center of mass as low as possible? [Based on a problem by Walter van B. Roberts and Martin Gardner.]

Contributors and Attributions

CHAPTER OVERVIEW

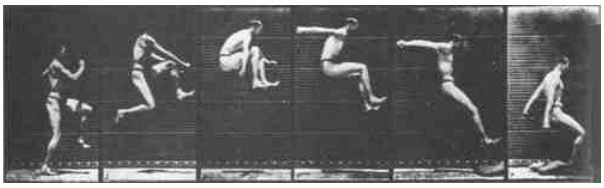
5: Conservation of Angular Momentum

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

5.1: Angular Momentum In Two Dimensions

Sure, and maybe the sun won't come up tomorrow." Of course, the sun only appears to go up and down because the earth spins, so the cliche should really refer to the unlikelihood of the earth's stopping its rotation abruptly during the night. Why can't it stop? It wouldn't violate conservation of momentum, because the earth's rotation doesn't add anything to its momentum. While California spins in one direction, some equally massive part of India goes the opposite way, canceling its momentum. A halt to Earth's rotation would entail a drop in kinetic energy, but that energy could simply be converted into some other form, such as heat.



a / The jumper can't move his legs counterclockwise without moving his arms clockwise. (Thomas Eakins.)

Other examples along these lines are not hard to find. An atom spins at the same rate for billions of years. A high-diver who is rotating when he comes off the board does not need to make any physical effort to continue rotating, and indeed would be unable to stop rotating before he hit the water.

These observations have the hallmarks of a conservation law:

- *A closed system is involved.* Nothing is making an effort to twist the earth, the hydrogen atom, or the high-diver. They are isolated from rotation-changing influences, i.e., they are closed systems.
- *Something remains unchanged.* There appears to be a numerical quantity for measuring rotational motion such that the total amount of that quantity remains constant in a closed system.
- *Something can be transferred back and forth without changing the total amount.* In the photo of the old-fashioned high jump, a, the jumper wants to get his feet out in front of him so he can keep from doing a "face plant" when he lands. Bringing his feet forward would involve a certain quantity of counterclockwise rotation, but he didn't start out with any rotation when he left the ground. Suppose we consider counterclockwise as positive and clockwise as negative. The only way his legs can acquire some positive rotation is if some other part of his body picks up an equal amount of negative rotation. This is why he swings his arms up behind him, clockwise.

What numerical measure of rotational motion is conserved? Car engines and old-fashioned LP records have speeds of rotation measured in rotations per minute (r.p.m.), but the number of rotations per minute (or per second) is not a conserved quantity. A twirling figure skater, for instance, can pull her arms in to increase her r.p.m.'s. The first section of this chapter deals with the numerical definition of the quantity of rotation that results in a valid conservation law.

When most people think of rotation, they think of a solid object like a wheel rotating in a circle around a fixed point. Examples of this type of rotation, called rigid rotation or rigid-body rotation, include a spinning top, a seated child's swinging leg, and a helicopter's spinning propeller. Rotation, however, is a much more general phenomenon, and includes noncircular examples such as a comet in an elliptical orbit around the sun, or a cyclone, in which the core completes a circle more quickly than the outer parts.

If there is a numerical measure of rotational motion that is a conserved quantity, then it must include nonrigid cases like these, since nonrigid rotation can be traded back and forth with rigid rotation. For instance, there is a trick for finding out if an egg is raw or hardboiled. If you spin a hardboiled egg and then stop it briefly with your finger, it stops dead. But if you do the same with a raw egg, it springs back into rotation because the soft interior was still swirling around within the momentarily motionless shell. The pattern of flow of the liquid part is presumably very complex and nonuniform due to the asymmetric shape of the egg and the different consistencies of the yolk and the white, but there is apparently some way to describe the liquid's total amount of rotation with a single number, of which some percentage is given back to the shell when you release it.

The best strategy is to devise a way of defining the amount of rotation of a single small part of a system. The amount of rotation of a system such as a cyclone will then be defined as the total of all the contributions from its many small parts.

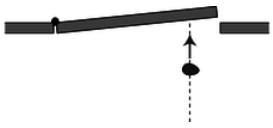


***Figure b**: An overhead view of a piece of putty being thrown at a door. Even though the putty is neither spinning nor traveling along a curve, we must define it has having some kind of "rotation" because it is able to make the door rotate.*

The quest for a conserved quantity of rotation even requires us to broaden the rotation concept to include cases where the motion doesn't repeat or even curve around. If you throw a piece of putty at a door, (Figure b), the door will recoil and start rotating. The putty was traveling straight, not in a circle, but if there is to be a general conservation law that can cover this situation, it appears that we must describe the putty as having had some "rotation," which it then gave up to the door. The best way of thinking about it is to attribute rotation to any moving object or part of an object that changes its angle in relation to the axis of rotation. In the putty-and-door example, the hinge of the door is the natural point to think of as an axis, and the putty changes its angle as seen by someone standing at the hinge, Figure c. For this reason, the conserved quantity we are investigating is called *angular momentum*. The symbol for angular momentum can't be "a" or "m," since those are used for acceleration and mass, so the letter L is arbitrarily chosen instead.

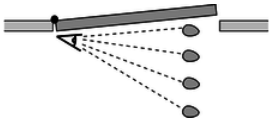


***Figure c:** As seen by someone standing at the axis, the putty changes its angular position. We therefore define it as having angular momentum.*

Imagine a 1 kg blob of putty, thrown at the door at a speed of 1 m/s, which hits the door at a distance of 1 m from the hinge. We define this blob to have 1 unit of angular momentum. When it hits the door, the door will recoil and start rotating. We can use the speed at which the door recoils as a measure of the angular momentum the blob brought in.[1]

Experiments show, not surprisingly, that a 2 kg blob thrown in the same way makes the door rotate twice as fast, so the angular momentum of the putty blob must be proportional to mass,

$$L \propto m.$$

Similarly, experiments show that doubling the velocity of the blob will have a doubling effect on the result, so its angular momentum must be proportional to its velocity as well,

$$L \propto mv.$$

You have undoubtedly had the experience of approaching a closed door with one of those bar-shaped handles on it and pushing on the wrong side, the side close to the hinges. You feel like an idiot, because you have so little leverage that you can hardly budge the door. The same would be true with the putty blob.

***Figure d**: A putty blob thrown directly at the axis has no angular motion, and therefore no angular momentum. It will not cause the door to rotate.*

Experiments would show that the amount of rotation the blob can give to the door is proportional to the distance, r, from the axis of rotation, so angular momentum must be proportional to r as well,

$$L \propto mvr.$$

We are almost done, but there is one missing ingredient. We know on grounds of symmetry that a putty ball thrown directly inward toward the hinge will have no angular momentum to give to the door. After all, there would not even be any way to decide whether

the ball's rotation was clockwise or counterclockwise in this situation. It is therefore only the component of the blob's velocity vector perpendicular to the door that should be counted in its angular momentum,

$$L = mv_{\perp}r.$$

More generally, $v_{\perp}$ should be thought of as the component of the object's velocity vector that is perpendicular to the line joining the object to the axis of rotation.

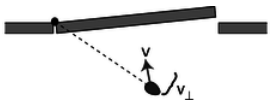


***Figure e**: Only the component of the velocity vector perpendicular to the line connecting the object to the axis should be counted into the definition of angular momentum.*

We find that this equation agrees with the definition of the original putty blob as having one unit of angular momentum, and we can now see that the units of angular momentum are $(\mathrm{kg{\cdot}m/s){\cdot}m}$, i.e., $\mathrm{kg{\cdot}m^2/s}$. Summarizing, we have

$$L = mv_{\perp}r[\text{angular momentum of a particle in two dimensions}],$$

where m is the particle's mass, $v_{\perp}$ is the component of its velocity vector perpendicular to the line joining it to the axis of rotation, and r is its distance from the axis. (Note that r is not necessarily the radius of a circle.) Positive and negative signs of angular momentum are used to describe opposite directions of rotation. The angular momentum of a finite-sized object or a system of many objects is found by dividing it up into many small parts, applying the equation to each part, and adding to find the total amount of angular momentum. (As implied by the word “particle,” matter isn't the only thing that can have angular momentum. Light can also have angular momentum, and the above equation would not apply to light.)

Conservation of angular momentum has been verified over and over again by experiment, and is now believed to be one of the most fundamental principles of physics, along with conservation of mass, energy, and momentum.

Example 1: A figure skater pulls her arms in

When a figure skater is twirling, there is very little friction between her and the ice, so she is essentially a closed system, and her angular momentum is conserved. If she pulls her arms in, she is decreasing r for all the atoms in her arms.

Figure f: A figure skater pulls in her arms so that she can execute a spin more rapidly.

It would violate conservation of angular momentum if she then continued rotating at the same speed, i.e., taking the same amount of time for each revolution, because her arms' contributions to her angular momentum would have decreased, and no other part of her would have increased its angular momentum. This is impossible because it would violate conservation of angular momentum. If her total angular momentum is to remain constant, the decrease in r for her arms must be compensated for by an overall increase in her rate of rotation. That is, by pulling her arms in, she substantially reduces the time for each rotation.

Example 2: Earth's slowing rotation and the receding moon

The earth's rotation is actually slowing down very gradually, with the kinetic energy being dissipated as heat by friction between the land and the tidal bulges raised in the seas by the earth's gravity. Does this mean that angular momentum is not really perfectly conserved? No, it just means that the earth is not quite a closed system by itself. If we consider the earth and moon as a system, then the angular momentum lost by the earth must be gained by the moon somehow. In fact very precise measurements of the distance between the earth and the moon have been carried out by bouncing laser beams off of a mirror left there by astronauts, and these measurements show that the moon is receding from the earth at a rate of 4 centimeters per year! The moon's greater value of r means that it has a greater angular momentum, and the increase turns out to be exactly the amount lost by the earth. In the days of the dinosaurs, the days were significantly shorter, and the moon was closer and appeared bigger in the sky.

Figure g: A view of the earth-moon system from above the north pole. All distances have been highly distorted for legibility.

But what force is causing the moon to speed up, drawing it out into a larger orbit? It is the gravitational forces of the earth's tidal bulges. In figure g, the earth's rotation is counterclockwise (arrow). The moon's gravity creates a bulge on the side near it, because its gravitational pull is stronger there, and an "anti-bulge" on the far side, since its gravity there is weaker. For simplicity, let's focus on the tidal bulge closer to the moon. Its frictional force is trying to slow down the earth's rotation, so its force on the earth's solid crust is toward the bottom of the figure. By Newton's third law, the crust must thus make a force on the bulge which is toward the top of the figure. This causes the bulge to be pulled forward at a slight angle, and the bulge's gravity therefore pulls the moon forward, accelerating its orbital motion about the earth and flinging it outward.

The result would obviously be extremely difficult to calculate directly, and this is one of those situations where a conservation law allows us to make precise quantitative statements about the outcome of a process when the calculation of the process itself would be prohibitively complex.

Restriction to rotation in a plane

Is angular momentum a vector, or a scalar? It does have a direction in space, but it's a direction of rotation, not a straight-line direction like the directions of vectors such as velocity or force. It turns out that there is a way of defining angular momentum as a vector, but in this section the examples will be confined to a single plane of rotation, i.e., effectively two-dimensional situations. In this special case, we can choose to visualize the plane of rotation from one side or the other, and to define clockwise and counterclockwise rotation as having opposite signs of angular momentum. "Effectively" two-dimensional means that we can deal with objects that aren't flat, as long as the velocity vectors of all their parts lie in a plane.

Discussion Questions

◊ Conservation of plain old momentum, p, can be thought of as the greatly expanded and modified descendant of Galileo's original principle of inertia, that no force is required to keep an object in motion. The principle of inertia is counterintuitive, and there are many situations in which it appears superficially that a force *is* needed to maintain motion, as maintained by Aristotle. Think of a situation in which conservation of angular momentum, L, also seems to be violated, making it seem incorrectly that something external must act on a closed system to keep its angular momentum from "running down."

4.1.2 Application to planetary motion

We now discuss the application of conservation of angular momentum to planetary motion, both because of its intrinsic importance and because it is a good way to develop a visual intuition for angular momentum.

Kepler's law of equal areas states that the area swept out by a planet in a certain length of time is always the same. Angular momentum had not been invented in Kepler's time, and he did not even know the most basic physical facts about the forces at work. He thought of this law as an entirely empirical and unexpectedly simple way of summarizing his data, a rule that succeeded in describing and predicting how the planets sped up and slowed down in their elliptical paths. It is now fairly simple, however, to show that the equal area law amounts to a statement that the planet's angular momentum stays constant.

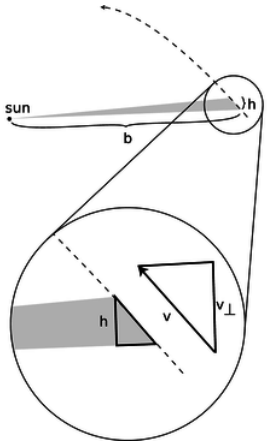


Figure h: The area swept out by a planet in its orbit.

There is no simple geometrical rule for the area of a pie wedge cut out of an ellipse, but if we consider a very short time interval, as shown in figure h, the shaded shape swept out by the planet is very nearly a triangle. We do know how to compute the area of a triangle. It is one half the product of the base and the height:

$$\text{area} = \frac{1}{2}bh.$$

We wish to relate this to angular momentum, which contains the variables r and $v_\perp$. If we consider the sun to be the axis of rotation, then the variable r is identical to the base of the triangle, $r = b$. Referring to the magnified portion of the figure, $v_\perp$ can be related to h, because the two right triangles are similar:

$$\frac{h}{\text{distance traveled}} = \frac{v_\perp}{|\mathbf{v}|}$$

The area can thus be rewritten as

$$\text{area} = \frac{1}{2}r\frac{v_\perp(\text{distance traveled})}{|\mathbf{v}|}.$$

The distance traveled equals $|\mathbf{v}|\Delta t$, so this simplifies to

$$\text{area} = \frac{1}{2}rv_\perp\Delta t.$$

We have found the following relationship between angular momentum and the rate at which area is swept out:

$$L = 2m\frac{\text{area}}{\Delta t}.$$

The factor of 2 in front is simply a matter of convention, since any conserved quantity would be an equally valid conserved quantity if you multiplied it by a constant. The factor of m was not relevant to Kepler, who did not know the planets' masses, and who was only describing the motion of one planet at a time.

We thus find that Kepler's equal-area law is equivalent to a statement that the planet's angular momentum remains constant. But wait, why should it remain constant? --- the planet is not a closed system, since it is being acted on by the sun's gravitational force. There are two valid answers. The first is that it is actually the total angular momentum of the sun plus the planet that is conserved. The sun, however, is millions of times more massive than the typical planet, so it accelerates very little in response to the planet's gravitational force. It is thus a good approximation to say that the sun doesn't move at all, so that no angular momentum is transferred between it and the planet.

The second answer is that to change the planet's angular momentum requires not just a force but a force applied in a certain way. Later in this section (starting on page 254) we discuss the transfer of angular momentum by a force, but the basic idea here is that a force directly in toward the axis does not change the angular momentum.

Discussion Questions

◊ Suppose an object is simply traveling in a straight line at constant speed. If we pick some point not on the line and call it the axis of rotation, is area swept out by the object at a constant rate?

◊

i / Discussion question B.

The figure is a strobe photo of a pendulum bob, taken from underneath the pendulum looking straight up. The black string can't be seen in the photograph. The bob was given a slight sideways push when it was released, so it did not swing in a plane. The bright spot marks the center, i.e., the position the bob would have if it hung straight down at us. Does the bob's angular momentum appear to remain constant if we consider the center to be the axis of rotation?

4.1.3 Two theorems about angular momentum

With plain old momentum, $\mathbf{p}$, we had the freedom to work in any inertial frame of reference we liked. The same object could have different values of momentum in two different frames, if the frames were not at rest with respect to each other. Conservation of momentum, however, would be true in either frame. As long as we employed a single frame consistently throughout a calculation, everything would work.

The same is true for angular momentum, and in addition there is an ambiguity that arises from the definition of an axis of rotation. For a wheel, the natural choice of an axis of rotation is obviously the axle, but what about an egg rotating on its side? The egg has an asymmetric shape, and thus no clearly defined geometric center. A similar issue arises for a cyclone, which does not even have a sharply defined shape, or for a complicated machine with many gears. The following theorem, the first of two presented in this section, explains how to deal with this issue. Although I have put descriptive titles above both theorems, they have no generally accepted names. The proofs, given on page 913, use the vector cross-product technique introduced in section 4.3, which greatly simplifies them.

The choice of axis theorem. It is entirely arbitrary what point one defines as the axis for purposes of calculating angular momentum. If a closed system's angular momentum is conserved when calculated with one choice of axis, then it will be conserved for any other choice of axis. Likewise, any inertial frame of reference may be used. The theorem also holds in the case where the system is not closed, but the total external force is zero.

Example 3: Colliding asteroids described with different axes

Observers on planets A and B both see the two asteroids colliding. The asteroids are of equal mass and their impact speeds are the same. Astronomers on each planet decide to define their own planet as the axis of rotation. Planet A is twice as far from the collision as planet B. The asteroids collide and stick. For simplicity, assume planets A and B are both at rest.

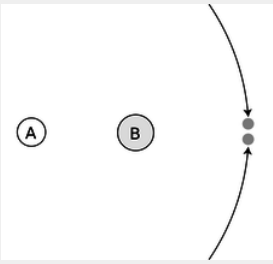


Figure j: Two asteroids collide.

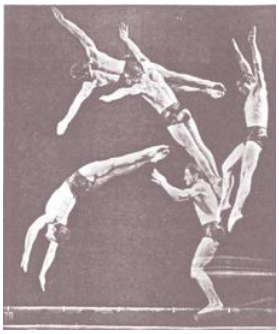
With planet A as the axis, the two asteroids have the same amount of angular momentum, but one has positive angular momentum and the other has negative. Before the collision, the total angular momentum is therefore zero. After the collision, the two asteroids will have stopped moving, and again the total angular momentum is zero. The total angular momentum both before and after the collision is zero, so angular momentum is conserved if you choose planet A as the axis.

The only difference with planet B as axis is that r is smaller by a factor of two, so all the angular momenta are halved. Even though the angular momenta are different than the ones calculated by planet A, angular momentum is still conserved.

The earth spins on its own axis once a day, but simultaneously travels in its circular one-year orbit around the sun, so any given part of it traces out a complicated loopy path. It would seem difficult to calculate the earth's angular momentum, but it turns out that there is an intuitively appealing shortcut: we can simply add up the angular momentum due to its spin plus that arising from its center of mass's circular motion around the sun. This is a special case of the following general theorem:

The spin theorem. An object's angular momentum with respect to some outside axis A can be found by adding up two parts:

1. The first part is the object's angular momentum found by using its own center of mass as the axis, i.e., the angular momentum the object has because it is spinning.
2. The other part equals the angular momentum that the object would have with respect to the axis A if it had all its mass concentrated at and moving with its center of mass.



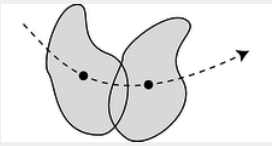
k / Everyone has a strong tendency to think of the diver as rotating about his own center of mass. However, he is flying in an arc, and he also has angular momentum because of this motion.

Example 4: A system with its center of mass at rest

In the special case of an object whose center of mass is at rest, the spin theorem implies that the object's angular momentum is the same regardless of what axis we choose. (This is an even stronger statement than the choice of axis theorem, which only guarantees that angular momentum is conserved for any given choice of axis, without specifying that it is the same for all such choices.)

Example 5: Angular momentum of a rigid object

▷ A motorcycle wheel has almost all its mass concentrated at the outside. If the wheel has mass m and radius r, and the time required for one revolution is T, what is the spin part of its angular momentum?



l / This rigid object has angular momentum both because it is spinning about its center of mass and because it is moving through space.

▷ This is an example of the commonly encountered special case of rigid motion, as opposed to the rotation of a system like a hurricane in which the different parts take different amounts of time to go around. We don't really have to go through a laborious process of adding up contributions from all the many parts of a wheel, because they are all at about the same distance from the axis, and are all moving around the axis at about the same speed. The velocity is all perpendicular to the spokes,

$$\begin{aligned} v_\perp &= (\text{circumference})/T \\ &= 2\pi r/T \end{aligned}$$

and the angular momentum of the wheel about its center is

$$\begin{aligned} L &= mv_\perp r \\ &= m(2\pi r/T)r \\ &= 2\pi mr^2/T. \end{aligned}$$

Note that although the factors of 2π in this expression is peculiar to a wheel with its mass concentrated on the rim, the proportionality to m/T would have been the same for any other rigidly rotating object. Although an object with a noncircular shape does not have a radius, it is also true in general that angular momentum is proportional to the square of the object's size for fixed values of m and T. For instance doubling an object's size doubles both the $v_\perp$ and r factors in the contribution of each of its parts to the total angular momentum, resulting in an overall factor of four increase.

4.1.4 Torque

Force is the rate of transfer of momentum. The corresponding quantity in the case of angular momentum is called torque (rhymes with “fork”). Where force tells us how hard we are pushing or pulling on something, torque indicates how hard we are twisting on it. Torque is represented by the Greek letter tau, τ, and the rate of change of an object's angular momentum equals the total torque acting on it,

$$\tau_{total} = dL/dt.$$

As with force and momentum, it often happens that angular momentum recedes into the background and we focus our interest on the torques. The torque-focused point of view is exemplified by the fact that many scientifically untrained but mechanically apt people know all about torque, but none of them have heard of angular momentum. Car enthusiasts eagerly compare engines' torques, and there is a tool called a torque wrench which allows one to apply a desired amount of torque to a screw and avoid overtightening it.

Torque distinguished from force

Of course a force is necessary in order to create a torque --- you can't twist a screw without pushing on the wrench --- but force and torque are two different things. One distinction between them is direction. We use positive and negative signs to represent forces in the two possible directions along a line. The direction of a torque, however, is clockwise or counterclockwise, not a linear direction.

The other difference between torque and force is a matter of leverage. A given force applied at a door's knob will change the door's angular momentum twice as rapidly as the same force applied halfway between the knob and the hinge. The same amount of force produces different amounts of torque in these two cases.

It's possible to have a zero total torque with a nonzero total force. An airplane with four jet engines would be designed so that their forces are balanced on the left and right. Their forces are all in the same direction, but the clockwise torques of two of the engines are canceled by the counterclockwise torques of the other two, giving zero total torque.

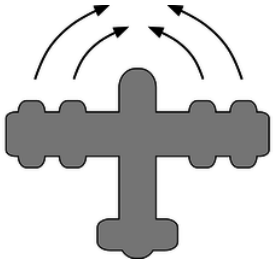


***Figure m**: The plane's four engines produce zero total torque but not zero total force.*

Conversely we can have zero total force and nonzero total torque. A merry-go-round's engine needs to supply a nonzero torque on it to bring it up to speed, but there is zero total force on it. If there was not zero total force on it, its center of mass would accelerate!

Relationship between force and torque

How do we calculate the amount of torque produced by a given force? Since it depends on leverage, we should expect it to depend on the distance between the axis and the point of application of the force. I'll work out an equation relating torque to force for a particular very simple situation, and give a more rigorous derivation on page 284, after developing some mathematical techniques that dramatically shorten and simplify the proof.

Consider a pointlike object which is initially at rest at a distance r from the axis we have chosen for defining angular momentum. We first observe that a force directly inward or outward, along the line connecting the axis to the object, does not impart any angular momentum to the object.

A force perpendicular to the line connecting the axis and the object does, however, make the object pick up angular momentum. Newton's second law gives

$$a = F/m,$$

and using $a = dv/dt$ we find the velocity the object acquires after a time dt,

$$dv = Fdt/m.$$

We're trying to relate force to a change in angular momentum, so we multiply both sides of the equation by mr to give

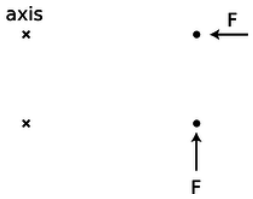
$$\begin{aligned} mdv\,r &= Fdt\,r \\ dL &= Fdt\,r. \end{aligned}$$

Dividing by dt gives the torque:

$$\begin{aligned} \frac{dL}{dt} &= Fr \\ \tau &= Fr. \end{aligned}$$

If a force acts at an angle other than 0 or 90° with respect to the line joining the object and the axis, it would be only the component of the force perpendicular to the line that would produce a torque,

$$\tau = F_\perp r.$$



n / The simple physical situation we use to derive an equation for torque. A force that points directly in at or out away from the axis produces neither clockwise nor counterclockwise angular momentum. A force in the perpendicular direction does transfer angular momentum.

Although this result was proved under a simplified set of circumstances, it is more generally valid:[2]

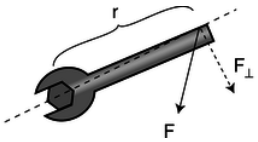
\mythmhdr{Relationship between force and torque} The rate at which a force transfers angular momentum to an object, i.e., the torque produced by the force, is given by

$$|\tau| = r|F_\perp|,$$

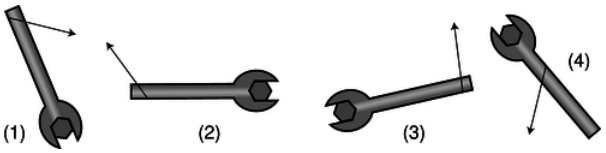
where r is the distance from the axis to the point of application of the force, and $F_\perp$ is the component of the force that is perpendicular to the line joining the axis to the point of application.

The equation is stated with absolute value signs because the positive and negative signs of force and torque indicate different things, so there is no useful relationship between them. The sign of the torque must be found by physical inspection of the case at hand.

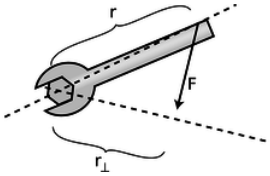
From the equation, we see that the units of torque can be written as newtons multiplied by meters. Metric torque wrenches are calibrated in $\mathrm{N{\cdot}m}$, but American ones use foot-pounds, which is also a unit of distance multiplied by a unit of force. We know from our study of mechanical work that newtons multiplied by meters equal joules, but torque is a completely different quantity from work, and nobody writes torques with units of joules, even though it would be technically correct.



o / The geometric relationships referred to in the relationship between force and torque.



p / Self-check.



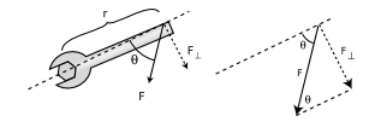
q / Visualizing torque in terms of $r_\perp$.

Exercise 5.1.1

Compare the magnitudes and signs of the four torques shown in figure p. (answer in the back of the PDF version of the book)

Example 6: How torque depends on the direction of the force

▷ How can the torque applied to the wrench in the figure be expressed in terms of r, $|F|$, and the angle θ?
▷ The force vector and its $F_\perp$ component form the hypotenuse and one leg of a right triangle,



and the interior angle opposite to $F_\perp$ equals θ. The absolute value of $F_\perp$ can thus be expressed as

$$F_\perp = |\mathbf{F}| \sin\theta,$$

leading to

$$|\tau| = r|\mathbf{F}| \sin\theta.$$

Sometimes torque can be more neatly visualized in terms of the quantity $r_\perp$ shown in the figure on the left, which gives us a third way of expressing the relationship between torque and force:

$$|\tau| = r_\perp |F|.$$

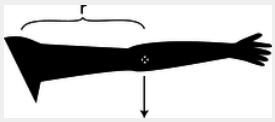
Of course you wouldn't want to go and memorize all three equations for torque. Starting from any one of them you could easily derive the other two using trigonometry. Familiarizing yourself with them can however clue you in to easier avenues of attack on certain problems.

The torque due to gravity

Up until now we've been thinking in terms of a force that acts at a single point on an object, such as the force of your hand on the wrench. This is of course an approximation, and for an extremely realistic calculation of your hand's torque on the wrench you might need to add up the torques exerted by each square millimeter where your skin touches the wrench. This is seldom necessary. But in the case of a gravitational force, there is never any single point at which the force is applied. Our planet is exerting a separate tug on every brick in the Leaning Tower of Pisa, and the total gravitational torque on the tower is the sum of the torques contributed by all the little forces. Luckily there is a trick that allows us to avoid such a massive calculation. It turns out that for purposes of computing the total gravitational torque on an object, you can get the right answer by just pretending that the whole gravitational force acts at the object's center of mass.

Example 7: Gravitational torque on an outstretched arm

▷ Your arm has a mass of 3.0 kg, and its center of mass is 30 cm from your shoulder. What is the gravitational torque on your arm when it is stretched out horizontally to one side, taking the shoulder to be the axis?



r / Example 7.

▷ The total gravitational force acting on your arm is

$$|\mathbf{F}| = (3.0\ \text{kg})(9.8\ \text{m/s}^2) = 29\ \text{N}.$$

For the purpose of calculating the gravitational torque, we can treat the force as if it acted at the arm's center of mass. The force is straight down, which is perpendicular to the line connecting the shoulder to the center of mass, so

$$F_\perp = |\mathbf{F}| = 29\ \text{N}.$$

Continuing to pretend that the force acts at the center of the arm, r equals 30 cm = 0.30 m, so the torque is

$$\tau = r\,F_\perp = 9\ \text{N}\cdot\text{m}.$$

Discussion Questions

◊ This series of discussion questions deals with past students' incorrect reasoning about the following problem.

Suppose a comet is at the point in its orbit shown in the figure. The only force on the comet is the sun's gravitational force. Throughout the question, define all torques and angular momenta using the sun as the axis.

(1) Is the sun producing a nonzero torque on the comet? Explain.

(2) Is the comet's angular momentum increasing, decreasing, or staying the same? Explain.

Explain what is wrong with the following answers. In some cases, the answer is correct, but the reasoning leading up to it is wrong.

(a) Incorrect answer to part (1): “Yes, because the sun is exerting a force on the comet, and the comet is a certain distance from the sun.”

(b) Incorrect answer to part (1): “No, because the torques cancel out.”

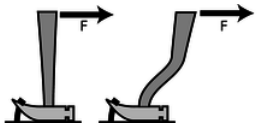
(c) Incorrect answer to part (2): “Increasing, because the comet is speeding up.”

u / Discussion question A.

◊ You whirl a rock over your head on the end of a string, and gradually pull in the string, eventually cutting the radius in half. What happens to the rock's angular momentum? What changes occur in its speed, the time required for one revolution, and its acceleration? Why might the string break?

◊ A helicopter has, in addition to the huge fan blades on top, a smaller propeller mounted on the tail that rotates in a vertical plane. Why?

◊



s / Discussion question D.

Which claw hammer would make it easier to get the nail out of the wood if the same force was applied in the same direction?

◊

t / Discussion question E.

The photo shows an amusement park ride whose two cars rotate in opposite directions. Why is this a good design?

4.1.5 Applications to statics

v / The windmills are not closed systems, but angular momentum is being transferred out of them at the same rate it is transferred in, resulting in constant angular momentum. To get an idea of the huge scale of the modern windmill farm, note the sizes of the trucks and trailers.

In chapter 2 I defined equilibrium as a situation where the interaction energy is minimized. This is the same as a condition of zero total force, or constant momentum. Thus a car is in equilibrium not just when it is parked but also when it is cruising down a straight road with constant momentum.

Likewise there are many cases where a system is not closed but maintains constant angular momentum. When a merry-go-round is running at constant angular momentum, the engine's torque is being canceled by the torque due to friction.

It's not enough for a boat not to sink --- we'd also like to avoid having it capsize. For this reason, we now redefine equilibrium as follows.

When an object has constant momentum and constant angular momentum, we say that it is in equilibrium. Again, this is a scientific redefinition of the common English word, since in ordinary speech nobody would describe a car spinning out on an icy road as being in equilibrium.

Very commonly, however, we are interested in cases where an object is not only in equilibrium but also at rest, and this corresponds more closely to the usual meaning of the word. Statics is the branch of physics concerned with problems such as these.

Solving statics problems is now simply a matter of applying and combining some things you already know:

- You know the behaviors of the various types of forces, for example that a frictional force is always parallel to the surface of contact.

- You know about vector addition of forces. It is the vector sum of the forces that must equal zero to produce equilibrium.

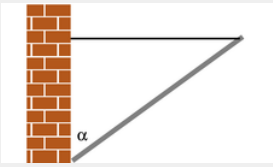
- You know about torque. The total torque acting on an object must be zero if it is to be in equilibrium.

- You know that the choice of axis is arbitrary, so you can make a choice of axis that makes the problem easy to solve.

In general, this type of problem could involve four equations in four unknowns: three equations that say the force components add up to zero, and one equation that says the total torque is zero. Most cases you'll encounter will not be this complicated. In the example below, only the equation for zero total torque is required in order to get an answer.

Example 8: A flagpole

▷ A 10-kg flagpole is being held up by a lightweight horizontal cable, and is propped against the foot of a wall as shown in the figure. If the cable is only capable of supporting a tension of 70 N, how great can the angle α be without breaking the cable?



w / Example 8.

▷ All three objects in the figure are supposed to be in equilibrium: the pole, the cable, and the wall. Whichever of the three objects we pick to investigate, all the forces and torques on it have to cancel out. It is not particularly helpful to analyze the forces and torques on the wall, since it has forces on it from the ground that are not given and that we don't want to find. We could study the forces and torques on the cable, but that doesn't let us use the given information about the pole. The object we need to analyze is the pole.

The pole has three forces on it, each of which may also result in a torque: (1) the gravitational force, (2) the cable's force, and (3) the wall's force.

We are free to define an axis of rotation at any point we wish, and it is helpful to define it to lie at the bottom end of the pole, since by that definition the wall's force on the pole is applied at $r = 0$ and thus makes no torque on the pole. This is good, because we don't know what the wall's force on the pole is, and we are not trying to find it.

With this choice of axis, there are two nonzero torques on the pole, a counterclockwise torque from the cable and a clockwise torque from gravity. Choosing to represent counterclockwise torques as positive numbers, and using the equation $|\boldsymbol{\tau}| = r|F|\sin\theta$, we have

$$r_{cable}|F_{cable}|\sin\theta_{cable} - r_{grav}|F_{grav}|\sin\theta_{grav} = 0.$$

A little geometry gives $\theta_{cable} = 90^\circ - \alpha$ and $\theta_{grav} = \alpha$, so

$$r_{cable}|F_{cable}|\sin(90^\circ - \alpha) - r_{grav}|F_{grav}|\sin\alpha = 0.$$

The gravitational force can be considered as acting at the pole's center of mass, i.e., at its geometrical center, so r_{cable} is twice r_{grav}, and we can simplify the equation to read

$$2|F_{cable}|\sin(90^\circ - \alpha) - |F_{grav}|\sin\alpha = 0.$$

These are all quantities we were given, except for α, which is the angle we want to find. To solve for α we need to use the trig identity $\sin(90^\circ - x) = \cos x$,

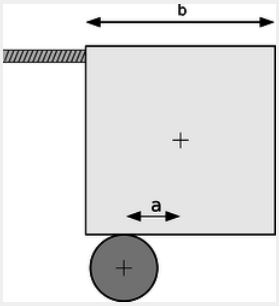
$$2|F_{cable}|\cos\alpha - |F_{grav}|\sin\alpha = 0,$$

which allows us to find

$$\begin{aligned}
\tan\alpha &= 2\frac{|\mathbf{F}_{cable}|}{|\mathbf{F}_{grav}|} \\
\alpha &= \tan^{-1}\left(2\frac{|\mathbf{F}_{cable}|}{|\mathbf{F}_{grav}|}\right) \\
&= \tan^{-1}\left(2\times\frac{70\text{ N}}{98\text{ N}}\right) \\
&= 55^\circ.
\end{aligned}$$

Example 9: Art!

▷ The abstract sculpture shown in figure x contains a cube of mass m and sides of length b. The cube rests on top of a cylinder, which is off-center by a distance a. Find the tension in the cable.



x / Example 9.

▷ There are four forces on the cube: a gravitational force mg, the force F_T from the cable, the upward normal force from the cylinder, F_N, and the horizontal static frictional force from the cylinder, F_s.

The total force on the cube in the vertical direction is zero:

$$F_N - mg = 0.$$

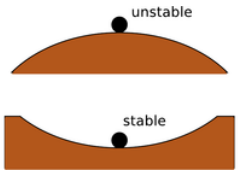
As our axis for defining torques, it's convenient to choose the point of contact between the cube and the cylinder, because then neither F_s nor F_N makes any torque. The cable's torque is counterclockwise, and the torque due to gravity is clockwise. and the cylinder's torque is clockwise. Letting counterclockwise torques be positive, and using the convenient equation $\tau = r_\perp F$, we find the equation for the total torque:

$$bF_T - F_N a = 0.$$

We could also write down the equation saying that the total horizontal force is zero, but that would bring in the cylinder's frictional force on the cube, which we don't know and don't need to find. We already have two equations in the two unknowns F_T and F_N, so there's no need to make it into three equations in three unknowns. Solving the first equation for $F_N = mg$, we then substitute into the second equation to eliminate F_N, and solve for $F_T = (a/b)mg$.

Why is one equilibrium stable and another unstable? Try pushing your own nose to the left or the right. If you push it a millimeter to the left, it responds with a gentle force to the right. If you push it a centimeter to the left, its force on your finger becomes much stronger. The defining characteristic of a stable equilibrium is that the farther the object is moved away from equilibrium, the stronger the force is that tries to bring it back.

The opposite is true for an unstable equilibrium. In the top figure, the ball resting on the round hill theoretically has zero total force on it when it is exactly at the top. But in reality the total force will not be exactly zero, and the ball will begin to move off to one side. Once it has moved, the net force on the ball is greater than it was, and it accelerates more rapidly. In an unstable equilibrium, the farther the object gets from equilibrium, the stronger the force that pushes it farther from equilibrium.



y / Stable and unstable equilibria.

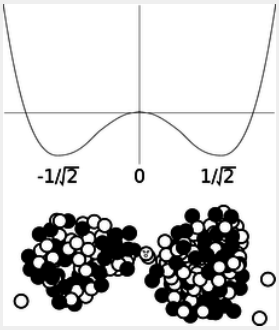
This idea can be rephrased in terms of energy. The difference between the stable and unstable equilibria shown in figure y is that in the stable equilibrium, the energy is at a minimum, and moving to either side of equilibrium will increase it, whereas the unstable equilibrium represents a maximum.

Note that we are using the term “stable” in a weaker sense than in ordinary speech. A domino standing upright is stable in the sense we are using, since it will not spontaneously fall over in response to a sneeze from across the room or the vibration from a passing truck. We would only call it unstable in the technical sense if it could be toppled by *any* force, no matter how small. In everyday usage, of course, it would be considered unstable, since the force required to topple it is so small.

z / The dancer's equilibrium is unstable. If she didn't constantly make tiny adjustments, she would tip over.

Example 10: Application of Calculus

▷ Nancy Neutron is living in a uranium nucleus that is undergoing fission. Nancy's nuclear energy as a function of position can be approximated by $U = x^4 - x^2$, where all the units and numerical constants have been suppressed for simplicity. Use calculus to locate the equilibrium points, and determine whether they are stable or unstable.



aa / Example 10.

▷ The equilibrium points occur where the U is at a minimum or maximum, and minima and maxima occur where the derivative (which equals minus the force on Nancy) is zero. This derivative is $dU/dx = 4x^3 - 2x$, and setting it equal to zero, we have $x = 0, \pm 1/\sqrt{2}$. Minima occur where the second derivative is positive, and maxima where it is negative. The second derivative is $12x^2 - 2$, which is negative at $x = 0$ (unstable) and positive at $x = \pm 1/\sqrt{2}$ (stable). Interpretation: the graph of U is shaped like a rounded letter `W,' with the two troughs representing the two halves of the splitting nucleus. Nancy is going to have to decide which half she wants to go with.

Contributors and Attributions

5.2: Rigid-Body Rotation

4.2.1 Kinematics

When a rigid object rotates, every part of it (every atom) moves in a circle, covering the same angle in the same amount of time, a. Every atom has a different velocity vector, b. Since all the velocities are different, we can't measure the speed of rotation of the top by giving a single velocity. We can, however, specify its speed of rotation consistently in terms of angle per unit time. Let the position of some reference point on the top be denoted by its angle θ, measured in a circle around the axis. For reasons that will become more apparent shortly, we measure all our angles in radians. Then the change in the angular position of any point on the top can be written as $d\theta$, and all parts of the top have the same value of $d\theta$ over a certain time interval dt. We define the angular velocity, ω (Greek omega),

$$\omega = \frac{d\theta}{dt} \tag{5.2.1}$$

$$[\text{definition of angular velocity; } \theta \text{ in units of radians}] \tag{5.2.2}$$

which is similar to, but not the same as, the quantity ω we defined earlier to describe vibrations. The relationship between ω and t is exactly analogous to that between x and t for the motion of a particle through space.

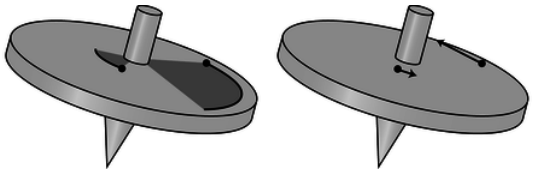


Figure 4.2.1: (left) he two atoms cover the same angle in a given time interval. (right) Their velocity vectors, however, differ in both magnitude and direction.

Exercise 5.2.1

If two different people chose two different reference points on the top in order to define θ=0, how would their θ-t graphs differ? What effect would this have on the angular velocities?

(answer in the back of the PDF version of the book)

The angular velocity has units of radians per second, rad/s. However, radians are not really units at all. The radian measure of an angle is defined, as the length of the circular arc it makes, divided by the radius of the circle. Dividing one length by another gives a unitless quantity, so anything with units of radians is really unitless. We can therefore simplify the units of angular velocity, and call them inverse seconds, s^{-1}.

Example 5.2.1: A 78-rpm record

▷ In the early 20th century, the standard format for music recordings was a plastic disk that held a single song and rotated at 78 rpm (revolutions per minute). What was the angular velocity of such a disk?

▷ If we measure angles in units of revolutions and time in units of minutes, then 78 rpm is the angular velocity. Using standard physics units of radians/second, however, we have

$$\frac{78 \text{ revolutions}}{1 \text{ minute}} \times \frac{2\pi \text{ radians}}{1 \text{ revolution}} \times \frac{1 \text{ minute}}{60 \text{ seconds}} = 8.2 \text{ s}^{-1}. \tag{5.2.3}$$

In the absence of any torque, a rigid body will rotate indefinitely with the same angular velocity. If the angular velocity is changing because of a torque, we define an angular acceleration,

$$\alpha = \frac{d\omega}{dt} \tag{5.2.4}$$

$$[\text{definition of angular acceleration}] \tag{5.2.5}$$

The symbol is the Greek letter alpha. The units of this quantity are $\mathrm{rad/s^2}$, or simply $\mathrm{s^{-2}}$.

The mathematical relationship between ω and θ is the same as the one between v and x, and similarly for α and a. We can thus make a system of analogies, c, and recycle all the familiar kinematic equations for constant-acceleration motion.

x ⟷ θ

v ⟷ ω

a ⟷ α

Figure 4.2.2: *Analogies between rotational and linear quantities.*

Example 5.2.2: The synodic period

Mars takes nearly twice as long as the Earth to complete an orbit. If the two planets are alongside one another on a certain day, then one year later, Earth will be back at the same place, but Mars will have moved on, and it will take more time for Earth to finish catching up. Angular velocities add and subtract, just as velocity vectors do. If the two planets' angular velocities are ω_1 and ω_2, then the angular velocity of one relative to the other is $\omega_1 - \omega_2$. The corresponding period, $1/(1/T_1 - 1/T_2)$ is known as the synodic period.

Example 5.2.3: A neutron star

▷ A neutron star is initially observed to be rotating with an angular velocity of $2.0\ \mathrm{s^{-1}}$, determined via the radio pulses it emits. If its angular acceleration is a constant $-1.0 \times 10^{-8}\ \mathrm{s^{-2}}$, how many rotations will it complete before it stops? (In reality, the angular acceleration is not always constant; sudden changes often occur, and are referred to as “starquakes!”)

▷ The equation $v_f^2 - v_i^2 = 2a\Delta x$ can be translated into $\omega_f^2 - \omega_i^2 = 2\alpha\Delta\theta$, giving

$$\begin{aligned}\Delta\theta &= (\omega_f^2 - \omega_i^2)/2\alpha \\ &= 2.0 \times 10^8 \text{ radians} \\ &= 3.2 \times 10^7 \text{ rotations.}\end{aligned}$$

4.2.2 Relations between angular quantities and motion of a point

It is often necessary to be able to relate the angular quantities to the motion of a particular point on the rotating object. As we develop these, we will encounter the first example where the advantages of radians over degrees become apparent.

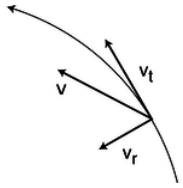


Figure 4.2.3: *We construct a coordinate system that coincides with the location and motion of the moving point of interest at a certain moment.*

The speed at which a point on the object moves depends on both the object's angular velocity ω and the point's distance r from the axis. We adopt a coordinate system, d, with an inward (radial) axis and a tangential axis. The length of the infinitesimal circular arc ds traveled by the point in a time interval dt is related to $d\theta$ by the definition of radian measure, $d\theta = ds/r$, where positive and negative values of ds represent the two possible directions of motion along the tangential axis. We then have $v_t = ds/dt = rd\theta/dt = \omega r$, or

$$v_t = \omega r \tag{5.2.6}$$

$$\text{tangential velocity of a point at a distance } r \text{ from the axis of rotation} \tag{5.2.7}$$

The radial component is zero, since the point is not moving inward or outward,

$$v_r = 0 \tag{5.2.8}$$

$$\text{radial velocity of a point at a distance } r \text{ from the axis of rotation} \tag{5.2.9}$$

Note that we had to use the definition of radian measure in this derivation. Suppose instead we had used units of degrees for our angles and degrees per second for angular velocities. The relationship between $d\theta_{degrees}$ and ds is $d\theta_{degrees} = (360/2\pi)s/r$, where the extra conversion factor of $(360/2\pi)$comes from that fact that there are 360 degrees in a full circle, which is equivalent to 2π radians. The equation for v_t would then have been $v_t = (2\pi/360)(\omega_{degrees\,per\,second})(r)$, which would have been much messier. Simplicity, then, is the reason for using radians rather than degrees; by using radians we avoid infecting all our equations with annoying conversion factors.

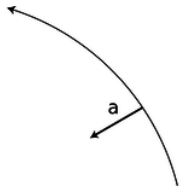


Figure 4.2.4: *Even if the rotating object has zero angular acceleration, every point on it has an acceleration towards the center.*

Since the velocity of a point on the object is directly proportional to the angular velocity, you might expect that its acceleration would be directly proportional to the angular acceleration. This is not true, however. Even if the angular acceleration is zero, i.e., if the object is rotating at constant angular velocity, every point on it will have an acceleration vector directed toward the axis, e. As derived on page 209, the magnitude of this acceleration is

$$a_r = \omega^2 r. \tag{5.2.10}$$

$$\text{radial acceleration of a point at a distance } r \text{ from the axis} \tag{5.2.11}$$

For the tangential component, any change in the angular velocity $d\omega$ will lead to a change $d\omega \cdot r$ in the tangential velocity, so it is easily shown that

$$a_t = \alpha r. \tag{5.2.12}$$

$$\text{tangential acceleration of a point at a distance } r \text{ from the axis} \tag{5.2.13}$$

Exercise 5.2.2

Positive and negative signs of ω represent rotation in opposite directions. Why does it therefore make sense physically that ω is raised to the first power in the equation for v_t and to the second power in the one for a_r?

(answer in the back of the PDF version of the book)

Example 5.2.3: Radial acceleration at the surface of the Earth

▷ What is your radial acceleration due to the rotation of the earth if you are at the equator?

▷ At the equator, your distance from the Earth's rotation axis is the same as the radius of the spherical Earth, 6.4×10^6 m. Your angular velocity is

$$\begin{aligned}\omega &= \frac{2\pi \text{ radians}}{1 \text{ day}} \\ &= 7.3 \times 10^{-5} \text{ s}^{-1},\end{aligned}$$

which gives an acceleration of

$$\begin{aligned}a_r &= \omega^2 r \\ &= 0.034 \text{ m/s}^2.\end{aligned}$$

The angular velocity was a very small number, but the radius was a very big number. Squaring a very small number, however, gives a very very small number, so the ω^2 factor "wins," and the final result is small.

If you're standing on a bathroom scale, this small acceleration is provided by the imbalance between the downward force of gravity and the slightly weaker upward normal force of the scale on your foot. The scale reading is therefore a little lower than it should be.

4.2.3 Dynamics

If we want to connect all this kinematics to anything dynamical, we need to see how it relates to torque and angular momentum. Our strategy will be to tackle angular momentum first, since angular momentum relates to motion, and to use the additive property of angular momentum: the angular momentum of a system of particles equals the sum of the angular momenta of all the individual particles. The angular momentum of one particle within our rigidly rotating object, $L = mv_{\perp}r$, can be rewritten as $L = r\,p\,\sin\,\theta$, where r and p are the magnitudes of the particle's $\mathbf{r}$ and momentum vectors, and θ is the angle between these two vectors. (The $\mathbf{r}$ vector points outward perpendicularly from the axis to the particle's position in space.) In rigid-body rotation the angle θ is 90°, so we have simply $L = rp$. Relating this to angular velocity, we have

$$L = rp = (r)(mv) = (r)(m\omega r) = mr^2\omega. \tag{5.2.14}$$

The particle's contribution to the total angular momentum is proportional to ω, with a proportionality constant mr^2. We refer to mr^2 as the particle's contribution to the object's total *moment of inertia*, I, where "moment" is used in the sense of "important," as in "momentous" --- a bigger value of I tells us the particle is more important for determining the total angular momentum. The total moment of inertia is

$$I = \sum m_i r_i^2 \tag{5.2.15}$$

$$[\text{definition of the moment of inertia; for rigid-body rotation in a plane; } r \text{ is the distance from the axis, measured perpendicular to the axis}] \tag{5.2.16}$$

The angular momentum of a rigidly rotating body is then

$$L = I\omega. \tag{5.2.17}$$

$$[\text{angular momentum of rigid-body rotation in a plane}] \tag{5.2.18}$$

Since torque is defined as dL/dt, and a rigid body has a constant moment of inertia, we have $\tau = dL/dt = Id\omega/dt = I\alpha$,

$$\tau = I\alpha \tag{5.2.19}$$

$$[\text{relationship between torque and angular acceleration for rigid-body rotation in a plane}] \tag{5.2.20}$$

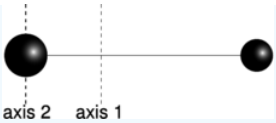
which is analogous to $F = ma$.

The complete system of analogies between linear motion and rigid-body rotation is given in figure f.

x ⟷ θ
v ⟷ ω
a ⟷ α
m ⟷ I
p ⟷ L
F ⟷ τ

Figure 4.2.5: Analogies between rotational and linear quantities.

Example 5.2.3: A barbell



g / Example 15

▷ The barbell shown in figure g consists of two small, dense, massive balls at the ends of a very light rod. The balls have masses of 2.0 kg and 1.0 kg, and the length of the rod is 3.0 m. Find the moment of inertia of the rod (1) for rotation about its center of mass, and (2) for rotation about the center of the more massive ball.

▷ (1) The ball's center of mass lies 1/3 of the way from the greater mass to the lesser mass, i.e., 1.0 m from one and 2.0 m from the other. Since the balls are small, we approximate them as if they were two pointlike particles. The moment of inertia is

$$\begin{aligned} I &= (2.0\ \text{kg})(1.0\ \text{m})^2 + (1.0\ \text{kg})(2.0\ \text{m})^2 \\ &= 2.0\ \text{kg·m}^2 + 4.0\ \text{kg·m}^2 \\ &= 6.0\ \text{kg·m}^2 \end{aligned}$$

Perhaps counterintuitively, the less massive ball contributes far more to the moment of inertia.

(2) The big ball theoretically contributes a little bit to the moment of inertia, since essentially none of its atoms are exactly at r=0. However, since the balls are said to be small and dense, we assume all the big ball's atoms are so close to the axis that we can ignore their small contributions to the total moment of inertia:

$$\begin{aligned} I &= (1.0\ \text{kg})(3.0\ \text{m})^2 \\ &= 9.0\ \text{kg·m}^2 \end{aligned}$$

This example shows that the moment of inertia depends on the choice of axis. For example, it is easier to wiggle a pen about its center than about one end.

Example 16: The parallel axis theorem

▷ Generalizing the previous example, suppose we pick any axis parallel to axis 1, but offset from it by a distance h. Part (2) of the previous example then corresponds to the special case of $h = -1.0$ m (negative being to the left). What is the moment of inertia about this new axis?

▷ The big ball's distance from the new axis is $(1.0\ \text{m})+h$, and the small one's is $(2.0\ \text{m})$-h. The new moment of inertia is

$$\begin{aligned} I &= (2.0\ \text{kg})[(1.0\ \text{m})+h]^2 + (1.0\ \text{kg})[(2.0\ \text{m}) - h]^2 \\ &= 6.0\ \text{kg·m}^2 + (4.0\ \text{kg·m})h - (4.0\ \text{kg·m})h + (3.0\ \text{kg})h^2. \end{aligned}$$

The constant term is the same as the moment of inertia about the center-of-mass axis, the first-order terms cancel out, and the third term is just the total mass multiplied by h^2. The interested reader will have no difficulty in generalizing this to any set of particles (problem 38, p. 294), resulting in the parallel axis theorem: If an object of total mass M rotates about a line at a distance h from its center of mass, then its moment of inertia equals $I_{cm} + Mh^2$, where I_{cm} is the moment of inertia for rotation about a parallel line through the center of mass.

Example 17: Scaling of the moment of inertia

▷ (1) Suppose two objects have the same mass and the same shape, but one is less dense, and larger by a factor k. How do their moments of inertia compare?

(2) What if the densities are equal rather than the masses?

▷ (1) This is like increasing all the distances between atoms by a factor k. All the r's become greater by this factor, so the moment of inertia is increased by a factor of k^2.

(2) This introduces an increase in mass by a factor of k^3, so the moment of inertia of the bigger object is greater by a factor of k^5.

4.2.4 Iterated integrals

In various places in this book, starting with subsection 4.2.5, we'll come across integrals stuck inside other integrals. These are known as iterated integrals, or double integrals, triple integrals, etc. Similar concepts crop up all the time even when you're not doing calculus, so let's start by imagining such an example. Suppose you want to count how many squares there are on a chess board, and you don't know how to multiply eight times eight. You could start from the upper left, count eight squares across, then continue with the second row, and so on, until you how counted every square, giving the result of 64. In slightly more formal mathematical language, we could write the following recipe: for each row, r, from 1 to 8, consider the columns, c, from 1 to 8, and add one to the count for each one of them. Using the sigma notation, this becomes

$$\sum_{r=1}^{8}\sum_{c=1}^{8}1. \tag{5.2.21}$$

If you're familiar with computer programming, then you can think of this as a sum that could be calculated using a loop nested inside another loop. To evaluate the result (again, assuming we don't know how to multiply, so we have to use brute force), we can first evaluate the inside sum, which equals 8, giving

$$\sum_{r=1}^{8}8. \tag{5.2.22}$$

Notice how the “dummy” variable c has disappeared. Finally we do the outside sum, over r, and find the result of 64.

Now imagine doing the same thing with the pixels on a TV screen. The electron beam sweeps across the screen, painting the pixels in each row, one at a time. This is really no different than the example of the chess board, but because the pixels are so small, you normally think of the image on a TV screen as continuous rather than discrete. This is the idea of an integral in calculus. Suppose we want to find the area of a rectangle of width a and height b, and we don't know that we can just multiply to get the area ab. The brute force way to do this is to break up the rectangle into a grid of infinitesimally small squares, each having width dx and height dy, and therefore the infinitesimal area $dA = dxdy$. For convenience, we'll imagine that the rectangle's lower left corner is at the origin. Then the area is given by this integral:

$$\begin{aligned}\text{area} &= \int_{y=0}^{b}\int_{x=0}^{a} dA \\ &= \int_{y=0}^{b}\int_{x=0}^{a} dxdy\end{aligned}$$

Notice how the leftmost integral sign, over y, and the rightmost differential, dy, act like bookends, or the pieces of bread on a sandwich. Inside them, we have the integral sign that runs over x, and the differential dx that matches it on the right. Finally, on the innermost layer, we'd normally have the thing we're integrating, but here's it's 1, so I've omitted it. Writing the lower limits of the integrals with $x =$ and $y =$ helps to keep it straight which integral goes with with differential. The result is

$$\begin{aligned}\text{area} &= \int_{y=0}^{b}\int_{x=0}^{a} dA \\ &= \int_{y=0}^{b}\int_{x=0}^{a} dxdy \\ &= \int_{y=0}^{b}\left(\int_{x=0}^{a} dx\right)dy \\ &= \int_{y=0}^{b} ady \\ &= a\int_{y=0}^{b} dy \\ &= ab.\end{aligned}$$

Example 18: Area of a triangle

▷ Find the area of a 45-45-90 right triangle having legs a.

▷ Let the triangle's hypotenuse run from the origin to the point (a, a), and let its legs run from the origin to $(0, a)$, and then to (a, a). In other words, the triangle sits on top of its hypotenuse. Then the integral can be set up the same way as the one before, but for a particular value of y, values of x only run from 0 (on the y axis) to y (on the hypotenuse). We then have

$$\begin{aligned}\text{area} &= \int_{y=0}^{a}\int_{x=0}^{y} dA \\ &= \int_{y=0}^{a}\int_{x=0}^{y} dxdy \\ &= \int_{y=0}^{a}\left(\int_{x=0}^{y} dx\right)dy \\ &= \int_{y=0}^{a} ydy \\ &= \frac{1}{2}a^2\end{aligned}$$

Note that in this example, because the upper end of the x values depends on the value of y, it makes a difference which order we do the integrals in. The x integral has to be on the inside, and we have to do it first.

Example 19: Volume of a cube

▷ Find the volume of a cube with sides of length a.

▷ This is a three-dimensional example, so we'll have integrals nested three deep, and the thing we're integrating is the volume $dV = dxdydz$.

$$\begin{aligned}\text{volume} &= \int_{z=0}^{a}\int_{y=0}^{a}\int_{x=0}^{a} dxdydz \\ &= \int_{z=0}^{a}\int_{y=0}^{a} adydz \\ &= a\int_{z=0}^{a}\int_{y=0}^{a} dydz \\ &= a\int_{z=0}^{a} adz \\ &= a^3\end{aligned}$$

Example 20: Area of a circle

▷ Find the area of a circle.

▷ To make it easy, let's find the area of a semicircle and then double it. Let the circle's radius be r, and let it be centered on the origin and bounded below by the x axis. Then the curved edge is given by the equation $r^2 = x^2 + y^2$, or $y = \sqrt{r^2 - x^2}$. Since the y integral's limit depends on x, the x integral has to be on the outside. The area is

$$\begin{aligned}\text{area} &= \int_{x=-r}^{r}\int_{y=0}^{\sqrt{r^2-x^2}} dydx \\ &= \int_{x=-r}^{r}\sqrt{r^2-x^2}\,dx \\ &= r\int_{x=-r}^{r}\sqrt{1-(x/r)^2}\,dx.\end{aligned}$$

$$\text{Substituting } u = x/r,$$

$$\text{area} = r^2\int_{u=-1}^{1}\sqrt{1-u^2}\,du$$

The definite integral equals π, as you can find using a trig substitution or simply by looking it up in a table, and the result is, as expected, $\pi r^2/2$ for the area of the semicircle.

4.2.5 Finding moments of inertia by integration

When calculating the moment of inertia of an ordinary-sized object with perhaps 10^{26} atoms, it would be impossible to do an actual sum over atoms, even with the world's fastest supercomputer. Calculus, however, offers a tool, the integral, for breaking a sum down to infinitely many small parts. If we don't worry about the existence of atoms, then we can use an integral to compute a moment of inertia as if the object was smooth and continuous throughout, rather than granular at the atomic level. Of course this granularity typically has a negligible effect on the result unless the object is itself an individual molecule. This subsection consists of three examples of how to do such a computation, at three distinct levels of mathematical complication.

Moment of inertia of a thin rod

What is the moment of inertia of a thin rod of mass M and length L about a line perpendicular to the rod and passing through its center? We generalize the discrete sum

$$I = \sum m_i r_i^2$$

to a continuous one,

$$\begin{aligned} I &= \int r^2 dm \\ &= \int_{-L/2}^{L/2} x^2 \, \frac{M}{L} dx \quad [r = |x|, \text{ so } r^2 = x^2] \\ &= \frac{1}{12} ML^2 \end{aligned}$$

In this example the object was one-dimensional, which made the math simple. The next example shows a strategy that can be used to simplify the math for objects that are three-dimensional, but possess some kind of symmetry.

Moment of inertia of a disk

What is the moment of inertia of a disk of radius b, thickness t, and mass M, for rotation about its central axis?

We break the disk down into concentric circular rings of thickness dr. Since all the mass in a given circular slice has essentially the same value of r (ranging only from r to $r+dr$), the slice's contribution to the total moment of inertia is simply $r^2 dm$. We then have

$$\begin{aligned} I &= \int r^2 dm \\ &= \int r^2 \rho dV, \end{aligned}$$

where $V = \pi b^2 t$ is the total volume, $\rho = M/V = M/\pi b^2 t$ is the density, and the volume of one slice can be calculated as the volume enclosed by its outer surface minus the volume enclosed by its inner surface, $dV = \pi(r+dr)^2 t - \pi r^2 t = 2\pi t r dr$.

$$\begin{aligned} I &= \int_0^b r^2 \frac{M}{\pi b^2 t} \, 2\pi t \, r dr \\ &= \frac{1}{2} Mb^2. \end{aligned}$$

In the most general case where there is no symmetry about the rotation axis, we must use iterated integrals, as discussed in subsection 4.2.4. The example of the disk possessed two types of symmetry with respect to the rotation axis: (1) the disk is the same when rotated through any angle about the axis, and (2) all slices perpendicular to the axis are the same. These two symmetries reduced the number of layers of integrals from three to one. The following example possesses only one symmetry, of type (2), and we simply set it up as a triple integral. You may not have seen multiple integrals yet in a math course. If so, just skim this example.

Moment of inertia of a cube

What is the moment of inertia of a cube of side b, for rotation about an axis that passes through its center and is parallel to four of its faces? Let the origin be at the center of the cube, and let x be the rotation axis.

$$
\begin{aligned}
I &= \int r^2 dm \\
&= \rho \int r^2 dV \\
&= \rho \int_{-b/2}^{b/2} \int_{-b/2}^{b/2} \int_{-b/2}^{b/2} \left(y^2 + z^2\right) dx dy dz \\
&= \rho b \int_{-b/2}^{b/2} \int_{-b/2}^{b/2} \left(y^2 + z^2\right) dy dz
\end{aligned}
$$

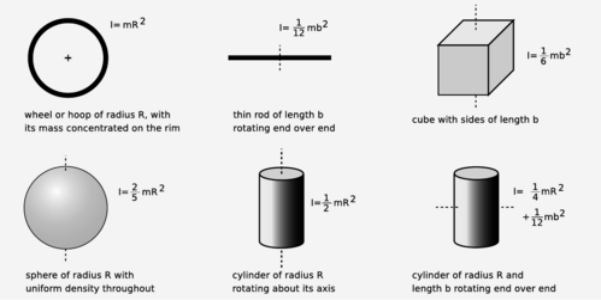
The fact that the last step is a trivial integral results from the symmetry of the problem. The integrand of the remaining double integral breaks down into two terms, each of which depends on only one of the variables, so we break it into two integrals,

$$
I = \rho b \int_{-b/2}^{b/2} \int_{-b/2}^{b/2} y^2 dy dz + \rho b \int_{-b/2}^{b/2} \int_{-b/2}^{b/2} z^2 dy dz \tag{5.2.23}
$$

which we know have identical results. We therefore only need to evaluate one of them and double the result:

$$
\begin{aligned}
I &= 2\rho b \int_{-b/2}^{b/2} \int_{-b/2}^{b/2} z^2 dy d\, z \\
&= 2\rho b^2 \int_{-b/2}^{b/2} z^2 dz \\
&= \frac{1}{6} \rho b^5 \\
&= \frac{1}{6} M b^2
\end{aligned}
$$

Figure h shows the moments of inertia of some shapes, which were evaluated with techniques like these.



h / Moments of inertia of some geometric shapes.

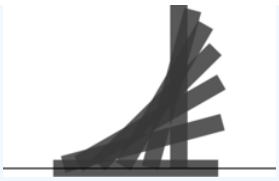
Example 5.2.9: The hammer throw

▷ In the men's Olympic hammer throw, a steel ball of radius 6.1 cm is swung on the end of a wire of length 1.22 m. What fraction of the ball's angular momentum comes from its rotation, as opposed to its motion through space?

▷ It's always important to solve problems symbolically first, and plug in numbers only at the end, so let the radius of the ball be b, and the length of the wire ℓ. If the time the ball takes to go once around the circle is T, then this is also the time it takes to revolve once around its own axis. Its speed is $v = 2\pi\ell/T$, so its angular momentum due to its motion through space is $mv\ell = 2\pi m\ell^2/T$. Its angular momentum due to its rotation around its own center is $(4\pi/5)mb^2/T$. The ratio of these two angular momenta is $(2/5)(b/\ell)^2 = 1.0 \times 10^{-3}$. The angular momentum due to the ball's spin is extremely small.

Example 5.2.10: Toppling a rod

▷ A rod of length b and mass m stands upright. We want to strike the rod at the bottom, causing it to fall and land flat. Find the momentum, p, that should be delivered, in terms of m, b, and g. Can this really be done without having the rod scrape on the floor?



i / Example 22.

▷ This is a nice example of a question that can very nearly be answered based only on units. Since the three variables, m, b, and g, all have different units, they can't be added or subtracted. The only way to combine them mathematically is by multiplication or division. Multiplying one of them by itself is exponentiation, so in general we expect that the answer must be of the form

$$p = Am^j b^k g^l, \tag{5.2.24}$$

where A, j, k, and l are unitless constants. The result has to have units of $\mathrm{kg \cdot m/s}$. To get kilograms to the first power, we need

$$j = 1, \tag{5.2.25}$$

meters to the first power requires

$$k + l = 1, \tag{5.2.26}$$

and seconds to the power -1 implies

$$l = 1/2. \tag{5.2.27}$$

We find $j = 1$, $k = 1/2$, and $l = 1/2$, so the solution must be of the form

$$p = Am\sqrt{bg}. \tag{5.2.28}$$

Note that no physics was required!

Consideration of units, however, won't help us to find the unitless constant A. Let t be the time the rod takes to fall, so that $(1/2)gt^2 = b/2$. If the rod is going to land exactly on its side, then the number of revolutions it completes while in the air must be 1/4, or 3/4, or 5/4, ..., but all the possibilities greater than 1/4 would cause the head of the rod to collide with the floor prematurely. The rod must therefore rotate at a rate that would cause it to complete a full rotation in a time $T = 4t$, and it has angular momentum $L = (\pi/6)mb^2/T$.

The momentum lost by the object striking the rod is p, and by conservation of momentum, this is the amount of momentum, in the horizontal direction, that the rod acquires. In other words, the rod will fly forward a little. However, this has no effect on the solution to the problem. More importantly, the object striking the rod loses angular momentum $bp/2$, which is also transferred to the rod. Equating this to the expression above for L, we find $p = (\pi/12)m\sqrt{bg}$.

Finally, we need to know whether this can really be done without having the foot of the rod scrape on the floor. The figure shows that the answer is no for this rod of finite width, but it appears that the answer would be yes for a sufficiently thin rod. This is analyzed further in homework problem 37 on page 294.

Contributors

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

5.3: Angular Momentum In Three Dimensions

Conservation of angular momentum produces some surprising phenomena when extended to three dimensions. Try the following experiment, for example. Take off your shoe, and toss it in to the air, making it spin along its long (toe-to-heel) axis. You should observe a nice steady pattern of rotation. The same happens when you spin the shoe about its shortest (top-to-bottom) axis. But something unexpected happens when you spin it about its third (left-to-right) axis, which is intermediate in length between the other two. Instead of a steady pattern of rotation, you will observe something more complicated, with the shoe changing its orientation with respect to the rotation axis.

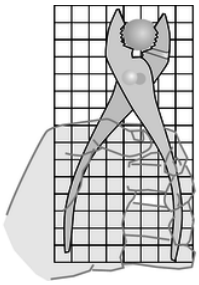
Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

5.4: Footnotes

1. We assume that the door is much more massive than the blob. Under this assumption, the speed at which the door recoils is much less than the original speed of the blob, so the blob has lost essentially all its angular momentum, and given it to the door.
2. A proof is given in example 28 on page 284.

5.E: Conservation of Angular Momentum (Exercises)



a / Problem 1.

1. The figure shows scale drawing of a pair of pliers being used to crack a nut, with an appropriately reduced centimeter grid. Warning: do not attempt this at home; it is bad manners. If the force required to crack the nut is 300 N, estimate the force required of the person's hand. (solution in the pdf version of the book)

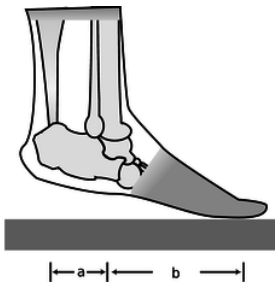
2. You are trying to loosen a stuck bolt on your RV using a big wrench that is 50 cm long. If you hang from the wrench, and your mass is 55 kg, what is the maximum torque you can exert on the bolt? (answer check available at lightandmatter.com)

3. A physical therapist wants her patient to rehabilitate his injured elbow by laying his arm flat on a table, and then lifting a 2.1 kg mass by bending his elbow. In this situation, the weight is 33 cm from his elbow. He calls her back, complaining that it hurts him to grasp the weight. He asks if he can strap a bigger weight onto his arm, only 17 cm from his elbow. How much mass should she tell him to use so that he will be exerting the same torque? (He is raising his forearm itself, as well as the weight.) (answer check available at lightandmatter.com)

4. An object thrown straight up in the air is momentarily at rest when it reaches the top of its motion. Does that mean that it is in equilibrium at that point? Explain.

5. An object is observed to have constant angular momentum. Can you conclude that no torques are acting on it? Explain. [Based on a problem by Serway and Faughn.]

{C}{C}

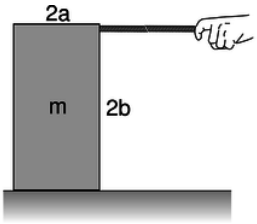


b / Problem 6.

6. A person of mass m stands on the ball of one foot. Find the tension in the calf muscle and the force exerted by the shinbones on the bones of the foot, in terms of m, g, a, and b. For simplicity, assume that all the forces are at 90-degree angles to the foot, i.e., neglect the angle between the foot and the floor.(answer check available at lightandmatter.com)

7. Two pointlike particles have the same momentum vector. Can you conclude that their angular momenta are the same? Explain. [Based on a problem by Serway and Faughn.]

{C}{C}

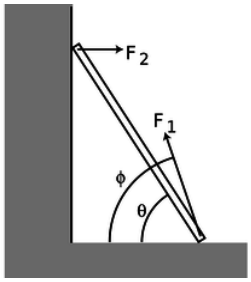


c / Problem 8.

8. The box shown in the figure is being accelerated by pulling on it with the rope.
(a) Assume the floor is frictionless. What is the maximum force that can be applied without causing the box to tip over? \hwhint{hwhint:tipbox}(answer check available at lightandmatter.com)
(b) Repeat part a, but now let the coefficient of friction be μ.(answer check available at lightandmatter.com)
(c) What happens to your answer to part b when the box is sufficiently tall? How do you interpret this?

9. A uniform ladder of mass m and length ℓ leans against a smooth wall, making an angle θ with respect to the ground. The dirt exerts a normal force and a frictional force on the ladder, producing a force vector with magnitude F_1 at an angle ϕ with respect to the ground. Since the wall is smooth, it exerts only a normal force on the ladder; let its magnitude be F_2.
(a) Explain why ϕ must be greater than θ. No math is needed.
(b) Choose any numerical values you like for m and ℓ, and show that the ladder can be in equilibrium (zero torque and zero total force vector) for θ=45.00° and ϕ=63.43°.

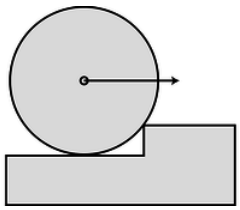
{C}{C}



d / Problems 9 and 10.

10. Continuing problem 9, find an equation for ϕ in terms of θ, and show that m and L do not enter into the equation. Do not assume any numerical values for any of the variables. You will need the trig identity $\sin(a-b) = \sin a \cos b - \sin b \cos a$. (As a numerical check on your result, you may wish to check that the angles given in problem 9b satisfy your equation.)(answer check available at lightandmatter.com)

{C}{C}



e / Problem 11.

11. (a) Find the minimum horizontal force which, applied at the axle, will pull a wheel over a step. Invent algebra symbols for whatever quantities you find to be relevant, and give your answer in symbolic form.
(b) Under what circumstances does your result become infinite? Give a physical interpretation. What happens to your answer when the height of the curb is zero? Does this make sense? \hwhint{hwhint:wheeloverstep}

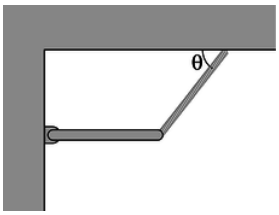
12. A ball is connected by a string to a vertical post. The ball is set in horizontal motion so that it starts winding the string around the post. Assume that the motion is confined to a horizontal plane, i.e., ignore gravity. Michelle and Astrid are trying to predict the final velocity of the ball when it reaches the post. Michelle says that according to conservation of angular momentum, the ball has

to speed up as it approaches the post. Astrid says that according to conservation of energy, the ball has to keep a constant speed. Who is right? [Hint: How is this different from the case where you whirl a rock in a circle on a string and gradually reel in the string?]

13. In the 1950's, serious articles began appearing in magazines like *Life* predicting that world domination would be achieved by the nation that could put nuclear bombs in orbiting space stations, from which they could be dropped at will. In fact it can be quite difficult to get an orbiting object to come down. Let the object have energy $E = K + U$ and angular momentum L. Assume that the energy is negative, i.e., the object is moving at less than escape velocity. Show that it can never reach a radius less than

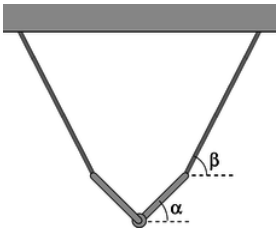
$$r_{min} = \frac{GMm}{2E}\left(-1 + \sqrt{1 + \frac{2EL^2}{G^2M^2m^3}}\right).$$

[Note that both factors are negative, giving a positive result.]



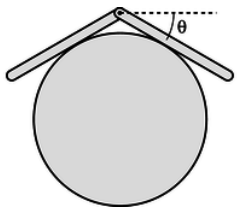
f / Problem 14.

14. (a) The bar of mass m is attached at the wall with a hinge, and is supported on the right by a massless cable. Find the tension, T, in the cable in terms of the angle θ.(answer check available at lightandmatter.com)
(b) Interpreting your answer to part a, what would be the best angle to use if we wanted to minimize the strain on the cable?
(c) Again interpreting your answer to part a, for what angles does the result misbehave mathematically? Interpet this physically.



g / Problem 15.

15. (a) The two identical rods are attached to one another with a hinge, and are supported by the two massless cables. Find the angle α in terms of the angle β, and show that the result is a purely geometric one, independent of the other variables involved. (answer check available at lightandmatter.com)
(b) Using your answer to part a, sketch the configurations for $\beta \to 0$, $\beta = 45^\circ$, and $\beta = 90^\circ$. Do your results make sense intuitively?

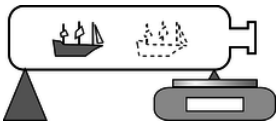


h / Problem 16.

16. Two bars of length ℓ are connected with a hinge and placed on a frictionless cylinder of radius r. (a) Show that the angle θ shown in the figure is related to the unitless ratio r/ℓ by the equation

$$\frac{r}{\ell} = \frac{\cos^2\theta}{2\,\tan\theta}.$$

(b) Discuss the physical behavior of this equation for very large and very small values of r/ℓ.



i / Problem 17.

17. You wish to determine the mass of a ship in a bottle without taking it out. Show that this can be done with the setup shown in the figure, with a scale supporting the bottle at one end, provided that it is possible to take readings with the ship slid to several different locations. Note that you can't determine the position of the ship's center of mass just by looking at it, and likewise for the bottle. In particular, you can't just say, “position the ship right on top of the fulcrum” or “position it right on top of the balance.”

18. Suppose that we lived in a universe in which Newton's law of gravity gave an interaction energy proportional to r^{-6}, rather than r^{-1}. Which, if any, of Kepler's laws would still be true? Which would be completely false? Which would be different, but in a way that could be calculated with straightforward algebra?

19. Use analogies to find the equivalents of the following equations for rotation in a plane:

$$\begin{aligned} KE &= p^2/2m \\ \Delta x &= v_0\Delta t + (1/2)a\Delta t^2 \\ W &= F\Delta x \end{aligned}$$

Example: $v = \Delta x/\Delta t \rightarrow \omega = \Delta\theta/\Delta t$

20. For a one-dimensional harmonic oscillator, the solution to the energy conservation equation,

$$U + K = \frac{1}{2}kx^2 + \frac{1}{2}mv^2 = \text{constant},$$

is an oscillation with frequency $\omega = \sqrt{k/m}$.

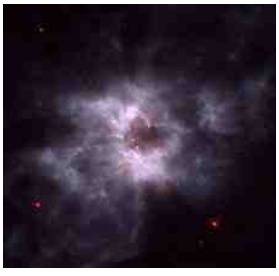
Now consider an analogous system consisting of a bar magnet hung from a thread, which acts like a magnetic compass. A normal compass is full of water, so its oscillations are strongly damped, but the magnet-on-a-thread compass has very little friction, and will oscillate repeatedly around its equilibrium direction. The magnetic energy of the bar magnet is

$$U = -Bm\cos\theta,$$

where B is a constant that measures the strength of the earth's magnetic field, m is a constant that parametrizes the strength of the magnet, and θ is the angle, measured in radians, between the bar magnet and magnetic north. The equilibrium occurs at $\theta = 0$, which is the minimum of U.

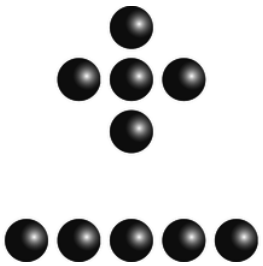
(a) Problem 19 on p. 291 gave some examples of how to construct analogies between rotational and linear motion. Using the same technique, translate the equation defining the linear quantity k to one that defines an analogous angular one κ (Greek letter kappa). Applying this to the present example, find an expression for κ. (Assume the thread is so thin that its stiffness does not have any significant effect compared to earth's magnetic field.)
(b) Find the frequency of the compass's vibrations.

21. (a) Find the angular velocities of the earth's rotation and of the earth's motion around the sun.(answer check available at lightandmatter.com)
(b) Which motion involves the greater acceleration?



j / Problem 22.

22. The sun turns on its axis once every 26.0 days. Its mass is 2.0×10^{30} kg and its radius is 7.0×10^{8} m. Assume it is a rigid sphere of uniform density.
(a) What is the sun's angular momentum? (answer check available at lightandmatter.com)
In a few billion years, astrophysicists predict that the sun will use up all its sources of nuclear energy, and will collapse into a ball of exotic, dense matter known as a white dwarf. Assume that its radius becomes 5.8×10^{6} m (similar to the size of the Earth.) Assume it does not lose any mass between now and then. (Don't be fooled by the photo, which makes it look like nearly all of the star was thrown off by the explosion. The visually prominent gas cloud is actually thinner than the best laboratory vacuum ever produced on earth. Certainly a little bit of mass is actually lost, but it is not at all unreasonable to make an approximation of zero loss of mass as we are doing.)
(b) What will its angular momentum be?
(c) How long will it take to turn once on its axis? (answer check available at lightandmatter.com)



k / Problem 23

23. Give a numerical comparison of the two molecules' moments of inertia for rotation in the plane of the page about their centers of mass.

24. A yo-yo of total mass m consists of two solid cylinders of radius R, connected by a small spindle of negligible mass and radius r. The top of the string is held motionless while the string unrolls from the spindle. Show that the acceleration of the yo-yo is $g/(1+R^2/2r^2)$.

25. Show that a sphere of radius R that is rolling without slipping has angular momentum and momentum in the ratio $L/p = (2/5)R$.

26. Suppose a bowling ball is initially thrown so that it has no angular momentum at all, i.e., it is initially just sliding down the lane. Eventually kinetic friction will bring its angular velocity up to the point where it is rolling without slipping. Show that the final velocity of the ball equals 5/7 of its initial velocity. You'll need the result of problem 25.

27. Find the angular momentum of a particle whose position is $\mathbf{r} = 3\hat{\mathbf{x}} - \hat{\mathbf{y}} + \hat{\mathbf{z}}$ (in meters) and whose momentum is $\mathbf{p} = -2\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}}$ (in kg·m/s). (answer check available at lightandmatter.com)

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

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

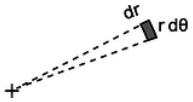
(answer check available at lightandmatter.com)

29. Prove property (3) of the vector cross product from the theorem on page 912.

30. 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.

31. 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 $(\mathbf{A} \times \mathbf{B}) \times \mathbf{C}$.

32. 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) $(\mathbf{A} \cdot \mathbf{B}) \times \mathbf{C}$



l / Problem 33

33. (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 $\theta+d\theta$.(answer check available at lightandmatter.com)
(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.(answer check available at lightandmatter.com)
(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.(answer check available at lightandmatter.com)

34. 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.(answer check available at lightandmatter.com)

35. The nucleus ^{168}Er (erbium-168) contains 68 protons (which is what makes it a nucleus of the element erbium) and 100 neutrons. It has an ellipsoidal shape like an American football, with one long axis and two short axes that are of equal diameter. Because this is a subatomic system, consisting of only 168 particles, its behavior shows some clear quantum-mechanical properties. It can only have certain energy levels, and it makes quantum leaps between these levels. Also, its angular momentum can only have certain values, which are all multiples of 2.109×10^{-34} $\mathrm{kg\cdot m^2/s}$. The table shows some of the observed angular momenta and energies of ^{168}Er, in SI units ($\mathrm{kg\cdot m^2/s}$ and joules).

$L\times10^{34}$	$E\times10^{14}$
0	0
2.109	1.2786
4.218	4.2311
6.327	8.7919
8.437	14.8731
10.546	22.3798
12.655	31.135
14.764	41.206
16.873	52.223

(a) These data can be described to a good approximation as a rigid end-over-end rotation. Estimate a single best-fit value for the moment of inertia from the data, and check how well the data agree with the assumption of rigid-body rotation.\hwhint{hwhint:erbium}(answer check available at lightandmatter.com)
(b) Check whether this moment of inertia is on the right order of magnitude. The moment of inertia depends on both the size and the shape of the nucleus. For the sake of this rough check, ignore the fact that the nucleus is not quite spherical. To estimate its size, use the fact that a neutron or proton has a volume of about 1 fm^3 (one cubic femtometer, where $1\ \mathrm{fm}=10^{-15}\ \mathrm{m}$), and assume they are closely packed in the nucleus.

36. (a) Prove the identity $\mathbf{a}\times(\mathbf{b}\times\mathbf{c})=\mathbf{b}(\mathbf{a}\cdot\mathbf{c})-\mathbf{c}(\mathbf{a}\cdot\mathbf{b})$ by expanding the product in terms of its components. Note that because the x, y, and z components are treated symmetrically in the definitions of the vector cross product, it is only necessary to carry out the proof for the x component of the result.
(b) Applying this to the angular momentum of a rigidly rotating body, $L=\int\mathbf{r}\times(\boldsymbol{\omega}\times\mathbf{r})dm$, show that the diagonal elements of the moment of inertia tensor can be expressed as, e.g., $I_{xx}=\int(y^2+z^2)dm$.
(c) Find the diagonal elements of the moment of inertia matrix of an ellipsoid with axes of lengths a, b, and c, in the principal-axis frame, and with the axis at the center.(answer check available at lightandmatter.com)

37. In example 22 on page 276, prove that if the rod is sufficiently thin, it can be toppled without scraping on the floor. (solution in the pdf version of the book)

38. Suppose an object has mass m, and moment of inertia I_{o} for rotation about some axis A passing through its center of mass. Prove that for an axis B, parallel to A and lying at a distance h from it, the object's moment of inertia is given by $I_{\text{o}} + mh^2$. This is known as the parallel axis theorem.

39. 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 $\frac{m}{18}(A^2+B^2-\mathbf{A}\cdot\mathbf{B})$. Hint: Combine the results of parts a and b with the result of problem 38.

40. 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. (answer check available at lightandmatter.com)

41. When we release an object such as a bicycle wheel or a coin on an inclined plane, we can observe a variety of different behaviors. Characterize these behaviors empirically and try to list the physical parameters that determine which behavior occurs. Try to form a conjecture about the behavior using simple closed-form expressions. Test your conjecture experimentally.

Contributors and Attributions

CHAPTER OVERVIEW

6: Thermodynamics

$S = k \log W$ -- *Inscription on the tomb of Ludwig Boltzmann, 1844-1906. Boltzmann, who originated the microscopic theory of thermodynamics, was driven to suicide by the criticism of his peers, who thought that physical theories shouldn't discuss purely hypothetical objects like atoms.*

In a developing country like China, a refrigerator is the mark of a family that has arrived in the middle class, and a car is the ultimate symbol of wealth. Both of these are *heat engines*: devices for converting between heat and other forms of energy. Unfortunately for the Chinese, neither is a very efficient device. Burning fossil fuels has made China's big cities the most polluted on the planet, and the country's total energy supply isn't sufficient to support American levels of energy consumption by more than a small fraction of China's population. Could we somehow manipulate energy in a more efficient way?

Conservation of energy is a statement that the total amount of energy is constant at all times, which encourages us to believe that any energy transformation can be undone --- indeed, the laws of physics you've learned so far don't even distinguish the past from the future. If you get in a car and drive around the block, the net effect is to consume some of the energy you paid for at the gas station, using it to heat the neighborhood. There would not seem to be any fundamental physical principle to prevent you from recapturing all that heat and using it again the next time you want to go for a drive. More modestly, why don't engineers design a car engine so that it recaptures the heat energy that would otherwise be wasted via the radiator and the exhaust?

Hard experience, however, has shown that designers of more and more efficient engines run into a brick wall at a certain point. The generators that the electric company uses to produce energy at an oil-fueled plant are indeed much more efficient than a car engine, but even if one is willing to accept a device that is very large, expensive, and complex, it turns out to be impossible to make a perfectly efficient heat engine --- not just impossible with present-day technology, but impossible due to a set of fundamental physical principles known as the science of *thermodynamics*. And thermodynamics isn't just a pesky set of constraints on heat engines. Without thermodynamics, there is no way to explain the direction of time's arrow --- why we can remember the past but not the future, and why it's easier to break Humpty Dumpty than to put him back together again.

Contributors and Attributions

Benjamin Crowell (Fullerton College).

6.1: Pressure and Temperature

When we heat an object, we speed up the mind-bogglingly complex random motion of its molecules. One method for taming complexity is the conservation laws, since they tell us that certain things must remain constant regardless of what process is going on. Indeed, the law of conservation of energy is also known as the first law of thermodynamics.

But as alluded to in the introduction to this chapter, conservation of energy by itself is not powerful enough to explain certain empirical facts about heat. A second way to sidestep the complexity of heat is to ignore heat's atomic nature and concentrate on quantities like temperature and pressure that tell us about a system's properties as a whole. This approach is called macroscopic in contrast to the microscopic method of attack. Pressure and temperature were fairly well understood in the age of Newton and Galileo, hundreds of years before there was any firm evidence that atoms and molecules even existed.

Unlike the conserved quantities such as mass, energy, momentum, and angular momentum, neither pressure nor temperature is additive. Two cups of coffee have twice the heat energy of a single cup, but they do not have twice the temperature. Likewise, the painful pressure on your eardrums at the bottom of a pool is not affected if you insert or remove a partition between the two halves of the pool.

We restrict ourselves to a discussion of pressure in fluids at rest and in equilibrium. In physics, the term “fluid” is used to mean either a gas or a liquid. The important feature of a fluid can be demonstrated by comparing with a cube of jello on a plate. The jello is a solid. If you shake the plate from side to side, the jello will respond by shearing, i.e., by slanting its sides, but it will tend to spring back into its original shape. A solid can sustain shear forces, but a fluid cannot. A fluid does not resist a change in shape unless it involves a change in volume.

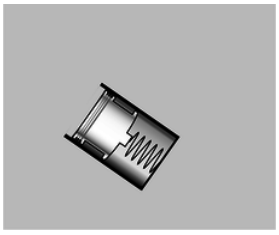
5.1.1 Pressure

If you're at the bottom of a pool, you can't relieve the pain in your ears by turning your head. The water's force on your eardrum is always the same, and is always perpendicular to the surface where the eardrum contacts the water. If your ear is on the east side of your head, the water's force is to the west. If you keep your ear in the same spot while turning around so your ear is on the north, the force will still be the same in magnitude, and it will change its direction so that it is still perpendicular to the eardrum: south. This shows that pressure has no direction in space, i.e., it is a scalar. The direction of the force is determined by the orientation of the surface on which the pressure acts, not by the pressure itself. A fluid flowing over a surface can also exert frictional forces, which are parallel to the surface, but the present discussion is restricted to fluids at rest.

Experiments also show that a fluid's force on a surface is proportional to the surface area. The vast force of the water behind a dam, for example, in proportion to the dam's great surface area. (The bottom of the dam experiences a higher proportion of its force.)

Based on these experimental results, it appears that the useful way to define pressure is as follows. The pressure of a fluid at a given point is defined as $F_\perp/A$, where A is the area of a small surface inserted in the fluid at that point, and $F_\perp$ is the component of the fluid's force on the surface which is perpendicular to the surface. (In the case of a moving fluid, fluid friction forces can act parallel to the surface, but we're only dealing with stationary fluids, so there is only an $F_\perp$.)

This is essentially how a pressure gauge works. The reason that the surface must be small is so that there will not be any significant difference in pressure between one part of it and another part. The SI units of pressure are evidently $\mathrm{N/m^2}$, and this combination can be abbreviated as the pascal, 1 Pa=1 $\mathrm{N/m^2}$. The pascal turns out to be an inconveniently small unit, so car tires, for example, normally have pressures imprinted on them in units of kilopascals.



a / A simple pressure gauge consists of a cylinder open at one end, with a piston and a spring inside. The depth to which the spring is depressed is a measure of the pressure. To determine the absolute pressure, the air needs to be pumped out of the interior of the gauge, so that there is no air pressure acting outward on the piston. In many practical gauges, the back of the piston is open to the atmosphere, so the pressure the gauge registers equals the pressure of the fluid minus the pressure of the atmosphere.

Example 1: Pressure in U.S. units

In U.S. units, the unit of force is the pound, and the unit of distance is the inch. The unit of pressure is therefore pounds per square inch, or p.s.i. (Note that the pound is not a unit of mass.)

Example 2: Atmospheric pressure in U.S. and metric units

▷ A figure that many people in the U.S. remember is that atmospheric pressure is about 15 pounds per square inch. What is this in metric units?

▷

$$\begin{aligned}(15\,\text{lb})/(1\,\text{in}^2) &= \frac{68\,\text{N}}{(0.0254\,\text{m})^2} \\ &= 1.0\times10^5\,\text{N/m}^2 \\ &= 100\,\text{kPa}\end{aligned}$$

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

6.2: Microscopic Description of An Ideal Gas

5.2.1 Evidence for the kinetic theory

Why does matter have the thermal properties it does? The basic answer must come from the fact that matter is made of atoms. How, then, do the atoms give rise to the bulk properties we observe? Gases, whose thermal properties are so simple, offer the best chance for us to construct a simple connection between the microscopic and macroscopic worlds.

A crucial observation is that although solids and liquids are nearly incompressible, gases can be compressed, as when we increase the amount of air in a car's tire while hardly increasing its volume at all. This makes us suspect that the atoms in a solid are packed shoulder to shoulder, while a gas is mostly vacuum, with large spaces between molecules. Most liquids and solids have densities about 1000 times greater than most gases, so evidently each molecule in a gas is separated from its nearest neighbors by a space something like 10 times the size of the molecules themselves.

If gas molecules have nothing but empty space between them, why don't the molecules in the room around you just fall to the floor? The only possible answer is that they are in rapid motion, continually rebounding from the walls, floor and ceiling. In section 2.4 I have already given some of the evidence for the kinetic theory of heat, which states that heat is the kinetic energy of randomly moving molecules. This theory was proposed by Daniel Bernoulli in 1738, and met with considerable opposition because it seemed as though the molecules in a gas would eventually calm down and settle into a thin film on the floor. There was no precedent for this kind of perpetual motion. No rubber ball, however elastic, rebounds from a wall with exactly as much energy as it originally had, nor do we ever observe a collision between balls in which none of the kinetic energy at all is converted to heat and sound. The analogy is a false one, however. A rubber ball consists of atoms, and when it is heated in a collision, the heat is a form of motion of those atoms. An individual molecule, however, cannot possess heat. Likewise sound is a form of bulk motion of molecules, so colliding molecules in a gas cannot convert their kinetic energy to sound. Molecules can indeed induce vibrations such as sound waves when they strike the walls of a container, but the vibrations of the walls are just as likely to impart energy to a gas molecule as to take energy from it. Indeed, this kind of exchange of energy is the mechanism by which the temperatures of the gas and its container become equilibrated.

5.2.2 Pressure, volume, and temperature

A gas exerts pressure on the walls of its container, and in the kinetic theory we interpret this apparently constant pressure as the averaged-out result of vast numbers of collisions occurring every second between the gas molecules and the walls. The empirical facts about gases can be summarized by the relation

$$PV \propto nT, [\text{ideal gas}] \tag{6.2.1}$$

which really only holds exactly for an ideal gas. Here n is the number of molecules in the sample of gas.

Example 6.2.1: Volume related to temperature

The proportionality of volume to temperature at fixed pressure was the basis for our definition of temperature.

Example 8: Pressure related to temperature

Pressure is proportional to temperature when volume is held constant. An example is the increase in pressure in a car's tires when the car has been driven on the freeway for a while and the tires and air have become hot.

We now connect these empirical facts to the kinetic theory of a classical ideal gas. For simplicity, we assume that the gas is monoatomic (i.e., each molecule has only one atom), and that it is confined to a cubical box of volume V, with L being the length of each edge and A the area of any wall. An atom whose velocity has an x component v_x will collide regularly with the left-hand wall, traveling a distance $2L$ parallel to the x axis between collisions with that wall. The time between collisions is $\Delta t = 2L/v_x$, and in each collision the x component of the atom's momentum is reversed from $-mv_x$ to mv_x. The total force on the wall is

$$F = \sum \frac{\Delta p_{x,i}}{\Delta t_i} [\text{monoatomic ideal gas}], \tag{6.2.2}$$

where the index i refers to the individual atoms. Substituting $\Delta p_{x,i} = 2mv_{x,i}$ and $\Delta t_i = 2L/v_{x,i}$, we have

$$F = \sum \frac{mv_{x,i}^2}{L} [\text{monoatomic ideal gas}]. \quad (6.2.3)$$

The quantity $mv_{x,i}^2$ is twice the contribution to the kinetic energy from the part of the atoms' center of mass motion that is parallel to the x axis. Since we're assuming a monoatomic gas, center of mass motion is the only type of motion that gives rise to kinetic energy. (A more complex molecule could rotate and vibrate as well.) If the quantity inside the sum included the y and z components, it would be twice the total kinetic energy of all the molecules. Since we expect the energy to be equally shared among x, y, and z motion,[1] the quantity inside the sum must therefore equal 2/3 of the total kinetic energy, so

$$F = \frac{2K_{total}}{3L} [\text{monoatomic ideal gas}]. \quad (6.2.4)$$

Dividing by A and using $AL = V$, we have

$$P = \frac{2K_{total}}{3V} [\text{monoatomic ideal gas}]. \quad (6.2.5)$$

This can be connected to the empirical relation $PV \propto nT$ if we multiply by V on both sides and rewrite K_{total} as $n\bar{K}$, where $\bar{K}$ is the average kinetic energy per molecule:

$$PV = \frac{2}{3} n\bar{K} [\text{monoatomic ideal gas}]. \quad (6.2.6)$$

For the first time we have an interpretation of temperature based on a microscopic description of matter: in a monoatomic ideal gas, the temperature is a measure of the average kinetic energy per molecule. The proportionality between the two is $\bar{K} = (3/2)kT$, where the constant of proportionality k, known as Boltzmann's constant, has a numerical value of 1.38×10^{-23} J/K. In terms of Boltzmann's constant, the relationship among the bulk quantities for an ideal gas becomes

$$PV = nkT, [\text{ideal gas}] \quad (6.2.7)$$

which is known as the ideal gas law. Although I won't prove it here, this equation applies to all ideal gases, even though the derivation assumed a monoatomic ideal gas in a cubical box. (You may have seen it written elsewhere as $PV = NRT$, where $N = n/N_A$ is the number of moles of atoms, $R = kN_A$, and $N_A = 6.0 \times 10^{23}$, called Avogadro's number, is essentially the number of hydrogen atoms in 1 g of hydrogen.)

Example 6.2.2: Pressure in a car tire

▷ After driving on the freeway for a while, the air in your car's tires heats up from 10°C to 35°C. How much does the pressure increase?

▷ The tires may expand a little, but we assume this effect is small, so the volume is nearly constant. From the ideal gas law, the ratio of the pressures is the same as the ratio of the absolute temperatures,

$$\begin{aligned} P_2/P_1 &= T_2/T_1 \\ &= (308 \text{ K})/(283 \text{ K}) \\ &= 1.09, \end{aligned}$$

or a 9% increase.

Discussion Questions

◊ Compare the amount of energy needed to heat 1 liter of helium by 1 degree with the energy needed to heat 1 liter of xenon. In both cases, the heating is carried out in a sealed vessel that doesn't allow the gas to expand. (The vessel is also well insulated.)

◊ Repeat discussion question A if the comparison is 1 kg of helium versus 1 kg of xenon (equal masses, rather than equal volumes).

◊ Repeat discussion question A, but now compare 1 liter of helium in a vessel of constant volume with the same amount of helium in a vessel that allows expansion beyond the initial volume of 1 liter. (This could be a piston, or a balloon.)

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

6.3: Entropy As a Macroscopic Quantity

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

6.4: Entropy As a Microscopic Quantity

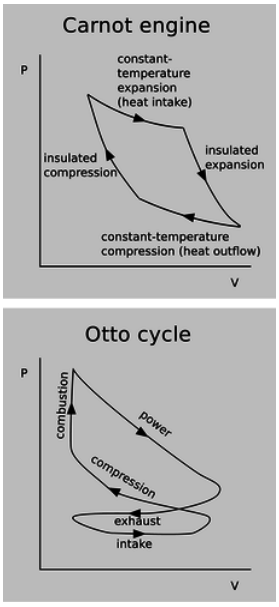
Contributors and Attributions

6.5: More About Heat Engines

So far, the only heat engine we've discussed in any detail has been a fictitious Carnot engine, with a monoatomic ideal gas as its working gas. As a more realistic example, figure l shows one full cycle of a cylinder in a standard gas-burning automobile engine. This four-stroke cycle is called the Otto cycle, after its inventor, German engineer Nikolaus Otto. The Otto cycle is more complicated than a Carnot cycle, in a number of ways:

- The working gas is physically pumped in and out of the cylinder through valves, rather than being sealed and reused indefinitely as in the Carnot engine.
- The cylinders are not perfectly insulated from the engine block, so heat energy is lost from each cylinder by conduction. This makes the engine less efficient that a Carnot engine, because heat is being discharged at a temperature that is not as cool as the environment.
- Rather than being heated by contact with an external heat reservoir, the air-gas mixture inside each cylinder is heated by internal combusion: a spark from a spark plug burns the gasoline, releasing heat.
- The working gas is not monoatomic. Air consists of diatomic molecules (N_2 and O_2), and gasoline of polyatomic molecules such as octane (C_8H_{18}).
- The working gas is not ideal. An ideal gas is one in which the molecules never interact with one another, but only with the walls of the vessel, when they collide with it. In a car engine, the molecules are interacting very dramatically with one another when the air-gas mixture explodes (and less dramatically at other times as well, since, for example, the gasoline may be in the form of microscopic droplets rather than individual molecules).

This is all extremely complicated, and it would be nice to have some way of understanding and visualizing the important properties of such a heat engine without trying to handle every detail at once. A good method of doing this is a type of graph known as a P-V diagram. As proved in homework problem 2, the equation $dW = Fdx$ for mechanical work can be rewritten as $dW = PdV$ in the case of work done by a piston. Here P represents the pressure of the working gas, and V its volume. Thus, on a graph of P versus V, the area under the curve represents the work done. When the gas expands, dx is positive, and the gas does positive work. When the gas is being compressed, dx is negative, and the gas does negative work, i.e., it absorbs energy.



a / P-V diagrams for a Carnot engine and an Otto engine.

Notice how, in the diagram of the Carnot engine in the top panel of figure a, the cycle goes clockwise around the curve, and therefore the part of the curve in which negative work is being done (arrowheads pointing to the left) are below the ones in which positive work is being done. This means that over all, the engine does a positive amount of work. This net work equals the area under the top part of the curve, minus the area under the bottom part of the curve, which is simply the area enclosed by the curve. Although the diagram for the Otto engine is more complicated, we can at least compare it on the same footing with the Carnot engine. The curve forms a figure-eight, because it cuts across itself. The top loop goes clockwise, so as in the case of the Carnot

engine, it represents positive work. The bottom loop goes counterclockwise, so it represents a net negative contribution to the work. This is because more work is expended in forcing out the exhaust than is generated in the intake stroke.

To make an engine as efficient as possible, we would like to make the loop have as much area as possible. What is it that determines the actual shape of the curve? First let's consider the constant-temperature expansion stroke that forms the top of the Carnot engine's P-V plot. This is analogous to the power stroke of an Otto engine. Heat is being sucked in from the hot reservoir, and since the working gas is always in thermal equilibrium with the hot reservoir, its temperature is constant. Regardless of the type of gas, we therefore have $PV = nkT$ with T held constant, and thus $P \propto V^{-1}$ is the mathematical shape of this curve --- a $y = 1/x$ graph, which is a hyperbola. This is all true regardless of whether the working gas is monoatomic, diatomic, or polyatomic. (The bottom of the loop is likewise of the form $P \propto V^{-1}$, but with a smaller constant of proportionality due to the lower temperature.)

Now consider the insulated expansion stroke that forms the right side of the curve for the Carnot engine. As shown on page 324, the relationship between pressure and temperature in an insulated compression or expansion is $T \propto P^b$, with $b = 2/5$, 2/7, or 1/4, respectively, for a monoatomic, diatomic, or polyatomic gas. For P as a function of V at constant T, the ideal gas law gives $P \propto T/V$, so $P \propto V^{-\gamma}$, where $\gamma = 1/(1-b)$ takes on the values 5/3, 7/5, and 4/3. The number γ can be interpreted as the ratio C_P/C_V, where C_P, the heat capacity at constant pressure, is the amount of heat required to raise the temperature of the gas by one degree while keeping its pressure constant, and C_V is the corresponding quantity under conditions of constant volume.

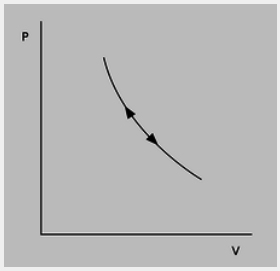
Example 22: The compression ratio

Operating along a constant-temperature stroke, the amount of mechanical work done by a heat engine can be calculated as follows:

$$\begin{aligned} PV &= nkT \\ \text{Setting } c = nkT \text{ to simplify the writing,} P &= cV^{-1} \\ W &= \int_{V_i}^{V_f} P dV \\ &= c\int_{V_i}^{V_f} V^{-1} dV \\ &= c\ln V_f - c\ln V_i \\ &= c\ln(V_f/V_i) \end{aligned}$$

The ratio V_f/V_i is called the *compression ratio* of the engine, and higher values result in more power along this stroke. Along an insulated stroke, we have $P \propto V^{-\gamma}$, with $\gamma \neq 1$, so the result for the work no longer has this perfect mathematical property of depending only on the ratio V_f/V_i. Nevertheless, the compression ratio is still a good figure of merit for predicting the performance of any heat engine, including an internal combustion engine. High compression ratios tend to make the working gas of an internal combustion engine heat up so much that it spontaneously explodes. When this happens in an Otto-cycle engine, it can cause ignition before the sparkplug fires, an undesirable effect known as pinging. For this reason, the compression ratio of an Otto-cycle automobile engine cannot normally exceed about 10. In a diesel engine, however, this effect is used intentionally, as an alternative to sparkplugs, and compression ratios can be 20 or more.

Example 23: Sound

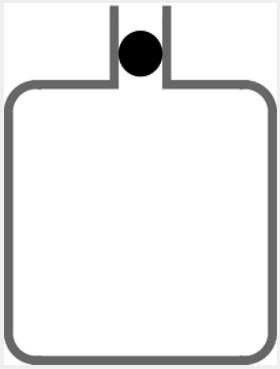


b / Example 23.

Figure b shows a P-V plot for a sound wave. As the pressure oscillates up and down, the air is heated and cooled by its compression and expansion. Heat conduction is a relatively slow process, so typically there is not enough time over each cycle for any significant amount of heat to flow from the hot areas to the cold areas. (This is analogous to insulated compression or expansion of a heat engine; in general, a compression or expansion of this type, with no transfer of heat, is called *adiabatic*.) The pressure and volume of a particular little piece of the air are therefore related according to $P \propto V^{-\gamma}$. The cycle of oscillation consists of motion back and forth along a single curve in the P-V plane, and since this curve encloses zero volume, no mechanical work is being done: the wave (under the assumed ideal conditions) propagates without any loss of energy due to friction.

The speed of sound is also related to γ. See example 13 on p. 375.

Example 24: Measuring γ using the “spring of air”



c / Example 24.

Figure c shows an experiment that can be used to measure the γ of a gas. When the mass m is inserted into bottle's neck, which has cross-sectional area A, the mass drops until it compresses the air enough so that the pressure is enough to support its weight. The observed frequency ω of oscillations about this equilibrium position y_o can be used to extract the γ of the gas.

$$\begin{aligned}\omega^2 &= \frac{k}{m} \\ &= -\frac{1}{m}\left.\frac{dF}{dy}\right|_{y_o} \\ &= -\frac{A}{m}\left.\frac{dP}{dy}\right|_{y_o} \\ &= -\frac{A^2}{m}\left.\frac{dP}{dV}\right|_{V_o}\end{aligned}$$

We make the bottle big enough so that its large surface-to-volume ratio prevents the conduction of any significant amount of heat through its walls during one cycle, so $P \propto V^{-\gamma}$, and $dP/dV = -\gamma P/V$. Thus,

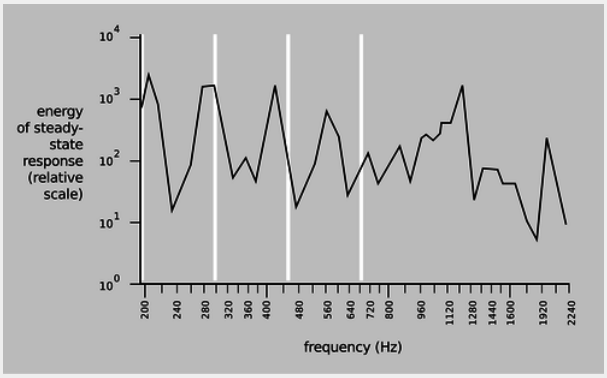
$$\omega^2 = \gamma\frac{A^2}{m}\frac{P_o}{V_o}$$

Example 25: The Helmholtz resonator

When you blow over the top of a beer bottle, you produce a pure tone. As you drink more of the beer, the pitch goes down. This is similar to example 24, except that instead of a solid mass m sitting inside the neck of the bottle, the moving mass is the air itself. As air rushes in and out of the bottle, its velocity is highest at the bottleneck, and since kinetic energy is proportional to the square of the velocity, essentially all of the kinetic energy is that of the air that's in the neck. In other words, we can replace m with $AL\rho$, where L is the length of the neck, and ρ is the density of the air. Substituting into the earlier result, we find that the resonant frequency is

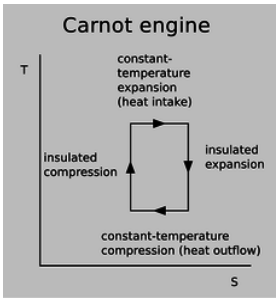
$$\omega^2 = \gamma \frac{P_0}{\rho} \frac{A}{LV_0}.$$

This is known as a Helmholtz resonator. As shown in figure d, a violin or an acoustic guitar has a Helmholtz resonance, since air can move in and out through the f-holes. Problem 10 is a more quantitative exploration of this.



d / The resonance curve of a 1713 Stradivarius violin, measured by Carleen Hutchins. There are a number of different resonance peaks, some strong and some weak; the ones near 200 and 400 Hz are vibrations of the wood, but the one near 300 Hz is a resonance of the air moving in and out through those holes shaped like the letter F. The white lines show the frequencies of the four strings.

We have already seen, based on the microscopic nature of entropy, that any Carnot engine has the same efficiency, and the argument only employed the assumption that the engine met the definition of a Carnot cycle: two insulated strokes, and two constant-temperature strokes. Since we didn't have to make any assumptions about the nature of the working gas being used, the result is evidently true for diatomic or polyatomic molecules, or for a gas that is not ideal. This result is surprisingly simple and general, and a little mysterious --- it even applies to possibilities that we have not even considered, such as a Carnot engine designed so that the working “gas” actually consists of a mixture of liquid droplets and vapor, as in a steam engine. How can it always turn out so simple, given the kind of mathematical complications that were swept under the rug in example 22? A better way to understand this result is by switching from P-V diagrams to a diagram of temperature versus entropy, as shown in figure e.



e / A T-S diagram for a Carnot engine.

An infinitesimal transfer of heat dQ gives rise to a change in entropy $dS = dQ/T$, so the area under the curve on a T-S plot gives the amount of heat transferred. The area under the top edge of the box in figure e, extending all the way down to the axis, represents the amount of heat absorbed from the hot reservoir, while the smaller area under the bottom edge represents the heat wasted into the cold reservoir. By conservation of energy, the area enclosed by the box therefore represents the amount of mechanical work being done, as for a P-V diagram. We can now see why the efficiency of a Carnot engine is independent of any of the physical details: the definition of a Carnot engine guarantees that the T-S diagram will be a rectangular box, and the efficiency depends only on the relative heights of the top and bottom of the box.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

6.6: Footnotes

1. This equal sharing will be justified more rigorously on page 322.
2. Even with smaller numbers of atoms, there is a problem with this kind of brute-force computation, because the tiniest measurement errors in the initial state would end up having large effects later on.
3. This is the same relation as the one on Boltzmann's tomb, just in a slightly different notation.

6.E: Thermodynamics (Exercises)

1. (a) Show that under conditions of standard pressure and temperature, the volume of a sample of an ideal gas depends only on the number of molecules in it.

(b) One mole is defined as 6.0×10^{23} atoms. Find the volume of one mole of an ideal gas, in units of liters, at standard temperature and pressure (0°C and 101 kPa). (answer check available at lightandmatter.com)

2. A gas in a cylinder expands its volume by an amount dV, pushing out a piston. Show that the work done by the gas on the piston is given by $dW = PdV$.

3. (a) A helium atom contains 2 protons, 2 electrons, and 2 neutrons. Find the mass of a helium atom. (answer check available at lightandmatter.com)
(b) Find the number of atoms in 1.0 kg of helium. (answer check available at lightandmatter.com)
(c) Helium gas is monoatomic. Find the amount of heat needed to raise the temperature of 1.0 kg of helium by 1.0 degree C. (This is known as helium's heat capacity at constant volume.) (answer check available at lightandmatter.com)

4. A sample of gas is enclosed in a sealed chamber. The gas consists of molecules, which are then split in half through some process such as exposure to ultraviolet light, or passing an electric spark through the gas. The gas returns to thermal equilibrium with the surrounding room. How does its pressure now compare with its pressure before the molecules were split?

5. Most of the atoms in the universe are in the form of gas that is not part of any star or galaxy: the intergalactic medium (IGM). The IGM consists of about 10^{-5} atoms per cubic centimeter, with a typical temperature of about 10^3 K. These are, in some sense, the density and temperature of the universe (not counting light, or the exotic particles known as “dark matter”). Calculate the pressure of the universe (or, speaking more carefully, the typical pressure due to the IGM).(answer check available at lightandmatter.com)

6. Estimate the pressure at the center of the Earth, assuming it is of constant density throughout. Note that g is not constant with respect to depth --- as shown in example 19 on page 105, g equals Gmr/b^3 for r, the distance from the center, less than b, the earth's radius.
(a) State your result in terms of G, m, and b.(answer check available at lightandmatter.com)
(b) Show that your answer from part a has the right units for pressure.
(c) Evaluate the result numerically. (answer check available at lightandmatter.com)
(d) Given that the earth's atmosphere is on the order of one thousandth the earth's radius, and that the density of the earth is several thousand times greater than the density of the lower atmosphere, check that your result is of a reasonable order of magnitude.

7. (a) Determine the ratio between the escape velocities from the surfaces of the earth and the moon. (answer check available at lightandmatter.com)
(b) The temperature during the lunar daytime gets up to about 130°\textup{C}. In the extremely thin (almost nonexistent) lunar atmosphere, estimate how the typical velocity of a molecule would compare with that of the same type of molecule in the earth's atmosphere. Assume that the earth's atmosphere has a temperature of 0°\textup{C}. (answer check available at lightandmatter.com)
(c) Suppose you were to go to the moon and release some fluorocarbon gas, with molecular formula $\mathrm{C}_n\mathrm{F}_{2n+2}$. Estimate what is the smallest fluorocarbon molecule (lowest n) whose typical velocity would be lower than that of an N_2 molecule on earth in proportion to the moon's lower escape velocity. The moon would be able to retain an atmosphere made of these molecules. (answer check available at lightandmatter.com)

8. Refrigerators, air conditioners, and heat pumps are heat engines that work in reverse. You put in mechanical work, and the effect is to take heat out of a cooler reservoir and deposit heat in a warmer one: $Q_L + W = Q_H$. As with the heat engines discussed previously, the efficiency is defined as the energy transfer you want (Q_L for a refrigerator or air conditioner, Q_H for a heat pump) divided by the energy transfer you pay for (W).

Efficiencies are supposed to be unitless, but the efficiency of an air conditioner is normally given in terms of an EER rating (or a more complex version called an SEER). The EER is defined as Q_L/W, but expressed in the barbaric units of of Btu/watt-hour. A typical EER rating for a residential air conditioner is about 10 Btu/watt-hour, corresponding to an efficiency of about 3. The standard temperatures used for testing an air conditioner's efficiency are $80^\circ F$ (27°C) inside and $95^\circ F$ (35°C) outside.

(a) What would be the EER rating of a reversed Carnot engine used as an air conditioner? (answer check available at lightandmatter.com)

(b) If you ran a 3-kW residential air conditioner, with an efficiency of 3, for one hour, what would be the effect on the total entropy of the universe? Is your answer consistent with the second law of thermodynamics? (answer check available at lightandmatter.com)

9. Even when resting, the human body needs to do a certain amount of mechanical work to keep the heart beating. This quantity is difficult to define and measure with high precision, and also depends on the individual and her level of activity, but it's estimated to be about 1 to 5 watts. Suppose we consider the human body as nothing more than a pump. A person who is just lying in bed all day needs about 1000 kcal/day worth of food to stay alive. (a) Estimate the person's thermodynamic efficiency as a pump, and (b) compare with the maximum possible efficiency imposed by the laws of thermodynamics for a heat engine operating across the difference between a body temperature of $37^\circ\mathrm{C}$ and an ambient temperature of $22^\circ\mathrm{C}$. (c) Interpret your answer.\hwans{hwans:heart-efficiency}

10. Example 25 on page 332 suggests analyzing the resonance of a violin at 300 Hz as a Helmholtz resonance. However, we might expect the equation for the frequency of a Helmholtz resonator to be a rather crude approximation here, since the f-holes are not long tubes, but slits cut through the face of the instrument, which is only about 2.5 mm thick. (a) Estimate the frequency that way anyway, for a violin with a volume of about 1.6 liters, and f-holes with a total area of 10 cm^2. (b) A common rule of thumb is that at an open end of an air column, such as the neck of a real Helmholtz resonator, some air beyond the mouth also vibrates as if it was inside the tube, and that this effect can be taken into account by adding 0.4 times the diameter of the tube for each open end (i.e., 0.8 times the diameter when both ends are open). Applying this to the violin's f-holes results in a huge change in L, since the ~ 7 mm width of the f-hole is considerably greater than the thickness of the wood. Try it, and see if the result is a better approximation to the observed frequency of the resonance.\hwans{hwans:violin-helmholtz}

Contributors and Attributions

CHAPTER OVERVIEW

7: Waves

Dandelion. Cello. Read those two words, and your brain instantly conjures a stream of associations, the most prominent of which have to do with vibrations. Our mental category of “dandelion-ness” is strongly linked to the color of light waves that vibrate about half a million billion times a second: yellow. The velvety throb of a cello has as its most obvious characteristic a relatively low musical pitch --- the note you're spontaneously imagining right now might be one whose sound vibrations repeat at a rate of a hundred times a second.

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

Evolution seems to have designed our two most important senses around the assumption that our environment is made of waves, whereas up until now, we've mostly taken the view that Nature can be understood by breaking her down into smaller and smaller parts, ending up with particles as her most fundamental building blocks. Does that work for light and sound? Sound waves are disturbances in air, which is made of atoms, but light, on the other hand, isn't a vibration of atoms. Light, unlike sound, can travel through a vacuum: if you're reading this by sunlight, you're taking advantage of light that had to make it through millions of miles of vacuum to get to you. Waves, then, are not just a trick that vibrating atoms can do. Waves are one of the basic phenomena of the universe. At the end of this book, we'll even see that the things we've been calling particles, such as electrons, are really waves![1]

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

7.1: Free Waves

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

7.2: Bounded Waves

Speech is what separates humans most decisively from animals. No other species can master syntax, and even though chimpanzees can learn a vocabulary of hand signs, there is an unmistakable difference between a human infant and a baby chimp: starting from birth, the human experiments with the production of complex speech sounds.

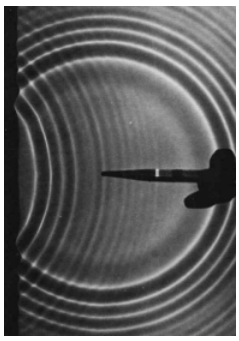
Since speech sounds are instinctive for us, we seldom think about them consciously. How do we control sound waves so skillfully? Mostly we do it by changing the shape of a connected set of hollow cavities in our chest, throat, and head. Somehow by moving the boundaries of this space in and out, we can produce all the vowel sounds. Up until now, we have been studying only those properties of waves that can be understood as if they existed in an infinite, open space with no boundaries. In this chapter we address what happens when a wave is confined within a certain space, or when a wave pattern encounters the boundary between two different media, such as when a light wave moving through air encounters a glass windowpane.

a / A cross-sectional view of a human body, showing the vocal tract.

6.2.1 Reflection, transmission, and absorption

Reflection and transmission

Sound waves can echo back from a cliff, and light waves are reflected from the surface of a pond. We use the word reflection, normally applied only to light waves in ordinary speech, to describe any such case of a wave rebounding from a barrier. Figure (a) shows a circular water wave being reflected from a straight wall. In this chapter, we will concentrate mainly on reflection of waves that move in one dimension, as in figure c/1.



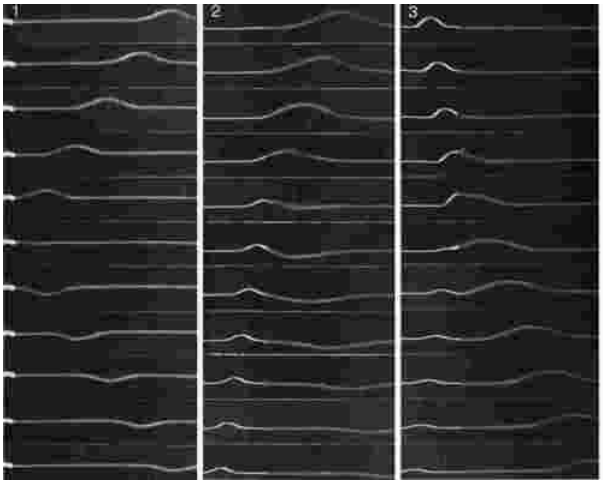
b / Circular water waves are reflected from a boundary on the left. (PSSC Physics)

Wave reflection does not surprise us. After all, a material object such as a rubber ball would bounce back in the same way. But waves are not objects, and there are some surprises in store.

First, only part of the wave is usually reflected. Looking out through a window, we see light waves that passed through it, but a person standing outside would also be able to see her reflection in the glass. A light wave that strikes the glass is partly reflected and partly transmitted (passed) by the glass. The energy of the original wave is split between the two. This is different from the behavior of the rubber ball, which must go one way or the other, not both.

Second, consider what you see if you are swimming underwater and you look up at the surface. You see your own reflection. This is utterly counterintuitive, since we would expect the light waves to burst forth to freedom in the wide-open air. A material projectile shot up toward the surface would never rebound from the water-air boundary!

What is it about the difference between two media that causes waves to be partly reflected at the boundary between them? Is it their density? Their chemical composition? Ultimately all that matters is the speed of the wave in the two media. A wave is partially reflected and partially transmitted at the boundary between media in which it has different speeds. For example, the speed of light waves in window glass is about 30% less than in air, which explains why windows always make reflections. Figure c shows examples of wave pulses being reflected at the boundary between two coil springs of different weights, in which the wave speed is different.



c / 1. A wave on a coil spring, initially traveling to the left, is reflected from the fixed end. 2. A wave in the lighter spring, where the wave speed is greater, travels to the left and is then partly reflected and partly transmitted at the boundary with the heavier coil spring, which has a lower wave speed. The reflection is inverted. 3. A wave moving to the right in the heavier spring is partly reflected at the boundary with the lighter spring. The reflection is uninverted. (PSSC Physics)

Reflections such as b and c/1, where a wave encounters a massive fixed object, can usually be understood on the same basis as cases like c/2 and c/3 where two media meet. Example c/1, for instance, is like a more extreme version of example c/2. If the heavy coil spring in c/2 was made heavier and heavier, it would end up acting like the fixed wall to which the light spring in c/1 has been attached.

Exercise 7.2.1

In figure c/1, the reflected pulse is upside-down, but its depth is just as big as the original pulse's height. How does the energy of the reflected pulse compare with that of the original?

(answer in the back of the PDF version of the book)

Example 8: Fish have internal ears.

Why don't fish have ear-holes? The speed of sound waves in a fish's body is not much different from their speed in water, so sound waves are not strongly reflected from a fish's skin. They pass right through its body, so fish can have internal ears.

Example 9: Whale songs traveling long distances

Sound waves travel at drastically different speeds through rock, water, and air. Whale songs are thus strongly reflected both at both the bottom and the surface. The sound waves can travel hundreds of miles, bouncing repeatedly between the bottom and the surface, and still be detectable. Sadly, noise pollution from ships has nearly shut down this cetacean version of the internet.

Example 10: Long-distance radio communication

Radio communication can occur between stations on opposite sides of the planet. The mechanism is entirely similar to the one explained in the previous example, but the three media involved are the earth, the atmosphere, and the ionosphere.

self-check:

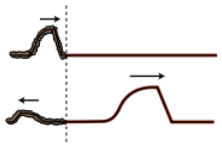
Sonar is a method for ships and submarines to detect each other by producing sound waves and listening for echoes. What properties would an underwater object have to have in order to be invisible to sonar?

(answer in the back of the PDF version of the book)

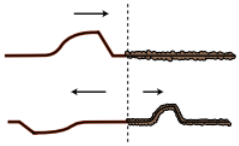
The use of the word “reflection” naturally brings to mind the creation of an image by a mirror, but this might be confusing, because we do not normally refer to “reflection” when we look at surfaces that are not shiny. Nevertheless, reflection is how we see the surfaces of all objects, not just polished ones. When we look at a sidewalk, for example, we are actually seeing the reflecting of the sun from the concrete. The reason we don't see an image of the sun at our feet is simply that the rough surface blurs the image so drastically.

Inverted and uninverted reflections

Notice how the pulse reflected back to the right in example c/2 comes back upside-down, whereas the one reflected back to the left in c/3 returns in its original upright form. This is true for other waves as well. In general, there are two possible types of reflections, a reflection back into a faster medium and a reflection back into a slower medium. One type will always be an inverting reflection and one noninverting.



d / An uninverted reflection. The reflected pulse is reversed front to back, but is not upside-down.

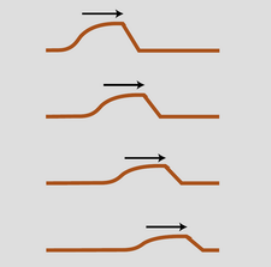


e / An inverted reflection. The reflected pulse is reversed both front to back and top to bottom.

It's important to realize that when we discuss inverted and uninverted reflections on a string, we are talking about whether the wave is flipped across the direction of motion (i.e., upside-down in these drawings). The reflected pulse will always be reversed front to back, as shown in figures d and e. This is because it is traveling in the other direction. The leading edge of the pulse is what gets reflected first, so it is still ahead when it starts back to the left --- it's just that “ahead” is now in the opposite direction.

Absorption

So far we have tacitly assumed that wave energy remains as wave energy, and is not converted to any other form. If this was true, then the world would become more and more full of sound waves, which could never escape into the vacuum of outer space. In reality, any mechanical wave consists of a traveling pattern of vibrations of some physical medium, and vibrations of matter always produce heat, as when you bend a coathangar back and forth and it becomes hot. We can thus expect that in mechanical waves such as water waves, sound waves, or waves on a string, the wave energy will gradually be converted into heat. This is referred to as absorption. The reduction in the wave's energy can also be described as a reduction in amplitude, the relationship between them being, as with a vibrating object, $E \propto A^2$.



f / A pulse traveling through a highly absorptive medium.

The wave suffers a decrease in amplitude, as shown in figure f. The decrease in amplitude amounts to the same fractional change for each unit of distance covered. For example, if a wave decreases from amplitude 2 to amplitude 1 over a distance of 1 meter, then after traveling another meter it will have an amplitude of 1/2. That is, the reduction in amplitude is exponential. This can be proved as follows. By the principle of superposition, we know that a wave of amplitude 2 must behave like the superposition of two identical waves of amplitude 1. If a single amplitude-1 wave would die down to amplitude 1/2 over a certain distance, then two amplitude-1 waves superposed on top of one another to make amplitude 1+1=2 must die down to amplitude 1/2+1/2=1 over the same distance.

Note

As a wave undergoes absorption, it loses energy. Does this mean that it slows down?

(answer in the back of the PDF version of the book)

In many cases, this frictional heating effect is quite weak. Sound waves in air, for instance, dissipate into heat extremely slowly, and the sound of church music in a cathedral may reverberate for as much as 3 or 4 seconds before it becomes inaudible. During this time it has traveled over a kilometer! Even this very gradual dissipation of energy occurs mostly as heating of the church's walls and by the leaking of sound to the outside (where it will eventually end up as heat). Under the right conditions (humid air and low frequency), a sound wave in a straight pipe could theoretically travel hundreds of kilometers before being noticeable attenuated.

In general, the absorption of mechanical waves depends a great deal on the chemical composition and microscopic structure of the medium. Ripples on the surface of antifreeze, for instance, die out extremely rapidly compared to ripples on water. For sound waves and surface waves in liquids and gases, what matters is the viscosity of the substance, i.e., whether it flows easily like water or mercury or more sluggishly like molasses or antifreeze. This explains why our intuitive expectation of strong absorption of sound in water is incorrect. Water is a very weak absorber of sound (viz. whale songs and sonar), and our incorrect intuition arises from focusing on the wrong property of the substance: water's high density, which is irrelevant, rather than its low viscosity, which is what matters.

Light is an interesting case, since although it can travel through matter, it is not itself a vibration of any material substance. Thus we can look at the star Sirius, 10^{14} km away from us, and be assured that none of its light was absorbed in the vacuum of outer space during its 9-year journey to us. The Hubble Space Telescope routinely observes light that has been on its way to us since the early history of the universe, billions of years ago. Of course the energy of light can be dissipated if it does pass through matter (and the light from distant galaxies is often absorbed if there happen to be clouds of gas or dust in between).

Example 11: Soundproofing

Typical amateur musicians setting out to soundproof their garages tend to think that they should simply cover the walls with the densest possible substance. In fact, sound is not absorbed very strongly even by passing through several inches of wood. A better strategy for soundproofing is to create a sandwich of alternating layers of materials in which the speed of sound is very different, to encourage reflection.
The classic design is alternating layers of fiberglass and plywood. The speed of sound in plywood is very high, due to its stiffness, while its speed in fiberglass is essentially the same as its speed in air. Both materials are fairly good sound absorbers, but sound waves passing through a few inches of them are still not going to be absorbed sufficiently. The point of combining them is that a sound wave that tries to get out will be strongly reflected at each of the fiberglass-plywood boundaries, and will bounce back and forth many times like a ping pong ball. Due to all the back-and-forth motion, the sound may end up traveling a total distance equal to ten times the actual thickness of the soundproofing before it escapes. This is the equivalent of having ten times the thickness of sound-absorbing material.

Example 12: Radio transmission

A radio transmitting station must have a length of wire or cable connecting the amplifier to the antenna. The cable and the antenna act as two different media for radio waves, and there will therefore be partial reflection of the waves as they come from the cable to the antenna. If the waves bounce back and forth many times between the amplifier and the antenna, a great deal of their energy will be absorbed. There are two ways to attack the problem. One possibility is to design the antenna so that the speed of the waves in it is as close as possible to the speed of the waves in the cable; this minimizes the amount of reflection. The other method is to connect the amplifier to the antenna using a type of wire or cable that does not strongly absorb the waves. Partial reflection then becomes irrelevant, since all the wave energy will eventually exit through the antenna.

Discussion Questions

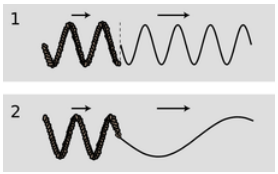
◊ A sound wave that underwent a pressure-inverting reflection would have its compressions converted to expansions and vice versa. How would its energy and frequency compare with those of the original sound? Would it sound any different? What happens if you swap the two wires where they connect to a stereo speaker, resulting in waves that vibrate in the opposite way?

6.2.2 Quantitative treatment of reflection

In this subsection we analyze the reasons why reflections occur at a speed-changing boundary, predict quantitatively the intensities of reflection and transmission, and discuss how to predict for any type of wave which reflections are inverting and which are uninverting.

Why reflection occurs

To understand the fundamental reasons for what does occur at the boundary between media, let's first discuss what doesn't happen. For the sake of concreteness, consider a sinusoidal wave on a string. If the wave progresses from a heavier portion of the string, in which its velocity is low, to a lighter-weight part, in which it is high, then the equation $v = f\lambda$ tells us that it must change its frequency, or its wavelength, or both. If only the frequency changed, then the parts of the wave in the two different portions of the string would quickly get out of step with each other, producing a discontinuity in the wave, g/1. This is unphysical, so we know that the wavelength must change while the frequency remains constant, g/2.



g / 1. A change in frequency without a change in wavelength would produce a discontinuity in the wave.\ 2. A simple change in wavelength without a reflection would result in a sharp kink in the wave.

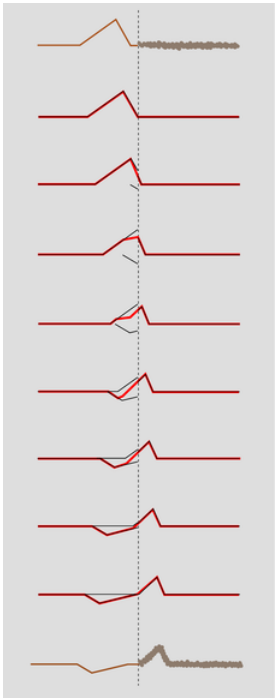
But there is still something unphysical about figure g/2. The sudden change in the shape of the wave has resulted in a sharp kink at the boundary. This can't really happen, because the medium tends to accelerate in such a way as to eliminate curvature. A sharp kink corresponds to an infinite curvature at one point, which would produce an infinite acceleration, which would not be consistent with the smooth pattern of wave motion envisioned in fig. g/2. Waves can have kinks, but not stationary kinks.

We conclude that without positing partial reflection of the wave, we cannot simultaneously satisfy the requirements of (1) continuity of the wave, and (2) no sudden changes in the slope of the wave. In other words, we assume that both the wave and its derivative are continuous functions.)

Does this amount to a proof that reflection occurs? Not quite. We have only proved that certain types of wave motion are not valid solutions. In the following subsection, we prove that a valid solution can always be found in which a reflection occurs. Now in physics, we normally assume (but seldom prove formally) that the equations of motion have a unique solution, since otherwise a given set of initial conditions could lead to different behavior later on, but the Newtonian universe is supposed to be deterministic. Since the solution must be unique, and we derive below a valid solution involving a reflected pulse, we will have ended up with what amounts to a proof of reflection.

Intensity of reflection

I will now show, in the case of waves on a string, that it is possible to satisfy the physical requirements given above by constructing a reflected wave, and as a bonus this will produce an equation for the proportions of reflection and transmission and a prediction as to which conditions will lead to inverted and which to uninverted reflection. We assume only that the principle of superposition holds, which is a good approximations for waves on a string of sufficiently small amplitude.



h / A pulse being partially reflected and partially transmitted at the boundary between two strings in which the wave speed is different. The top drawing shows the pulse heading to the right, toward the heaver string. For clarity, all but the first and last drawings are schematic. Once the reflected pulse begins to emerge from the boundary, it adds together with the trailing parts of the incident pulse. Their sum, shown as a wider line, is what is actually observed.

Let the unknown amplitudes of the reflected and transmitted waves be R and T, respectively. An inverted reflection would be represented by a negative value of R. We can without loss of generality take the incident (original) wave to have unit amplitude. Superposition tells us that if, for instance, the incident wave had double this amplitude, we could immediately find a corresponding solution simply by doubling R and T.

Just to the left of the boundary, the height of the wave is given by the height 1 of the incident wave, plus the height R of the part of the reflected wave that has just been created and begun heading back, for a total height of $1+R$. On the right side immediately next to the boundary, the transmitted wave has a height T. To avoid a discontinuity, we must have

$$1+R=T.$$

Next we turn to the requirement of equal slopes on both sides of the boundary. Let the slope of the incoming wave be s immediately to the left of the junction. If the wave was 100% reflected, and without inversion, then the slope of the reflected wave

would be $-s$, since the wave has been reversed in direction. In general, the slope of the reflected wave equals $-sR$, and the slopes of the superposed waves on the left side add up to $s - sR$. On the right, the slope depends on the amplitude, T, but is also changed by the stretching or compression of the wave due to the change in speed. If, for example, the wave speed is twice as great on the right side, then the slope is cut in half by this effect. The slope on the right is therefore $s(v_1/v_2)T$, where v_1 is the velocity in the original medium and v_2 the velocity in the new medium. Equality of slopes gives $s - sR = s(v_1/v_2)T$, or

$$1 - R = \frac{v_1}{v_2}T.$$

Solving the two equations for the unknowns R and T gives

$$R = \frac{v_2 - v_1}{v_2 + v_1} \text{ and} T \quad = \frac{2v_2}{v_2 + v_1}.$$

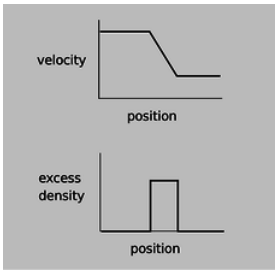
The first equation shows that there is no reflection unless the two wave speeds are different, and that the reflection is inverted in reflection back into a fast medium.

The energies of the transmitted and reflected wavers always add up to the same as the energy of the original wave. There is never any abrupt loss (or gain) in energy when a wave crosses a boundary; conversion of wave energy to heat occurs for many types of waves, but it occurs throughout the medium. The equation for T, surprisingly, allows the amplitude of the transmitted wave to be greater than 1, i.e., greater than that of the incident wave. This does not violate conservation of energy, because this occurs when the second string is less massive, reducing its kinetic energy, and the transmitted pulse is broader and less strongly curved, which lessens its potential energy.

Inverted and uninverted reflections in general (optional)

For waves on a string, reflections back into a faster medium are inverted, while those back into a slower medium are uninverted. Is this true for all types of waves? The rather subtle answer is that it depends on what property of the wave you are discussing.

Let's start by considering wave disturbances of freeway traffic. Anyone who has driven frequently on crowded freeways has observed the phenomenon in which one driver taps the brakes, starting a chain reaction that travels backward down the freeway as each person in turn exercises caution in order to avoid rear-ending anyone. The reason why this type of wave is relevant is that it gives a simple, easily visualized example of how our description of a wave depends on which aspect of the wave we have in mind. In steadily flowing freeway traffic, both the density of cars and their velocity are constant all along the road. Since there is no disturbance in this pattern of constant velocity and density, we say that there is no wave. Now if a wave is touched off by a person tapping the brakes, we can either describe it as a region of high density or as a region of decreasing velocity.



i / A wave pattern in freeway traffic.

The freeway traffic wave is in fact a good model of a sound wave, and a sound wave can likewise be described either by the density (or pressure) of the air or by its speed. Likewise many other types of waves can be described by either of two functions, one of which is often the derivative of the other with respect to position.

Now let's consider reflections. If we observe the freeway wave in a mirror, the high-density area will still appear high in density, but velocity in the opposite direction will now be described by a negative number. A person observing the mirror image will draw the same density graph, but the velocity graph will be flipped across the x axis, and its original region of negative slope will now have positive slope. Although I don't know any physical situation that would correspond to the reflection of a traffic wave, we can immediately apply the same reasoning to sound waves, which often do get reflected, and determine that a reflection can either be density-inverting and velocity-uninverting or density-uninverting and velocity-inverting.

j / In the mirror image, the areas of positive excess traffic density are still positive, but the velocities of the cars have all been reversed, so areas of positive excess velocity have been turned into negative ones.

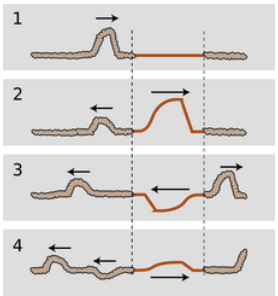
This same type of situation will occur over and over as one encounters new types of waves, and to apply the analogy we need only determine which quantities, like velocity, become negated in a mirror image and which, like density, stay the same.

A light wave, for instance, consists of a traveling pattern of electric and magnetic fields. All you need to know in order to analyze the reflection of light waves is how electric and magnetic fields behave under reflection; you don't need to know any of the detailed physics of electricity and magnetism. An electric field can be detected, for example, by the way one's hair stands on end. The direction of the hair indicates the direction of the electric field. In a mirror image, the hair points the other way, so the electric field is apparently reversed in a mirror image. The behavior of magnetic fields, however, is a little tricky. The magnetic properties of a bar magnet, for instance, are caused by the aligned rotation of the outermost orbiting electrons of the atoms. In a mirror image, the direction of rotation is reversed, say from clockwise to counterclockwise, and so the magnetic field is reversed twice: once simply because the whole picture is flipped and once because of the reversed rotation of the electrons. In other words, magnetic fields do not reverse themselves in a mirror image. We can thus predict that there will be two possible types of reflection of light waves. In one, the electric field is inverted and the magnetic field uninverted (example 23, p. 699). In the other, the electric field is uninverted and the magnetic field inverted.

6.2.3 Interference effects

If you look at the front of a pair of high-quality binoculars, you will notice a greenish-blue coating on the lenses. This is advertised as a coating to prevent reflection. Now reflection is clearly undesirable --- we want the light to go in the binoculars --- but so far I've described reflection as an unalterable fact of nature, depending only on the properties of the two wave media. The coating can't change the speed of light in air or in glass, so how can it work? The key is that the coating itself is a wave medium. In other words, we have a three-layer sandwich of materials: air, coating, and glass. We will analyze the way the coating works, not because optical coatings are an important part of your education but because it provides a good example of the general phenomenon of wave interference effects.

There are two different interfaces between media: an air-coating boundary and a coating-glass boundary. Partial reflection and partial transmission will occur at each boundary. For ease of visualization let's start by considering an equivalent system consisting of three dissimilar pieces of string tied together, and a wave pattern consisting initially of a single pulse.

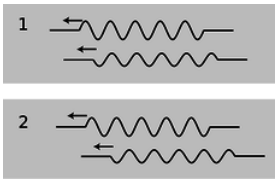


k / A pulse encounters two boundaries.

Figure k/1 shows the incident pulse moving through the heavy rope, in which its velocity is low. When it encounters the lighter-weight rope in the middle, a faster medium, it is partially reflected and partially transmitted. (The transmitted pulse is bigger, but nevertheless has only part of the original energy.) The pulse transmitted by the first interface is then partially reflected and partially transmitted by the second boundary, k/3. In figure k/4, two pulses are on the way back out to the left, and a single pulse is heading off to the right. (There is still a weak pulse caught between the two boundaries, and this will rattle back and forth, rapidly getting too weak to detect as it leaks energy to the outside with each partial reflection.)

Note how, of the two reflected pulses in k/4, one is inverted and one uninverted. One underwent reflection at the first boundary (a reflection back into a slower medium is uninverted), but the other was reflected at the second boundary (reflection back into a faster medium is inverted).

Now let's imagine what would have happened if the incoming wave pattern had been a long sinusoidal wave train instead of a single pulse. The first two waves to reemerge on the left could be in phase, l/1, or out of phase, l/2, or anywhere in between. The amount of lag between them depends entirely on the width of the middle segment of string. If we choose the width of the middle string segment correctly, then we can arrange for destructive interference to occur, l/2, with cancellation resulting in a very weak reflected wave.



l / A sine wave has been reflected at two different boundaries, and the two reflections interfere.

This whole analysis applies directly to our original case of optical coatings. Visible light from most sources does consist of a stream of short sinusoidal wave-trains such as the ones drawn above. The only real difference between the waves-on-a-rope example and the case of an optical coating is that the first and third media are air and glass, in which light does not have the same speed. However, the general result is the same as long as the air and the glass have light-wave speeds that are either both greater than the coating's or both less than the coating's.

The business of optical coatings turns out to be a very arcane one, with a plethora of trade secrets and “black magic” techniques handed down from master to apprentice. Nevertheless, the ideas you have learned about waves in general are sufficient to allow you to come to some definite conclusions without any further technical knowledge. The self-check and discussion questions will direct you along these lines of thought.

self-check:

Color corresponds to wavelength of light waves. Is it possible to choose a thickness for an optical coating that will produce destructive interference for all colors of light?

(answer in the back of the PDF version of the book)

This example was typical of a wide variety of wave interference effects. With a little guidance, you are now ready to figure out for yourself other examples such as the rainbow pattern made by a compact disc or by a layer of oil on a puddle.

Discussion Questions

- Is it possible to get *complete* destructive interference in an optical coating, at least for light of one specific wavelength?
- Sunlight consists of sinusoidal wave-trains containing on the order of a hundred cycles back-to-back, for a length of something like a tenth of a millimeter. What happens if you try to make an optical coating thicker than this?
- Suppose you take two microscope slides and lay one on top of the other so that one of its edges is resting on the corresponding edge of the bottom one. If you insert a sliver of paper or a hair at the opposite end, a wedge-shaped layer of air will exist in the middle, with a thickness that changes gradually from one end to the other. What would you expect to see if the slides were illuminated from above by light of a single color? How would this change if you gradually lifted the lower edge of the top slide until the two slides were finally parallel?
- An observation like the one described in discussion question C was used by Newton as evidence *against* the wave theory of light! If Newton didn't know about inverting and noninverting reflections, what would have seemed inexplicable to him about the region where the air layer had zero or nearly zero thickness?

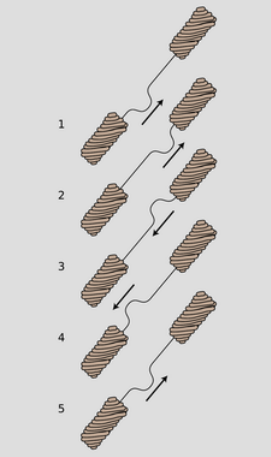
6.2.4 Waves bounded on both sides

In the example of the previous section, it was theoretically true that a pulse would be trapped permanently in the middle medium, but that pulse was not central to our discussion, and in any case it was weakening severely with each partial reflection. Now consider a guitar string. At its ends it is tied to the body of the instrument itself, and since the body is very massive, the behavior of

the waves when they reach the end of the string can be understood in the same way as if the actual guitar string was attached on the ends to strings that were extremely massive. Reflections are most intense when the two media are very dissimilar. Because the wave speed in the body is so radically different from the speed in the string, we should expect nearly 100% reflection.

n / We model a guitar string attached to the guitar's body at both ends as a light-weight string attached to extremely heavy strings at its ends.

Although this may seem like a rather bizarre physical model of the actual guitar string, it already tells us something interesting about the behavior of a guitar that we would not otherwise have understood. The body, far from being a passive frame for attaching the strings to, is actually the exit path for the wave energy in the strings. With every reflection, the wave pattern on the string loses a tiny fraction of its energy, which is then conducted through the body and out into the air. (The string has too little cross-section to make sound waves efficiently by itself.) By changing the properties of the body, moreover, we should expect to have an effect on the manner in which sound escapes from the instrument. This is clearly demonstrated by the electric guitar, which has an extremely massive, solid wooden body. Here the dissimilarity between the two wave media is even more pronounced, with the result that wave energy leaks out of the string even more slowly. This is why an electric guitar with no electric pickup can hardly be heard at all, and it is also the reason why notes on an electric guitar can be sustained for longer than notes on an acoustic guitar.



m / A pulse bounces back and forth.

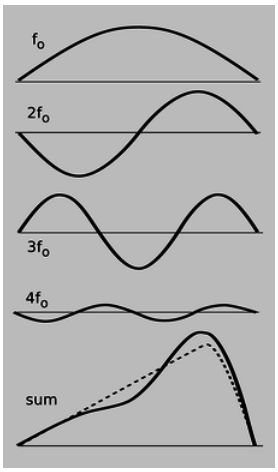
If we initially create a disturbance on a guitar string, how will the reflections behave? In reality, the finger or pick will give the string a triangular shape before letting it go, and we may think of this triangular shape as a very broad “dent” in the string which will spread out in both directions. For simplicity, however, let's just imagine a wave pattern that initially consists of a single, narrow pulse traveling up the neck, m/1. After reflection from the top end, it is inverted, m/3. Now something interesting happens: figure m/5 is identical to figure m/1. After two reflections, the pulse has been inverted twice and has changed direction twice. It is now back where it started. The motion is periodic. This is why a guitar produces sounds that have a definite sensation of pitch.

self-check:

Notice that from m/1 to m/5, the pulse has passed by every point on the string exactly twice. This means that the total distance it has traveled equals 2 L, where L is the length of the string. Given this fact, what are the period and frequency of the sound it produces, expressed in terms of L and v, the velocity of the wave?

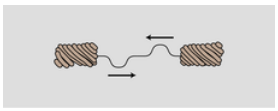
(answer in the back of the PDF version of the book)

Note that if the waves on the string obey the principle of superposition, then the velocity must be independent of amplitude, and the guitar will produce the same pitch regardless of whether it is played loudly or softly. In reality, waves on a string obey the principle of superposition approximately, but not exactly. The guitar, like just about any acoustic instrument, is a little out of tune when played loudly. (The effect is more pronounced for wind instruments than for strings, but wind players are able to compensate for it.)



p / Any wave can be made by superposing sine waves.

Now there is only one hole in our reasoning. Suppose we somehow arrange to have an initial setup consisting of two identical pulses heading toward each other, as in figure (o). They will pass through each other, undergo a single inverting reflection, and come back to a configuration in which their positions have been exactly interchanged. This means that the period of vibration is half as long. The frequency is twice as high.



o / The period of this double-pulse pattern is half of what we'd otherwise expect.

This might seem like a purely academic possibility, since nobody actually plays the guitar with two picks at once! But in fact it is an example of a very general fact about waves that are bounded on both sides. A mathematical theorem called Fourier's theorem states that any wave can be created by superposing sine waves. Figure p shows how even by using only four sine waves with appropriately chosen amplitudes, we can arrive at a sum which is a decent approximation to the realistic triangular shape of a guitar string being plucked. The one-hump wave, in which half a wavelength fits on the string, will behave like the single pulse we originally discussed. We call its frequency f_o. The two-hump wave, with one whole wavelength, is very much like the two-pulse example. For the reasons discussed above, its frequency is $2f_o$. Similarly, the three-hump and four-hump waves have frequencies of $3f_o$ and $4f_o$.

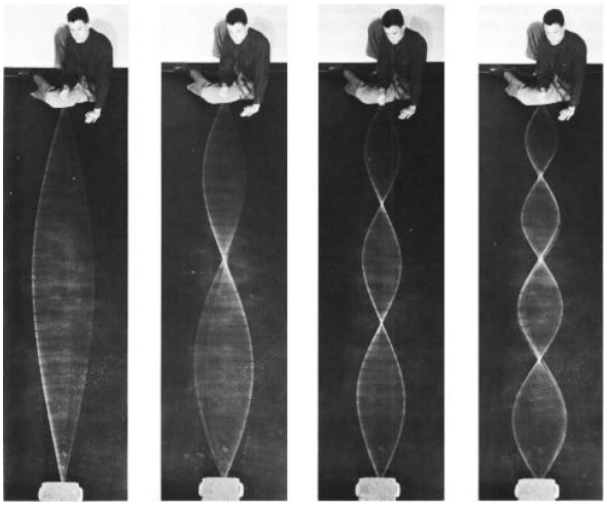
Theoretically we would need to add together infinitely many such wave patterns to describe the initial triangular shape of the string exactly, although the amplitudes required for the very high frequency parts would be very small, and an excellent approximation could be achieved with as few as ten waves.

We thus arrive at the following very general conclusion. Whenever a wave pattern exists in a medium bounded on both sides by media in which the wave speed is very different, the motion can be broken down into the motion of a (theoretically infinite) series of sine waves, with frequencies f_o, $2f_o$, $3f_o$, ... Except for some technical details, to be discussed below, this analysis applies to a vast range of sound-producing systems, including the air column within the human vocal tract. Because sounds composed of this kind of pattern of frequencies are so common, our ear-brain system has evolved so as to perceive them as a single, fused sensation of tone.

Musical applications

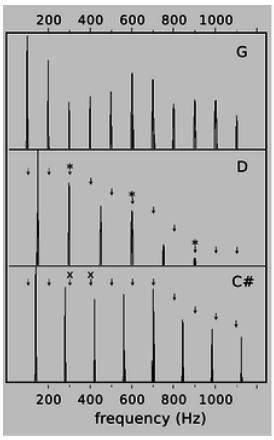
Many musicians claim to be able to pick out by ear several of the frequencies $2f_o$, $3f_o$, ..., called overtones or *harmonics* of the fundamental f_o, but they are kidding themselves. In reality, the overtone series has two important roles in music, neither of which

depends on this fictitious ability to “hear out” the individual overtones.



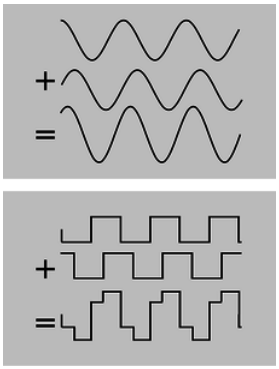
s / Standing waves on a rope. (PSSC Physics.)

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



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

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



r / If you take a sine wave and make a copy of it shifted over, their sum is still a sine wave. The same is not true for a square wave.

Standing waves

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

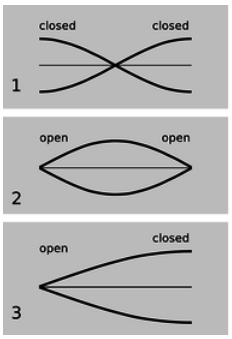
If you think about it, it's not even obvious that sine waves should be able to do this trick. After all, waves are supposed to travel at a set speed, aren't they? The speed isn't supposed to be zero! Well, we can actually think of a standing wave as a superposition of a moving sine wave with its own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

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

Standing-wave patterns of air columns

The air column inside a wind instrument behaves very much like the wave-on-a-string example we've been concentrating on so far, the main difference being that we may have either inverting or noninverting reflections at the ends.

Some organ pipes are closed at both ends. The speed of sound is different in metal than in air, so there is a strong reflection at the closed ends, and we can have standing waves. These reflections are both density-noninverting, so we get symmetric standing-wave patterns, such as the one shown in figure u/1.



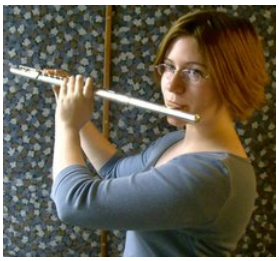
u / Graphs of excess density versus position for the lowest-frequency standing waves of three types of air columns. Points on the axis have normal air density.

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

Figure t shows the sound waves in and around a bamboo Japanese flute called a shakuhachi, which is *open* at both ends of the air column. We can only have a standing wave pattern if there are reflections at the ends, but that is very counterintuitive --- why is there any reflection at all, if the sound wave is free to emerge into open space, and there is no change in medium? Recall the reason why we got reflections at a change in medium: because the wavelength changes, so the wave has to readjust itself from one pattern to another, and the only way it can do that without developing a kink is if there is a reflection. Something similar is happening here. The only difference is that the wave is adjusting from being a plane wave to being a spherical wave. The reflections at the open ends are density-inverting, u/2, so the wave pattern is pinched off at the ends. Comparing panels 1 and 2 of the figure, we see that although the wave pattens are different, in both cases the wavelength is the same: in the lowest-frequency standing wave, half a wavelength fits inside the tube. Thus, it isn't necessary to memorize which type of reflection is inverting and which is inverting. It's only necessary to know that the tubes are symmetric.

v / A pan pipe is an asymmetric air column, open at the top and closed at the bottom.

Finally, we can have an asymmetric tube: closed at one end and open at the other. A common example is the pan pipes, v, which are closed at the bottom and open at the top. The standing wave with the lowest frequency is therefore one in which 1/4 of a wavelength fits along the length of the tube, as shown in figure u/3.



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

Sometimes an instrument's physical appearance can be misleading. A concert flute, w, is closed at the mouth end and open at the other, so we would expect it to behave like an asymmetric air column; in reality, it behaves like a symmetric air column open at both ends, because the embouchure hole (the hole the player blows over) acts like an open end. The clarinet and the saxophone look similar, having a moutpiece and reed at one end and an open end at the other, but they act different. In fact the clarinet's air

column has patterns of vibration that are asymmetric, the saxophone symmetric. The discrepancy comes from the difference between the conical tube of the sax and the cylindrical tube of the clarinet. The adjustment of the wave pattern from a plane wave to a spherical wave is more gradual at the flaring bell of the saxophone.

self-check:

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

(answer in the back of the PDF version of the book)

Example 13: The speed of sound

We can get a rough and ready derivation of the equation for the speed of sound by analyzing the standing waves in a cylindrical air column as a special type of Helmholtz resonance (example 25 on page 332), in which the cavity happens to have the same cross-sectional area as the neck. Roughly speaking, the regions of maximum density variation act like the cavity. The regions of minimum density variation, on the other hand, are the places where the velocity of the air is varying the most; these regions throttle back the speed of the vibration, because of the inertia of the moving air. If the cylinder has cross-sectional area A, then the "cavity" and "neck" parts of the wave both have lengths of something like $\lambda/2$, and the volume of the "cavity" is about $A\lambda/2$. We get $v = f\lambda = (\dots)\sqrt{\gamma P_0/\rho}$, where the factor $(\dots)$ represents numerical stuff that we can't possibly hope to have gotten right with such a crude argument. The correct result is in fact $v = \sqrt{\gamma P_0/\rho}$. Isaac Newton attempted the same calculation, but didn't understand the thermodynamic effects involved, and therefore got a result that didn't have the correct factor of $\sqrt{\gamma}$.

Contributors and Attributions

7.3: Footnotes

1. Speaking more carefully, I should say that every basic building block of light and matter has both wave and particle properties.

7.4: Problems

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

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

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

(a) the amplitude and frequency were doubled while the velocity remained the same;

(b) the frequency and velocity were both doubled while the amplitude remained unchanged;

(c) the wavelength and amplitude were reduced by a factor of three while the velocity was doubled.

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

a / Problem 3

3. (a) The graph shows the height of a water wave pulse as a function of position. Draw a graph of height as a function of time for a specific point on the water. Assume the pulse is traveling to the right.
(b) Repeat part a, but assume the pulse is traveling to the left.
(c) Now assume the original graph was of height as a function of time, and draw a graph of height as a function of position, assuming the pulse is traveling to the right.
(d) Repeat part c, but assume the pulse is traveling to the left.
Explain all your answers. [Problem by Arnold Arons.]

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

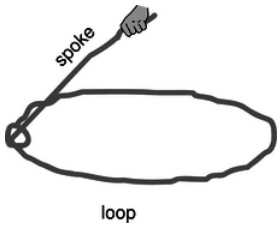
b / Problem 5.

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

[Problem by Arnold Arons.]

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

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



c / Problem 8.

8. The simplest trick with a lasso is to spin a flat loop in a horizontal plane. The whirling loop of a lasso is kept under tension mainly due to its own rotation. Although the spoke's force on the loop has an inward component, we'll ignore it. The purpose of this problem, which is based on one by A.P. French, is to prove a cute fact about wave disturbances moving around the loop. As far as I know, this fact has no practical implications for trick roping! Let the loop have radius r and mass per unit length μ, and let its angular velocity be ω.
(a) Find the tension, T, in the loop in terms of r, μ, and ω. Assume the loop is a perfect circle, with no wave disturbances on it yet. \hwhint{hwhint:lasso} \hwans{hwans:lasso}
(b) Find the velocity of a wave pulse traveling around the loop. Discuss what happens when the pulse moves is in the same direction as the rotation, and when it travels contrary to the rotation.

9. A string hangs vertically, free at the bottom and attached at the top.
(a) Find the velocity of waves on the string as a function of the distance from the bottom
(b) Find the acceleration of waves on the string.\hwans{hwans:wave-on-hanging-string}
(c) Interpret your answers to parts a and b for the case where a pulse comes down and reaches the end of the string. What happens next? Check your answer against experiment and conservation of energy.

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

11. Light travels faster in warmer air. Use this fact to explain the formation of a mirage appearing like the shiny surface of a pool of water when there is a layer of hot air above a road.

12. (a) Compute the amplitude of light that is reflected back into air at an air-water interface, relative to the amplitude of the incident wave. Assume that the light arrives in the direction directly perpendicular to the surface.The speeds of light in air and water are 3.0×10^8 and 2.2×10^8 m/s, respectively.
(b) Find the energy of the reflected wave as a fraction of the incident energy. \hwhint{hwhint:airwaterrefl}(answer check available at lightandmatter.com)

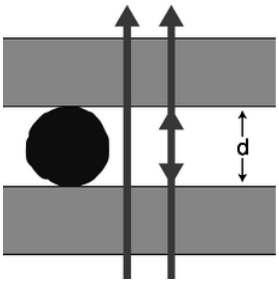
13. A concert flute produces its lowest note, at about 262 Hz, when half of a wavelength fits inside its tube. Compute the length of the flute. \hwans{hwans:flute}

14. (a) A good tenor saxophone player can play all of the following notes without changing her fingering, simply by altering the tightness of her lips: E♭ (150 Hz), E♭ (300 Hz), B♭ (450 Hz), and E♭ (600 Hz). How is this possible? (I'm not asking you to analyze the coupling between the lips, the reed, the mouthpiece, and the air column, which is very complicated.)
(b) Some saxophone players are known for their ability to use this technique to play “freak notes,” i.e., notes above the normal range of the instrument. Why isn't it possible to play notes below the normal range using this technique?

C	261.6 Hz
D	293.7
E	329.6
F	349.2
G	392.0
A	440.0
flat	466.2

Problem 15.

15. The table gives the frequencies of the notes that make up the key of F major, starting from middle C and going up through all seven notes.
(a) Calculate the first four or five harmonics of C and G, and determine whether these two notes will be consonant or dissonant. (Recall that harmonics that differ by about 1-10% cause dissonance.)
(b) Do the same for C and B♭.



d / Problem 16.

16. A Fabry-Perot interferometer, shown in the figure being used to measure the diameter of a thin filament, consists of two glass plates with an air gap between them. As the top plate is moved up or down with a screw, the light passing through the plates goes through a cycle of constructive and destructive interference, which is mainly due to interference between rays that pass straight through and those that are reflected twice back into the air gap. (Although the dimensions in this drawing are distorted for legibility, the glass plates would really be much thicker than the length of the wave-trains of light, so no interference effects would be observed due to reflections within the glass.)
(a) If the top plate is cranked down so that the thickness, d, of the air gap is much less than the wavelength λ of the light, i.e., in the limit $d \to 0$, what is the phase relationship between the two rays? (Recall that the phase can be inverted by a reflection.) Is the interference constructive, or destructive?
(b) If d is now slowly increased, what is the first value of d for which the interference is the same as at $d \to 0$? Express your answer in terms of λ.
(c) Suppose the apparatus is first set up as shown in the figure. The filament is then removed, and n cycles of brightening and dimming are counted while the top plate is brought down to $d = 0$. What is the thickness of the filament, in terms of n and λ?
Based on a problem by D.J. Raymond.

17. (a) A wave pulse moves into a new medium, where its velocity is greater by a factor α. Find an expression for the fraction, f, of the wave energy that is transmitted, in terms of α. Note that, as discussed in the text, you cannot simply find f by squaring the amplitude of the transmitted wave. \hwans{hwans:maxtransmission}
(b) Suppose we wish to transmit a pulse from one medium to another, maximizing the fraction of the wave energy transmitted. To do so, we sandwich another layer in between them, so that the wave moves from the initial medium, where its velocity is v_1, through the intermediate layer, where it is v_2, and on into the final layer, where it becomes v_3. What is the optimal value of v_2? (Assume that the middle layer is thicker than the length of the pulse, so there are no interference effects. Also, although there will be later echoes that are transmitted after multiple reflections back and forth across the middle layer, you are only to optimize the strength of the transmitted pulse that is first to emerge. In other words, it's simply a matter of applying your answer from part a twice to find the amount that finally gets through.) \hwans{hwans:maxtransmission}

Contributors and Attributions

CHAPTER OVERVIEW

8: Relativity

Contributors and Attributions

8.1: 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...”

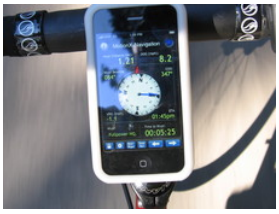


Figure a: / This Global Positioning System (GPS) system, running on a smartphone attached to a bike's handlebar, depends on Einstein's theory of relativity. Time flows at a different rate aboard a GPS satellite than it does on the bike, and the GPS software has to take this into account.

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.

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

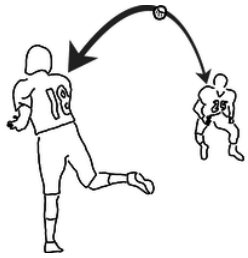
7.1.1 The correspondence principle

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.

7.1.2 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.



c / Newton's laws do not distinguish past from future. The football could travel in either direction while obeying Newton's laws.

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.

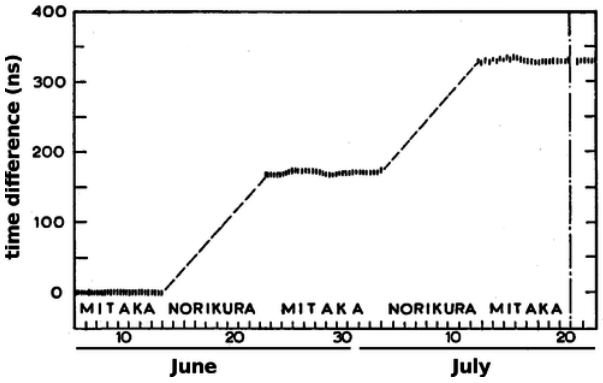
7.1.3 Time distortion arising from motion and gravity

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.

***Figure 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.*

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.



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. Cf. problem 25, p. 443.

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 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 until section 7.4, and that means that what we want to focus on right now is the distortion of time due to motion, not gravity.

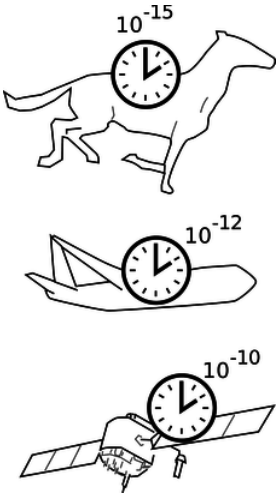


Figure f: The correspondence principle requires that the relativistic distortion of time become small for small velocities.

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.

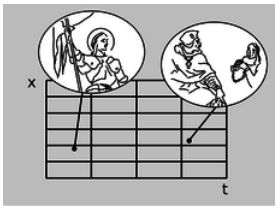
Contributors and Attributions

8.2: Distortion of Space and Time

7.2.1 The Lorentz transformation

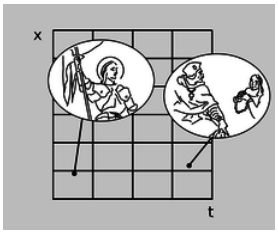
Relativity says that when two observers are in different frames of reference, each observer considers the other one's perception of time to be distorted. We'll also see that something similar happens to their observations of distances, so both space and time are distorted. What exactly is this distortion? 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 a.



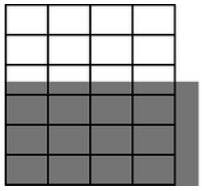
a / Two events are given as points on a graph of position versus time. Joan of Arc helps to restore Charles VII to the throne. At a later time and a different position, Joan of Arc is sentenced to death.

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 b.



b / A change of units distorts an x-t graph. This graph depicts exactly the same events as figure a. 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.

We're going to have quite a few examples of this type, so I'll adopt the convention shown in figure c for depicting them. Figure c summarizes the relationship between figures a and b in a more compact form. The gray rectangle represents the original coordinate grid of figure a, while the grid of black lines represents the new version from figure b. Omitting the grid from the gray rectangle makes the diagram easier to decode visually.



c / A convention we'll use to represent a distortion of time and space.

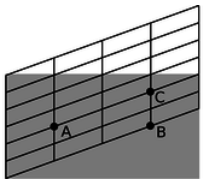
Our goal of unraveling the mysteries of special relativity amounts to nothing more than finding out how to draw a diagram like c in the case where the two different sets of coordinates represent measurements of time and space made by two different observers, each in motion relative to the other. Galileo 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 equally valid.
4. Causality holds, in the sense described on page 381.
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.

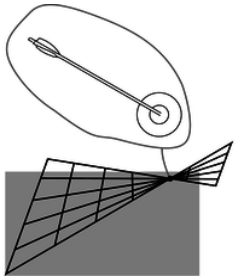


d / A Galilean version of the relationship between two frames of reference. As in all such graphs in this chapter, the original coordinates, represented by the gray rectangle, have a time axis that goes to the right, and a position axis that goes straight up.

If it were not for property 5, we could imagine that figure d 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 d 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 one answer, which is the one found by Einstein in 1905. To see why this is, let's work by a process of elimination.

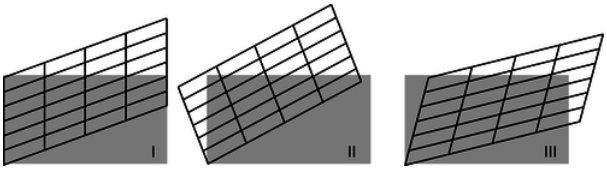


e / A transformation that leads to disagreements about whether two events occur at the same time and place. This is not just a matter of opinion. Either the arrow hit the bull's-eye or it didn't.

Figure e 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 e 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.

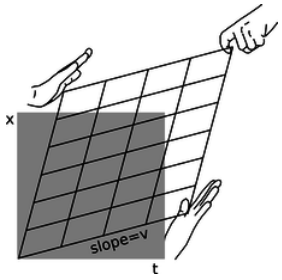
The next type of transformation we want to kill off is shown in figure f, in which the grid lines curve, but never cross one another. The trouble with this one is that it violates property 1, the uniformity of time and space. The transformation is unusually "twisty" at A, whereas at B it's much more 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 f violates property 1.

The examples in figures e and f 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 g, that have this property. Case I is the Galilean transformation of figure d on page 386, which we've already ruled out.



g / 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.

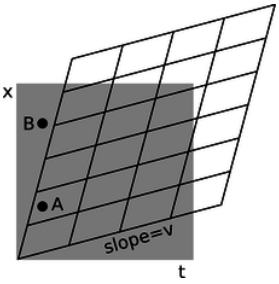
Case II can also be discarded. Here every point on the grid rotates counterclockwise. What physical parameter would determine the amount of rotation? The only thing that could be relevant would be v, the relative velocity of the motion of the two frames of reference with respect to one another. But if 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.



h / In the units that are most convenient for relativity, the transformation has symmetry about a 45-degree diagonal line.

The only remaining possibility is case III, which I've redrawn in figure h with a couple of changes. This is the one that Einstein predicted in 1905. The transformation is known as the Lorentz transformation, after HendrikLorentz (1853-1928), who partially anticipated Einstein's work, without arriving at the correct interpretation. The distortion is a kind of smooshing and stretching, as

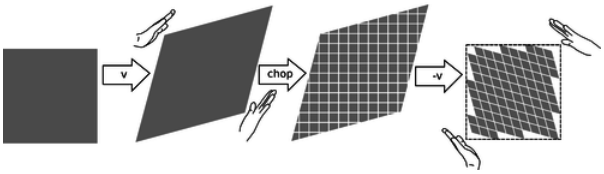
suggested by the hands. Also, we've already seen in figures a-c on page 385 that we're free to stretch or compress everything as much as we like in the horizontal and vertical directions, because this simply corresponds to choosing different units of measurement for time and distance. In figure h I've 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.



i / 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.

As in our discussion of the Galilean transformation, slopes are interpreted as velocities, and the slope of the near-horizontal lines in figure i 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 simultaneous 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 i 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 h 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 j.



j / 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 385, which states that all directions in space have the same properties, so $R(v) = 1$.

7.2.2 The γ factor

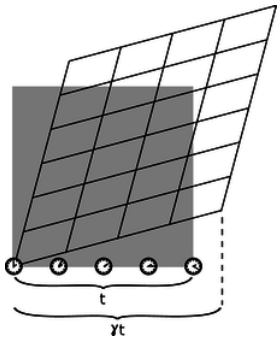
With a little algebra and geometry (homework problem 7, page 439), one can use the equal-area property to show that the factor γ (Greek letter gamma) defined in figure k is given by the equation

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

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. 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

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

Here's why we care about γ. Figure k 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:

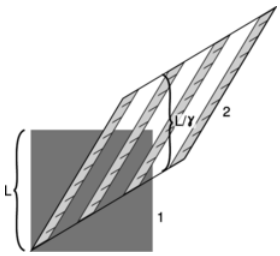


k / The γ factor.

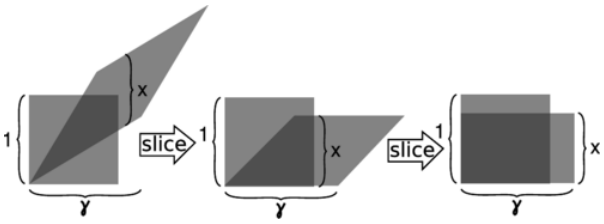
Time dilation

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.

As proved in figures l and m, lengths are also distorted:



l / 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 m, the length contracts from L to L/γ.



m / This figure proves, as claimed in figure l, 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.

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.

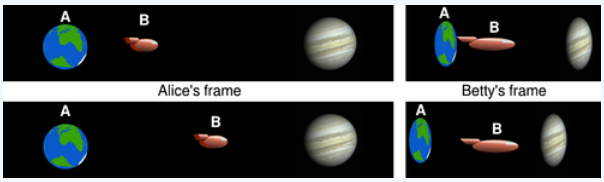
Exercise 8.2.1

What is γ when $v = 0$? What does this mean?

(answer in the back of the PDF version of the book)

Example 8.2.1: An interstellar road trip

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.



n / Example 1.

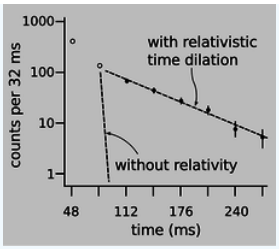
Betty experiences time dilation. At this speed, her γ is 2.0, so that the voyage will only seem to her to last 7 years. But there is perfect symmetry between Alice's and Betty's frames of reference, so Betty agrees with Alice on their relative speed; Betty sees herself as being at rest, while the sun 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.

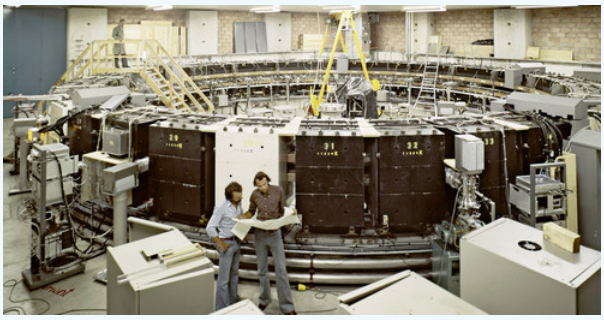
Example 8.2.2: Large time dilation

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 p shows a 1974 experiment[2] 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 2.197 μs 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 p, 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 o 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.



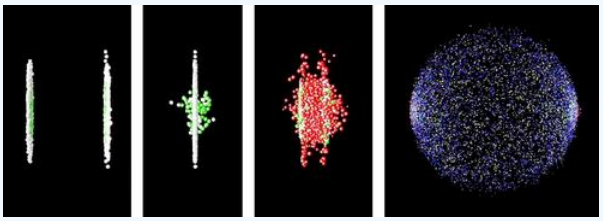
o / Muons accelerated to nearly c undergo radioactive decay much more slowly than they would according to an observer at rest with respect to the muons. The first two data-points (unfilled circles) were subject to large systematic errors.



p / Apparatus used for the test of relativistic time dilation described in example 2. The prominent black and white blocks are large magnets surrounding a circular pipe with a vacuum inside. (c) 1974 by CERN.

Example 8.2.3: An example of length contraction

Figure q shows an artist's rendering of the length contraction for the collision of two gold nuclei at relativistic speeds in the RHIC accelerator in Long Island, New York, which went on line in 2000. 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, in which experimenters hope to observe a new form of matter.

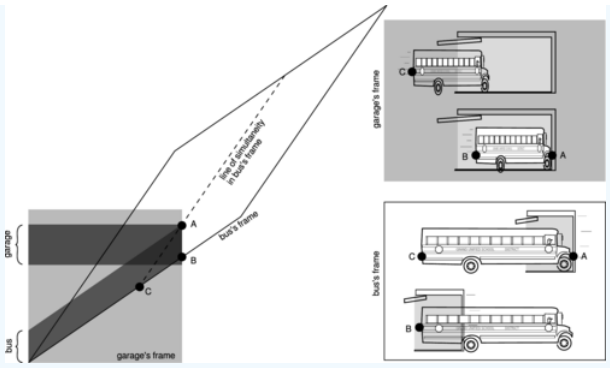


q / Colliding nuclei show relativistic length contraction.

Example 8.2.1: The garage paradox

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 r, 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.



r / Example 4: 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.

7.2.3 The universal speed c

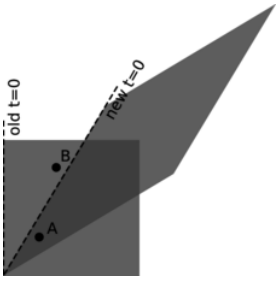
Let's think a little more about the role of the 45-degree diagonal in the Lorentz transformation. Slopes on these graphs are interpreted as velocities. This line has a slope of 1 in relativistic units, but that slope corresponds to c in ordinary metric units. We already know that the relativistic distance 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.

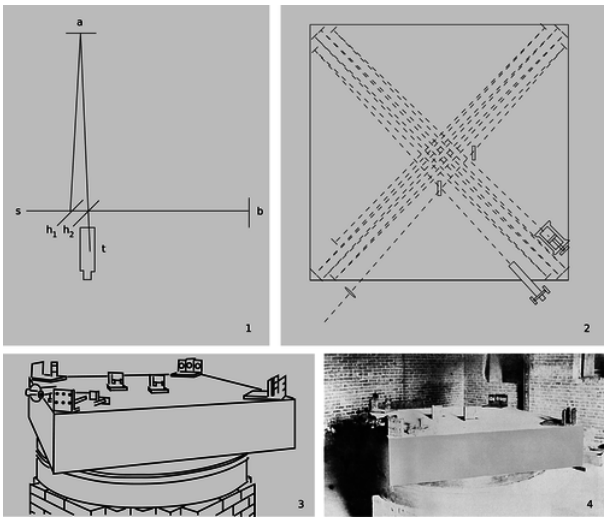
A universal speed limit

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 438). 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 s.



s / 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.



t / 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.

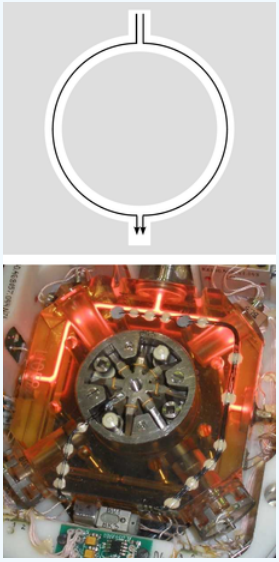
Now consider a beam of light. We're used to talking casually about the "speed of light," but what does that really mean? Motion is relative, so normally if we want to talk about a velocity, we have to specify what it's measured relative to. A sound wave has a certain speed relative to the air, and a 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.

Example 8.2.4: The ring laser gyroscope

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 u 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.



u / A ring laser gyroscope.

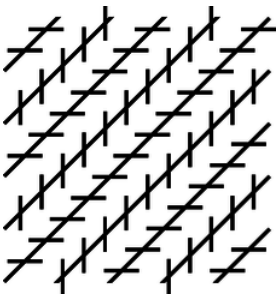
Example 8.2.6: No frequency-dependence

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 observation, 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

◊ 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?

◊



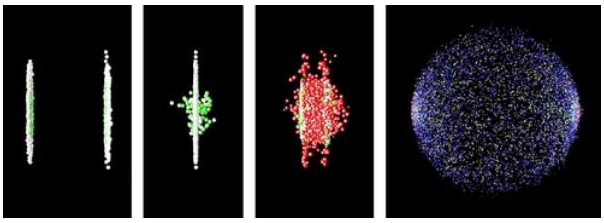
Discussion question B.

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?

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

◊ 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?

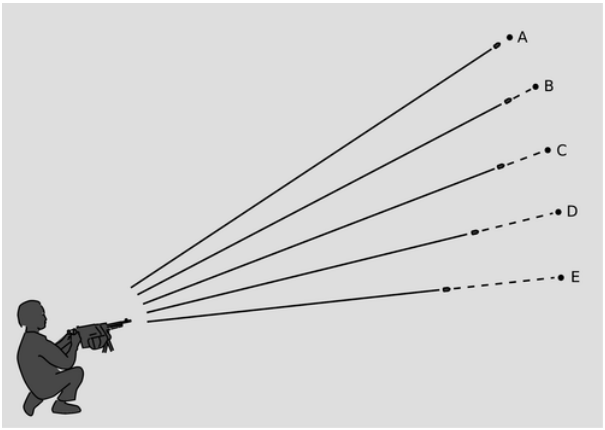
◊ Figure q from page 392, depicting the collision of two nuclei at the RHIC accelerator, is reproduced below. What would the shapes of the two nuclei look like to a microscopic observer riding on the left-hand nucleus? To an observer riding on the right-hand one? Can they agree on what is happening? 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?



v / Discussion question E: colliding nuclei show relativistic length contraction.

◊ 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?

◊ The machine-gunner in the figure sends out a spray of bullets. 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. The chain of impacts extends across space at a speed greater than c. Does this violate special relativity?



Discussion question G.

7.2.4 No action at a distance

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...”

Time delays in forces exerted at a distance

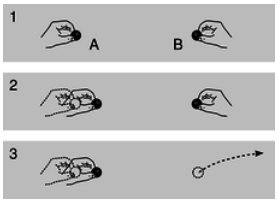
Relativity forbids Newton's instantaneous action at a distance. For suppose that instantaneous action at a distance existed. It would then be possible to send signals from one place in the universe to another without any time lag. This would allow perfect synchronization of all clocks. But the Hafele-Keating 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 (ch. 10-11). 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 7.2.3 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.

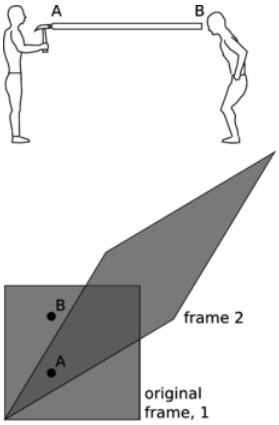
Even empty space, then, is not perfectly featureless. It has measurable properties. For example, we can drop a rock in order to measure the direction of the gravitational field, or use a magnetic compass to find the direction of the magnetic field. This concept made a deep impression on Einstein as a child. He recalled that as a five-year-old, the gift of a magnetic compass convinced him that there was “something behind things, something deeply hidden.”

More evidence that fields of force are real: they carry energy.



w / Fields carry energy.

The smoking-gun argument for this strange notion of traveling force ripples comes from the fact that they carry energy. In figure x/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.



x / Discussion question E.

If Alice chooses to move her ball closer to Betty's, x/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.

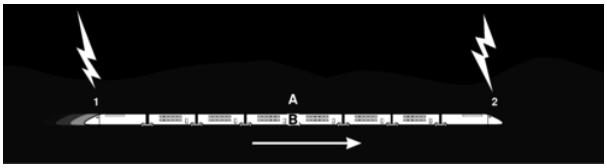
But now suppose, x/3, that Betty decides to play a trick on Alice by tossing B far away just as Alice is getting ready to move A. We have already established that Alice can't feel B's motion instantaneously, so the electric forces must actually be propagated by an electric *field*. Of course this experiment 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

◊ 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?

◊ 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.

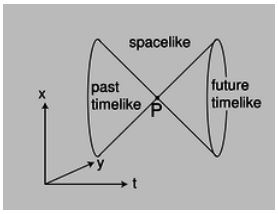


◊ Resolve the following paradox by drawing a spacetime diagram (i.e., a graph of x versus t). Andy and Beth are in motion relative to one another at a significant fraction of c. As they pass by each other, they exchange greetings, and Beth tells Andy that she is going to blow up a stick of dynamite one 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?

◊ 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?

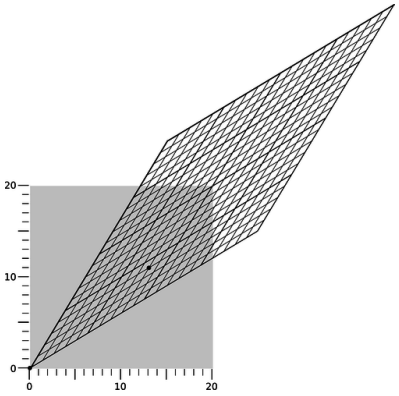
◊ The rod in the figure is perfectly rigid. At event A, the hammer strikes one end of the rod. At event B, the other end moves. Since the rod is perfectly rigid, it can't compress, so A and B are simultaneous. In frame 2, B happens before A. Did the motion at the right end *cause* the person on the left to decide to pick up the hammer and use it?

7.2.5 The light cone



y / 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 aa.



aa / Example 9.

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. \myoptionalsubsection[2]{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, $\mathrm{AB} \cong \mathrm{CD}$, 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 $\mathrm{AB} \cong \mathrm{CD}$ 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 1 on p. 391 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 $\overline{\mathrm{AB}}^2 = \Delta x^2 + \Delta y^2$.

Definition of the interval in relativity

Given points separated by coordinate differences Δx, Δy, Δz, 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 $\overline{\mathrm{AB}}$ 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 8 below, it is the same regardless of our frame of reference, i.e., regardless of our choice of coordinates.

Example 8.2.7: Pioneer 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 X 10^{12} m	3.951 X 10^{12} m	0.237 X 10^{12} m
3.7869120000 X 10^{8} s	2.420 X 10^{12} m	8.827 X 10^{12} m	0.488 X 10^{12} m

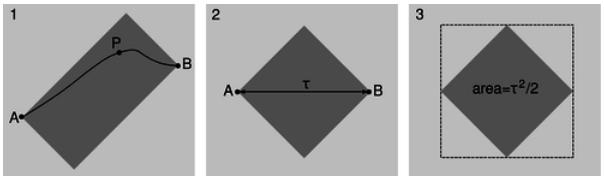
Compare the time elapsed on the spacecraft to the time in a frame of reference tied to the sun.

▷ 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 X 10^{8} s	0.212 X 10^{4} s	1.626 X 10^{4} s	0.084 X 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.

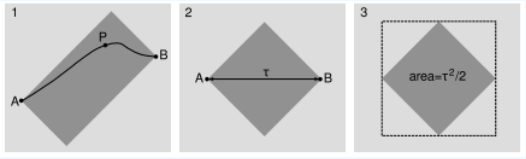


z / Light-rectangles, example 8.

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 $\tau^2/2$.

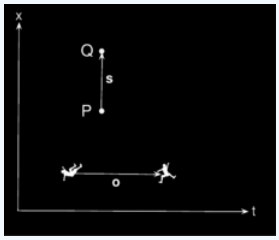
Example 8.2.8: Invariance of the interval

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 ab/1. On p. 389 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 ab/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 ab/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.



ab / Example 8.

Example 8.2.9: A numerical example of invariance



ac / Example 9.

Figure ac 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

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

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

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

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

Four-vectors and the inner product

Example 7 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.

Example 10: The twin paradox

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 1, p. 391).

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. 381).

Betty's track in the x-t plane (her “world-line” in relativistic jargon) consists of vectors $\mathbf{b}$ and $\mathbf{c}$ strung end-to-end (figure ad). We could adopt a frame of reference in which Betty was at rest during $\mathbf{b}$ (i.e., $b_x = 0$), but there is no frame in which $\mathbf{b}$ and $\mathbf{c}$ are parallel, so there is no frame in which Betty was at rest during *both* $\mathbf{b}$ and $\mathbf{c}$. This resolves the paradox.

We have already established by other methods that Betty ages less that 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 ad are

$$\begin{aligned}\mathbf{a} &= (6, 1)\\ \mathbf{b} &= (3, 2)\\ \mathbf{c} &= (3, -1),\end{aligned}$$

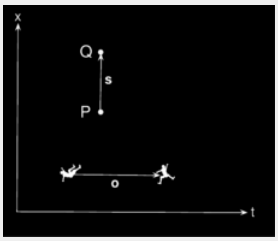
where the components are given in the order (t, x). The time experienced by Alice is then

$$|\mathbf{a}| = \sqrt{6^2 - 1^2} = 5.9,$$

which is greater than the Betty's elapsed time

$$|\mathbf{b}| + |\mathbf{c}| = \sqrt{3^2 - 2^2} + \sqrt{3^2 - (-1)^2} = 5.1.$$

Example 11: Simultaneity using inner products

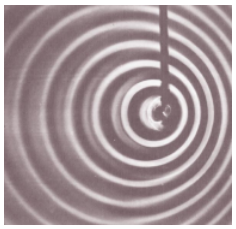


ac / Example 11.

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.

Doppler shifts of light and addition of velocities

When Doppler shifts happen to ripples on a pond or the sound waves from an airplane, they can depend on the relative motion of three different objects: the source, the receiver, and the medium. But light waves don't have a medium. Therefore Doppler shifts of light can only depend on the relative motion of the source and observer.



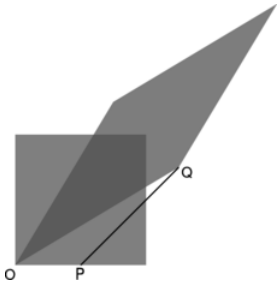
ad / The pattern of waves made by a point source moving to the right across the water. Note the shorter wavelength of the forward-emitted waves and the longer wavelength of the backward-going ones.

One simple case is the one in which the relative motion of the source and the receiver is perpendicular to the line connecting them. That is, the motion is transverse. Nonrelativistic Doppler shifts happen because the distance between the source and receiver is changing, so in nonrelativistic physics we don't expect any Doppler shift at all when the motion is transverse, and this is what is in fact observed to high precision. For example, the photo shows shortened and lengthened wavelengths to the right and left, along the source's line of motion, but an observer above or below the source measures just the normal, unshifted wavelength and frequency. But relativistically, we have a time dilation effect, so for light waves emitted transversely, there is a Doppler shift of $1/\gamma$ in frequency (or γ in wavelength).

The other simple case is the one in which the relative motion of the source and receiver is longitudinal, i.e., they are either approaching or receding from one another. For example, distant galaxies are receding from our galaxy due to the expansion of the universe, and this expansion was originally detected because Doppler shifts toward the red (low-frequency) end of the spectrum were observed.

Nonrelativistically, we would expect the light from such a galaxy to be Doppler shifted down in frequency by some factor, which would depend on the relative velocities of three different objects: the source, the wave's medium, and the receiver. Relativistically, things get simpler, because light isn't a vibration of a physical medium, so the Doppler shift can only depend on a single velocity v, which is the rate at which the separation between the source and the receiver is increasing.

ae / A graphical representation of the Lorentz transformation for a velocity of $(3/5)c$. The long diagonal is stretched by a factor of two, the short one is half its former length, and the area is the same as before.



af / At event O, the source and the receiver are on top of each other, so as the source emits a wave crest, it is received without any time delay. At P, the source emits another wave crest, and at Q the receiver receives it.

The square in figure af is the “graph paper” used by someone who considers the source to be at rest, while the parallelogram plays a similar role for the receiver. The figure is drawn for the case where $v = 3/5$ (in units where $c = 1$), and in this case the stretch factor of the long diagonal is 2. To keep the area the same, the short diagonal has to be squished to half its original size. But now it's a matter of simple geometry to show that OP equals half the width of the square, and this tells us that the Doppler shift is a factor of 1/2 in frequency. That is, the squish factor of the short diagonal is interpreted as the Doppler shift. To get this as a general equation for velocities other than 3/5, one can show by straightforward fiddling with the result of part c of problem 7 on p. 439 that the Doppler shift is

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

Here $v > 0$ is the case where the source and receiver are getting farther apart, $v < 0$ the case where they are approaching. (This is the opposite of the sign convention used in subsection 6.1.5. It is convenient to change conventions here so that we can use positive

values of v in the case of cosmological red-shifts, which are the most important application.)

Suppose that Alice stays at home on earth while her twin Betty takes off in her rocket ship at 3/5 of the speed of light. When I first learned relativity, the thing that caused me the most pain was understanding how each observer could say that the other was the one whose time was slow. It seemed to me that if I could take a pill that would speed up my mind and my body, then naturally I would see everybody *else* as being slow. Shouldn't the same apply to relativity? But suppose Alice and Betty get on the radio and try to settle who is the fast one and who is the slow one. Each twin's voice sounds slooooowed doooowwwwn to the other. If Alice claps her hands twice, at a time interval of one second by her clock, Betty hears the hand-claps coming over the radio two seconds apart, but the situation is exactly symmetric, and Alice hears the same thing if Betty claps. Each twin analyzes the situation using a diagram identical to ah, and attributes her sister's observations to a complicated combination of time distortion, the time taken by the radio signals to propagate, and the motion of her twin relative to her.

self-check:

Turn your book upside-down and reinterpret figure ah.

(answer in the back of the PDF version of the book)

Example 12: A symmetry property of the Doppler effect

Suppose that A and B are at rest relative to one another, but C is moving along the line between A and B. A transmits a signal to C, who then retransmits it to B. The signal accumulates two Doppler shifts, and the result is their product $D(v)D(-v)$. But this product must equal 1, so we must have $D(-v)D(v)=1$, which can be verified directly from the equation.

Example 13: The Ives-Stilwell experiment

The result of example 12 was the basis of one of the earliest laboratory tests of special relativity, by Ives and Stilwell in 1938. They observed the light emitted by excited by a beam of H_2^+ and H_3^+ ions with speeds of a few tenths of a percent of c. Measuring the light from both ahead of and behind the beams, they found that the product of the Doppler shifts $D(v)D(-v)$ was equal to 1, as predicted by relativity. If relativity had been false, then one would have expected the product to differ from 1 by an amount that would have been detectable in their experiment. In 2003, Saathoff et al. carried out an extremely precise version of the Ives-Stilwell technique with Li^+ ions moving at 6.4% of c. The frequencies observed, in units of MHz, were:

f_o	= 546466918.8±0.4
	(unshifted frequency)
$f_oD(-v)$	= 582490203.44±.09
	(shifted frequency, forward)
$f_o\,D(v)$	= 512671442.9±0.5
	(shifted frequency, backward)
$\sqrt{f_oD(-v)\cdot f_oD(v)}$	=546466918.6±0.3

The results show incredibly precise agreement between f_o and $\sqrt{f_oD(-v)\cdot f_oD(v)}$, as expected relativistically because $D(v)D(-v)$ is supposed to equal 1. The agreement extends to 9 significant figures, whereas if relativity had been false there should have been a relative disagreement of about $v^2=.004$, i.e., a discrepancy in the third significant figure. The spectacular agreement with theory has made this experiment a lightning rod for anti-relativity kooks.

We saw on p. 394 that relativistic velocities should not be expected to be exactly additive, and problem 1 on p. 438 verifies this in the special case where A moves relative to B at $0.6c$ and B relative to C at $0.6c$ --- the result *not* being $1.2c$. The relativistic Doppler shift provides a simple way of deriving a general equation for the relativistic combination of velocities; problem 17 on p. 442 guides you through the steps of this derivation, and the result is given on p. 936.

Contributors and Attributions

8.3: 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 $E = mc^2$.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

8.4: General Relativity (optional)

What you've learned so far about relativity is known as the special theory of relativity, which is compatible with three of the four known forces of nature: electromagnetism, the strong nuclear force, and the weak nuclear force. Gravity, however, can't be shoehorned into the special theory. In order to make gravity work, Einstein had to generalize relativity. The resulting theory is known as the general theory of relativity.[5]

Contributors and Attributions

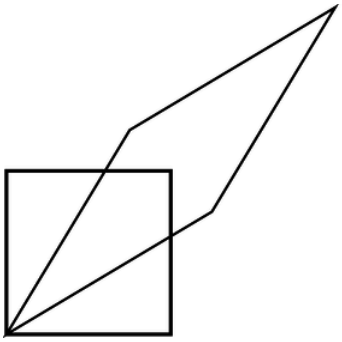
Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

8.5: Footnotes

1. A. Einstein, “On the Electrodynamics of Moving Bodies,” *Annalen der Physik* 17 (1905), p. 891, tr. Saha and Bose.
2. Bailey at al., Nucl. Phys. B150(1979) 1
3. http://arxiv.org/abs/0908.1832
4. A 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.
5. Einstein originally described the distinction between the two theories by saying that the special theory applied to nonaccelerating frames of reference, while the general one allowed any frame at all. The modern consensus is that Einstein was misinterpreting his own theory, and that special relativity actually handles accelerating frames just fine.
6. arxiv.org/abs/1310.8214

8.E: Relativity (Exercises)

1. The figure illustrates a Lorentz transformation using the conventions employed in section 7.2. 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.(answer check available at lightandmatter.com)



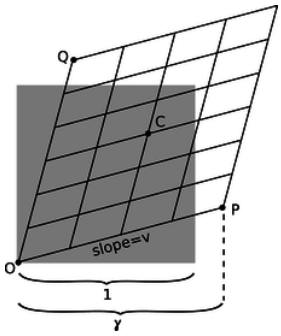
2. Astronauts in three different spaceships are communicating with each other. Those aboard ships A and B agree on the rate at which time is passing, but they disagree with the ones on ship C.
(a) Alice is aboard ship A. How does she describe the motion of her own ship, in its frame of reference?
(b) 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.)(answer check available at lightandmatter.com)

5. In example 2 on page 391, 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.(answer check available at lightandmatter.com)

6. 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.(answer check available at lightandmatter.com)



a / Problem 7.

7. In this homework problem, you'll fill in the steps of the algebra required in order to find the equation for γ on page 389. To keep the algebra simple, let the time t in figure k 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 $(\gamma, v\gamma)$. (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 $\gamma = 1/\sqrt{1-v^2}$.(answer check available at lightandmatter.com)

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

neutron	1.67495×10^{-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. (answer check available at lightandmatter.com)
(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 (see chapter 7). A neutrino, like an antineutrino, has negligible mass.

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

9. (a) Find a relativistic equation for the velocity of an object in terms of its mass and momentum (eliminating γ).(answer check available at lightandmatter.com)
(b) Show that your result is approximately the same as the classical value, p/m, at low velocities.
(c) Show that very large momenta result in speeds close to the speed of light.

10. (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.

11. 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 18 on page 418, but instead of very low, nonrelativistic 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.(answer check available at lightandmatter.com)
(d) Simplify your answer to part c for the case where the speed is ultrarelativistic.(answer check available at lightandmatter.com)
(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 $\mathrm{kg{\cdot}m/s}$.(answer check available at lightandmatter.com)
(f) Discuss how your answer in part e relates to the correspondence principle.

12. As discussed in chapter 6, the speed at which a disturbance travels along a string under tension is given by $v=\sqrt{T/\mu}$, where μ is the mass per unit length, and T is the tension.
(a) Suppose a string has a density ρ, and a cross-sectional area A. Find an expression for the maximum tension that could possibly exist in the string without producing $v>c$, which is impossible according to relativity. Express your answer in terms of ρ, A, and c. The interpretation is that relativity puts a limit on how strong any material can be.(answer check available at lightandmatter.com)
(b) Every substance has a tensile strength, defined as the force per unit area required to break it by pulling it apart. The tensile strength is measured in units of $\mathrm{N/m^2}$, which is the same as the pascal (Pa), the mks unit of pressure. Make a numerical estimate of the maximum tensile strength allowed by relativity in the case where the rope is made out of ordinary matter, with a density on the same order of magnitude as that of water. (For comparison, kevlar has a tensile strength of about 4×10^9 Pa, and there is speculation that fibers made from carbon nanotubes could have values as high as 6×10^{10} Pa.)(answer check available at lightandmatter.com)
(c) A black hole is a star that has collapsed and become very dense, so that its gravity is too strong for anything ever to escape from it. For instance, the escape velocity from a black hole is greater than c, so a projectile can't be shot out of it. Many people, when they hear this description of a black hole in terms of an escape velocity greater than c, wonder why it still wouldn't be possible to extract an object from a black hole by other means than launching it out as a projectile. For example, suppose we lower an astronaut into a black hole on a rope, and then pull him back out again. Why might this not work?

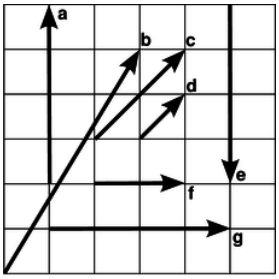
13. (a) A charged particle is surrounded by a uniform electric field. Starting from rest, it is accelerated by the field to speed v after traveling a distance d. Now it is allowed to continue for a further distance $3d$, for a total displacement from the start of $4d$. What speed will it reach, assuming classical physics?
(b) Find the relativistic result for the case of $v=c/2$.

14. Problem 14 has been deleted.

15. Expand the equation $K=m(\gamma-1)$ in a Taylor series, and find the first two nonvanishing terms. Show that the first term is the nonrelativistic expression for kinetic energy.

16. Expand the relativistic equation for momentum in a Taylor series, and find the first two nonvanishing terms. Show that the first term is the classical expression.

17. (solution in the pdf version of the book) As promised in subsection 7.2.8, this problem will lead you through the steps of finding an equation for the combination of velocities in relativity, generalizing the numerical result found in problem 1. Suppose that A moves relative to B at velocity u, and B relative to C at v. We want to find A's velocity w relative to C, in terms of u and v. Suppose that A emits light with a certain frequency. This will be observed by B with a Doppler shift $D(u)$. C detects a further shift of $D(v)$ relative to B. We therefore expect the Doppler shifts to multiply, $D(w)=D(u)D(v)$, and this provides an implicit rule for determining w if u and v are known. (a) Using the expression for D given in section 7.2.8, write down an equation relating u, v, and w. (b) Solve for w in terms of u and v. (c) Show that your answer to part b satisfies the correspondence principle.



b / Problem 18.

18. 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?

19. 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?

20. 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.

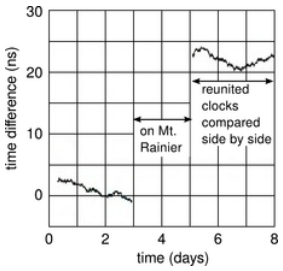
Example a href="#eg:no-pair-prod-in-vacuum">23 on p. 420 shows that pair production can't occur in a vacuum due to conservation of the energy-momentum four-vector. What about the other two processes? Can the photoelectric effect occur without the presence of some third particle such as an atomic nucleus? Can Compton scattering happen without a third particle?

21. 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.

22. Prove, as claimed in the caption of figure a on p. 424, that $S-180^\circ = 4(s-180^\circ)$, where S is the sum of the angles of the large equilateral triangle and s is the corresponding sum for one of the four small ones.(solution in the pdf version of the book)

23. If a two-dimensional being lived on the surface of a cone, would it say that its space was curved, or not?

24. (a) Verify that the equation $1-gh/c^2$ for the gravitational Doppler shift and gravitational time dilation has units that make sense. (b) Does this equation satisfy the correspondence principle?



c / Problem 25b. Redrawn from Van Baak, Physics Today 60 (2007) 16.

25. (a) Calculate the Doppler shift to be expected in the Pound-Rebka experiment described on p. 429. (b) In the 1978 Iijima mountain-valley experiment (p. 384), analysis was complicated by the clock's sensitivity to pressure, humidity, and temperature. A cleaner version of the experiment was done in 2005 by hobbyist Tom Van Baak. He put his kids and three of his atomic clocks in a minivan and drove from Bellevue, Washington to a lodge on Mount Rainier, 1340 meters higher in elevation. At home, he compared the clocks to others that had stayed at his house. Verify that the effect shown in the graph is as predicted by general relativity.

26. The International Space Station orbits at an altitude of about 350 km and a speed of about 8000 m/s relative to the ground. Compare the gravitational and kinematic time dilations. Over all, does time run faster on the ISS than on the ground, or more slowly?

27. Section 7.4.3 presented a Newtonian estimate of how compact an object would have to be in order to be a black hole. Although this estimate is not really right, it turns out to give the right answer to within about a factor of 2. To roughly what size would the earth have to be compressed in order to become a black hole?

28. Clock A sits on a desk. Clock B is tossed up in the air from the same height as the desk and then comes back down. Compare the elapsed times. \hwhint{hwhint:tossed-clock} (solution in the pdf version of the book)

29. The angular defect d of a triangle (measured in radians) is defined as $s-\pi$, where s is the sum of the interior angles. The angular defect is proportional to the area A of the triangle. Consider the geometry measured by a two-dimensional being who lives

on the surface of a sphere of radius R. First find some triangle on the sphere whose area and angular defect are easy to calculate. Then determine the general equation for d in terms of A and R.(answer check available at lightandmatter.com)

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

CHAPTER OVERVIEW

9: Atoms and Electromagnetism

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

9.1: The Electric Glue

Where the telescope ends, the microscope begins. Which of the two has the grander view? -- *Victor Hugo*

His father died during his mother's pregnancy. Rejected by her as a boy, he was packed off to boarding school when she remarried. He himself never married, but in middle age he formed an intense relationship with a much younger man, a relationship that he terminated when he underwent a psychotic break. Following his early scientific successes, he spent the rest of his professional life mostly in frustration over his inability to unlock the secrets of alchemy.

The man being described is Isaac Newton, but not the triumphant Newton of the standard textbook hagiography. Why dwell on the sad side of his life? To the modern science educator, Newton's lifelong obsession with alchemy may seem an embarrassment, a distraction from his main achievement, the creation the modern science of mechanics. To Newton, however, his alchemical researches were naturally related to his investigations of force and motion. What was radical about Newton's analysis of motion was its universality: it succeeded in describing both the heavens and the earth with the same equations, whereas previously it had been assumed that the sun, moon, stars, and planets were fundamentally different from earthly objects. But Newton realized that if science was to describe all of nature in a unified way, it was not enough to unite the human scale with the scale of the universe: he would not be satisfied until he fit the microscopic universe into the picture as well.

It should not surprise us that Newton failed. Although he was a firm believer in the existence of atoms, there was no more experimental evidence for their existence than there had been when the ancient Greeks first posited them on purely philosophical grounds. Alchemy labored under a tradition of secrecy and mysticism. Newton had already almost single-handedly transformed the fuzzy-headed field of “natural philosophy” into something we would recognize as the modern science of physics, and it would be unjust to criticize him for failing to change alchemy into modern chemistry as well. The time was not ripe. The microscope was a new invention, and it was cutting-edge science when Newton's contemporary Hooke discovered that living things were made out of cells.

8.1.1 The quest for the atomic force

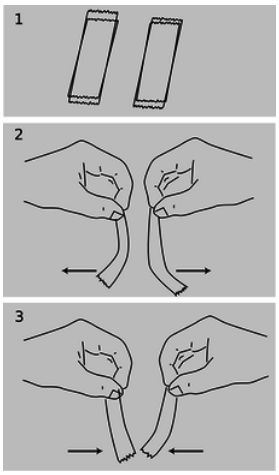
Newton was not the first of the age of reason. He was the last of the magicians. -- *John Maynard Keynes*

Newton's quest

Nevertheless it will be instructive to pick up Newton's train of thought and see where it leads us with the benefit of modern hindsight. In uniting the human and cosmic scales of existence, he had reimagined both as stages on which the actors were objects (trees and houses, planets and stars) that interacted through attractions and repulsions. He was already convinced that the objects inhabiting the microworld were atoms, so it remained only to determine what kinds of forces they exerted on each other.

His next insight was no less brilliant for his inability to bring it to fruition. He realized that the many human-scale forces --- friction, sticky forces, the normal forces that keep objects from occupying the same space, and so on --- must all simply be expressions of a more fundamental force acting between atoms. Tape sticks to paper because the atoms in the tape attract the atoms in the paper. My house doesn't fall to the center of the earth because its atoms repel the atoms of the dirt under it.

Here he got stuck. It was tempting to think that the atomic force was a form of gravity, which he knew to be universal, fundamental, and mathematically simple. Gravity, however, is always attractive, so how could he use it to explain the existence of both attractive and repulsive atomic forces? The gravitational force between objects of ordinary size is also extremely small, which is why we never notice cars and houses attracting us gravitationally. It would be hard to understand how gravity could be responsible for anything as vigorous as the beating of a heart or the explosion of gunpowder. Newton went on to write a million words of alchemical notes filled with speculation about some other force, perhaps a “divine force” or “vegetative force” that would for example be carried by the sperm to the egg.



a / Four pieces of tape are prepared, 1, as described in the text. Depending on which combination is tested, the interaction can be either repulsive, 2, or attractive, 3.

Luckily, we now know enough to investigate a different suspect as a candidate for the atomic force: electricity. Electric forces are often observed between objects that have been prepared by rubbing (or other surface interactions), for instance when clothes rub against each other in the dryer. A useful example is shown in figure a/1: stick two pieces of tape on a tabletop, and then put two more pieces on top of them. Lift each pair from the table, and then separate them. The two top pieces will then repel each other, a/2, as will the two bottom pieces. A bottom piece will attract a top piece, however, a/3. Electrical forces like these are similar in certain ways to gravity, the other force that we already know to be fundamental:

- Electrical forces are *universal*. Although some substances, such as fur, rubber, and plastic, respond more strongly to electrical preparation than others, all matter participates in electrical forces to some degree. There is no such thing as a "nonelectric" substance. Matter is both inherently gravitational and inherently electrical.
- Experiments show that the electrical force, like the gravitational force, is an *inverse square* force. That is, the electrical force between two spheres is proportional to $1/r^2$, where r is the center-to-center distance between them.

Furthermore, electrical forces make more sense than gravity as candidates for the fundamental force between atoms, because we have observed that they can be either attractive or repulsive.

8.1.2 Charge, electricity and magnetism

Charge

"Charge" is the technical term used to indicate that an object has been prepared so as to participate in electrical forces. This is to be distinguished from the common usage, in which the term is used indiscriminately for anything electrical. For example, although we speak colloquially of "charging" a battery, you may easily verify that a battery has no charge in the technical sense, e.g., it does not exert any electrical force on a piece of tape that has been prepared as described in the previous section.

Two types of charge

We can easily collect reams of data on electrical forces between different substances that have been charged in different ways. We find for example that cat fur prepared by rubbing against rabbit fur will attract glass that has been rubbed on silk. How can we make any sense of all this information? A vast simplification is achieved by noting that there are really only two types of charge. Suppose we pick cat fur rubbed on rabbit fur as a representative of type A, and glass rubbed on silk for type B. We will now find that there is no "type C." Any object electrified by any method is either A-like, attracting things A attracts and repelling those it repels, or B-like, displaying the same attractions and repulsions as B. The two types, A and B, always display opposite interactions. If A displays an attraction with some charged object, then B is guaranteed to undergo repulsion with it, and vice-versa.

The coulomb

Although there are only two types of charge, each type can come in different amounts. The metric unit of charge is the coulomb (rhymes with "drool on"), defined as follows:

One Coulomb (C) is the amount of charge such that a force of 9.0×10^9 N occurs between two point-like objects with charges of 1 C separated by a distance of 1 m.

The notation for an amount of charge is q. The numerical factor in the definition is historical in origin, and is not worth memorizing. The definition is stated for point-like, i.e., very small, objects, because otherwise different parts of them would be at different distances from each other.

A model of two types of charged particles

Experiments show that all the methods of rubbing or otherwise charging objects involve two objects, and both of them end up getting charged. If one object acquires a certain amount of one type of charge, then the other ends up with an equal amount of the other type. Various interpretations of this are possible, but the simplest is that the basic building blocks of matter come in two flavors, one with each type of charge. Rubbing objects together results in the transfer of some of these particles from one object to the other. In this model, an object that has not been electrically prepared may actually possesses a great deal of *both* types of charge, but the amounts are equal and they are distributed in the same way throughout it. Since type A repels anything that type B attracts, and vice versa, the object will make a total force of zero on any other object. The rest of this chapter fleshes out this model and discusses how these mysterious particles can be understood as being internal parts of atoms.

Use of positive and negative signs for charge

Because the two types of charge tend to cancel out each other's forces, it makes sense to label them using positive and negative signs, and to discuss the *total* charge of an object. It is entirely arbitrary which type of charge to call negative and which to call positive. Benjamin Franklin decided to describe the one we've been calling “A” as negative, but it really doesn't matter as long as everyone is consistent with everyone else. An object with a total charge of zero (equal amounts of both types) is referred to as electrically $neutral$.

self-check:

Criticize the following statement: “There are two types of charge, attractive and repulsive.”

(answer in the back of the PDF version of the book)

A large body of experimental observations can be summarized as follows:

Coulomb's law: The magnitude of the force acting between pointlike charged objects at a center-to-center distance r is given by the equation

$$|\mathbf{F}| = k\frac{|q_1||q_2|}{r^2},$$

where the constant k equals 9.0×10^9 $\mathrm{N{\cdot}m^2/C^2}$. The force is attractive if the charges are of different signs, and repulsive if they have the same sign.

Clever modern techniques have allowed the $1/r^2$ form of Coulomb's law to be tested to incredible accuracy, showing that the exponent is in the range from 1.9999999999999998 to 2.0000000000000002.

Note that Coulomb's law is closely analogous to Newton's law of gravity, where the magnitude of the force is Gm_1m_2/r^2, except that there is only one type of mass, not two, and gravitational forces are never repulsive. Because of this close analogy between the two types of forces, we can recycle a great deal of our knowledge of gravitational forces. For instance, there is an electrical equivalent of the shell theorem: the electrical forces exerted externally by a uniformly charged spherical shell are the same as if all the charge was concentrated at its center, and the forces exerted internally are zero.

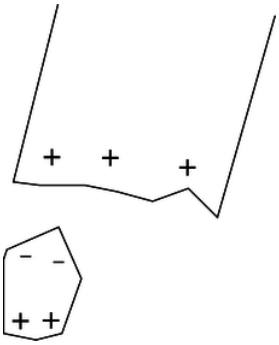
Conservation of charge

An even more fundamental reason for using positive and negative signs for electrical charge is that experiments show that charge is conserved according to this definition: in any closed system, the total amount of charge is a constant. This is why we observe that rubbing initially uncharged substances together always has the result that one gains a certain amount of one type of charge, while the other acquires an equal amount of the other type. Conservation of charge seems natural in our model in which matter is made of positive and negative particles. If the charge on each particle is a fixed property of that type of particle, and if the particles themselves can be neither created nor destroyed, then conservation of charge is inevitable.

Electrical forces involving neutral objects

b / A charged piece of tape attracts uncharged pieces of paper from a distance, and they leap up to it.

As shown in figure b, an electrically charged object can attract objects that are uncharged. How is this possible? The key is that even though each piece of paper has a total charge of zero, it has at least some charged particles in it that have some freedom to move. Suppose that the tape is positively charged, c. Mobile particles in the paper will respond to the tape's forces, causing one end of the paper to become negatively charged and the other to become positive. The attraction between the paper and the tape is now stronger than the repulsion, because the negatively charged end is closer to the tape.



c / The paper has zero total charge, but it does have charged particles in it that can move.

self-check:

What would have happened if the tape was negatively charged? (answer in the back of the PDF version of the book)

The path ahead

We have begun to encounter complex electrical behavior that we would never have realized was occurring just from the evidence of our eyes. Unlike the pulleys, blocks, and inclined planes of mechanics, the actors on the stage of electricity and magnetism are invisible phenomena alien to our everyday experience. For this reason, the flavor of the second half of your physics education is dramatically different, focusing much more on experiments and techniques. Even though you will never actually see charge moving through a wire, you can learn to use an ammeter to measure the flow.

Students also tend to get the impression from their first semester of physics that it is a dead science. Not so! We are about to pick up the historical trail that leads directly to the cutting-edge physics research you read about in the newspaper. The atom-smashing experiments that began around 1900, which we will be studying in this chapter, were not that different from the ones of the year 2000 --- just smaller, simpler, and much cheaper.

Magnetic forces

A detailed mathematical treatment of magnetism won't come until much later in this book, but we need to develop a few simple ideas about magnetism now because magnetic forces are used in the experiments and techniques we come to next. Everyday magnets come in two general types. Permanent magnets, such as the ones on your refrigerator, are made of iron or substances like steel that contain iron atoms. (Certain other substances also work, but iron is the cheapest and most common.) The other type of magnet, an example of which is the ones that make your stereo speakers vibrate, consist of coils of wire through which electric charge flows. Both types of magnets are able to attract iron that has not been magnetically prepared, for instance the door of the refrigerator.

A single insight makes these apparently complex phenomena much simpler to understand: magnetic forces are interactions between moving charges, occurring in addition to the electric forces. Suppose a permanent magnet is brought near a magnet of the coiled-wire type. The coiled wire has moving charges in it because we force charge to flow. The permanent magnet also has moving charges in it, but in this case the charges that naturally swirl around inside the iron. (What makes a magnetized piece of iron different from a block of wood is that the motion of the charge in the wood is random rather than organized.) The moving charges in the coiled-wire magnet exert a force on the moving charges in the permanent magnet, and vice-versa.

The mathematics of magnetism is significantly more complex than the Coulomb force law for electricity, which is why we will wait until chapter 11 before delving deeply into it. Two simple facts will suffice for now:

1. If a charged particle is moving in a region of space near where other charged particles are also moving, their magnetic force on it is directly proportional to its velocity.
2. The magnetic force on a moving charged particle is always perpendicular to the direction the particle is moving.

Example 1: A magnetic compass

The Earth is molten inside, and like a pot of boiling water, it roils and churns. To make a drastic oversimplification, electric charge can get carried along with the churning motion, so the Earth contains moving charge. The needle of a magnetic compass is itself a small permanent magnet. The moving charge inside the earth interacts magnetically with the moving charge inside the compass needle, causing the compass needle to twist around and point north.

Example 2: A television tube

A TV picture is painted by a stream of electrons coming from the back of the tube to the front. The beam scans across the whole surface of the tube like a reader scanning a page of a book. Magnetic forces are used to steer the beam. As the beam comes from the back of the tube to the front, up-down and left-right forces are needed for steering. But magnetic forces cannot be used to get the beam up to speed in the first place, since they can only push perpendicular to the electrons' direction of motion, not forward along it.

Discussion Questions

◊ An electrically charged piece of tape will be attracted to your hand. Does that allow us to tell whether the mobile charged particles in your hand are positive or negative, or both?

◊ If the electrical attraction between two pointlike objects at a distance of 1 m is 9×10^9 N, why can't we infer that their charges are $+1$ and -1 C? What further observations would we need to do in order to prove this?

$$\frac{m_{\mathrm{He}}}{m_{\mathrm{H}}} = 3.97$$

$$\frac{m_{\mathrm{Ne}}}{m_{\mathrm{H}}} = 20.01$$

$$\frac{m_{\mathrm{Sc}}}{m_{\mathrm{H}}} = 44.60$$

Examples of masses of atoms compared to that of hydrogen. Note how some, but not all, are close to integers.

8.1.3 Atoms

I was brought up to look at the atom as a nice, hard fellow, red or grey in color according to taste. -- *Rutherford*

Atomism

The Greeks have been kicked around a lot in the last couple of millennia: dominated by the Romans, bullied during the crusades by warlords going to and from the Holy Land, and occupied by Turkey until recently. It's no wonder they prefer to remember their salad days, when their best thinkers came up with concepts like democracy and atoms. Greece is democratic again after a period of military dictatorship, and an atom is proudly pictured on one of their coins. That's why it hurts me to have to say that the ancient Greek hypothesis that matter is made of atoms was pure guesswork. There was no real experimental evidence for atoms, and the 18th-century revival of the atom concept by Dalton owed little to the Greeks other than the name, which means “unsplittable.” Subtracting even more cruelly from Greek glory, the name was shown to be inappropriate in 1897 when physicist J.J. Thomson

proved experimentally that atoms had even smaller things inside them, which could be extracted. (Thomson called them “electrons.”) The “unsplittable” was splittable after all.

But that's getting ahead of our story. What happened to the atom concept in the intervening two thousand years? Educated people continued to discuss the idea, and those who were in favor of it could often use it to give plausible explanations for various facts and phenomena. One fact that was readily explained was conservation of mass. For example, if you mix 1 kg of water with 1 kg of dirt, you get exactly 2 kg of mud, no more and no less. The same is true for the a variety of processes such as freezing of water, fermenting beer, or pulverizing sandstone. If you believed in atoms, conservation of mass made perfect sense, because all these processes could be interpreted as mixing and rearranging atoms, without changing the total number of atoms. Still, this is nothing like a proof that atoms exist.

If atoms did exist, what types of atoms were there, and what distinguished the different types from each other? Was it their sizes, their shapes, their weights, or some other quality? The chasm between the ancient and modern atomisms becomes evident when we consider the wild speculations that existed on these issues until the present century. The ancients decided that there were four types of atoms, earth, water, air and fire; the most popular view was that they were distinguished by their shapes. Water atoms were spherical, hence water's ability to flow smoothly. Fire atoms had sharp points, which was why fire hurt when it touched one's skin. (There was no concept of temperature until thousands of years later.) The drastically different modern understanding of the structure of atoms was achieved in the course of the revolutionary decade stretching 1895 to 1905. The main purpose of this chapter is to describe those momentous experiments.

Atoms, light, and everything else

Although I tend to ridicule ancient Greek philosophers like Aristotle, let's take a moment to praise him for something. If you read Aristotle's writings on physics (or just skim them, which is all I've done), the most striking thing is how careful he is about classifying phenomena and analyzing relationships among phenomena. The human brain seems to naturally make a distinction between two types of physical phenomena: objects and motion of objects. When a phenomenon occurs that does not immediately present itself as one of these, there is a strong tendency to conceptualize it as one or the other, or even to ignore its existence completely. For instance, physics teachers shudder at students' statements that “the dynamite exploded, and force came out of it in all directions.” In these examples, the nonmaterial concept of force is being mentally categorized as if it was a physical substance. The statement that “winding the clock stores motion in the spring” is a miscategorization of electrical energy as a form of motion. An example of ignoring the existence of a phenomenon altogether can be elicited by asking people why we need lamps. The typical response that “the lamp illuminates the room so we can see things,” ignores the necessary role of light coming into our eyes from the things being illuminated.

If you ask someone to tell you briefly about atoms, the likely response is that “everything is made of atoms,” but we've now seen that it's far from obvious which “everything” this statement would properly refer to. For the scientists of the early 1900s who were trying to investigate atoms, this was not a trivial issue of definitions. There was a new gizmo called the vacuum tube, of which the only familiar example today is the picture tube of a TV. In short order, electrical tinkerers had discovered a whole flock of new phenomena that occurred in and around vacuum tubes, and given them picturesque names like “x-rays,” “cathode rays,” “Hertzian waves,” and “N-rays.” These were the types of observations that ended up telling us that we know about matter, but fierce controversies ensued over whether these were themselves forms of matter.

Let's bring ourselves up to the level of classification of phenomena employed by physicists in the year 1900. They recognized three categories:

- *Matter* has mass, can have kinetic energy, and can travel through a vacuum, transporting its mass and kinetic energy with it. Matter is conserved, both in the sense of conservation of mass and conservation of the number of atoms of each element. Atoms can't occupy the same space as other atoms, so a convenient way to prove something is not a form of matter is to show that it can pass through a solid material, in which the atoms are packed together closely.
- *Light* has no mass, always has energy, and can travel through a vacuum, transporting its energy with it. Two light beams can penetrate through each other and emerge from the collision without being weakened, deflected, or affected in any other way. Light can penetrate certain kinds of matter, e.g., glass.
- The third category is everything that doesn't fit the definition of light or matter. This catch-all category includes, for example, time, velocity, heat, and force.

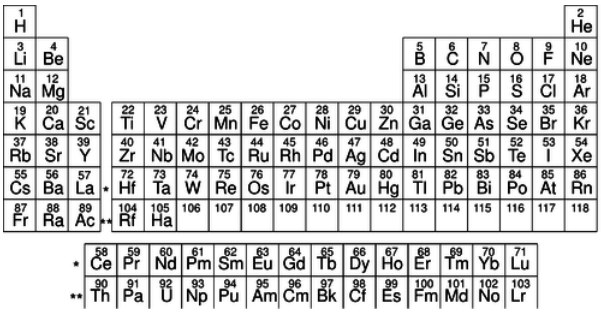
The chemical indexelements, chemicalelements

How would one find out what types of atoms there were? Today, it doesn't seem like it should have been very difficult to work out an experimental program to classify the types of atoms. For each type of atom, there should be a corresponding element, i.e., a pure substance made out of nothing but that type of atom. Atoms are supposed to be unsplittable, so a substance like milk could not possibly be elemental, since churning it vigorously causes it to split up into two separate substances: butter and whey. Similarly, rust could not be an element, because it can be made by combining two substances: iron and oxygen. Despite its apparent reasonableness, no such program was carried out until the eighteenth century. The ancients presumably did not do it because observation was not universally agreed on as the right way to answer questions about nature, and also because they lacked the necessary techniques or the techniques were the province of laborers with low social status, such as smiths and miners. Alchemists were hindered by atomism's reputation for subversiveness, and by a tendency toward mysticism and secrecy. (The most celebrated challenge facing the alchemists, that of converting lead into gold, is one we now know to be impossible, since lead and gold are both elements.)

By 1900, however, chemists had done a reasonably good job of finding out what the elements were. They also had determined the ratios of the different atoms' masses fairly accurately. A typical technique would be to measure how many grams of sodium (Na) would combine with one gram of chlorine (Cl) to make salt (NaCl). (This assumes you've already decided based on other evidence that salt consisted of equal numbers of Na and Cl atoms.) The masses of individual atoms, as opposed to the mass ratios, were known only to within a few orders of magnitude based on indirect evidence, and plenty of physicists and chemists denied that individual atoms were anything more than convenient symbols.

Making sense of the elements

As the information accumulated, the challenge was to find a way of systematizing it; the modern scientist's aesthetic sense rebels against complication. This hodgepodge of elements was an embarrassment. One contemporary observer, William Crookes, described the elements as extending “before us as stretched the wide Atlantic before the gaze of Columbus, mocking, taunting and murmuring strange riddles, which no man has yet been able to solve.” It wasn't long before people started recognizing that many atoms' masses were nearly integer multiples of the mass of hydrogen, the lightest element. A few excitable types began speculating that hydrogen was the basic building block, and that the heavier elements were made of clusters of hydrogen. It wasn't long, however, before their parade was rained on by more accurate measurements, which showed that not all of the elements had atomic masses that were near integer multiples of hydrogen, and even the ones that were close to being integer multiples were off by one percent or so.



e / A modern periodic table. Elements in the same column have similar chemical properties. The modern atomic numbers, discussed in section 8.2, were not known in Mendeleev's time, since the table could be flipped in various ways.

Chemistry professor Dmitri Mendeleev, preparing his lectures in 1869, wanted to find some way to organize his knowledge for his students to make it more understandable. He wrote the names of all the elements on cards and began arranging them in different ways on his desk, trying to find an arrangement that would make sense of the muddle. The row-and-column scheme he came up with is essentially our modern periodic table. The columns of the modern version represent groups of elements with similar chemical properties, and each row is more massive than the one above it. Going across each row, this almost always resulted in placing the atoms in sequence by weight as well. What made the system significant was its predictive value. There were three places where Mendeleev had to leave gaps in his checkerboard to keep chemically similar elements in the same column. He predicted that elements would exist to fill these gaps, and extrapolated or interpolated from other elements in the same column to predict their numerical properties, such as masses, boiling points, and densities. Mendeleev's professional stock skyrocketed when

his three elements (later named gallium, scandium and germanium) were discovered and found to have very nearly the properties he had predicted.

One thing that Mendeleev's table made clear was that mass was not the basic property that distinguished atoms of different elements. To make his table work, he had to deviate from ordering the elements strictly by mass. For instance, iodine atoms are lighter than tellurium, but Mendeleev had to put iodine after tellurium so that it would lie in a column with chemically similar elements.

Direct proof that atoms existed

The success of the kinetic theory of heat was taken as strong evidence that, in addition to the motion of any object as a whole, there is an invisible type of motion all around us: the random motion of atoms within each object. But many conservatives were not convinced that atoms really existed. Nobody had ever seen one, after all. It wasn't until generations after the kinetic theory of heat was developed that it was demonstrated conclusively that atoms really existed and that they participated in continuous motion that never died out.

The smoking gun to prove atoms were more than mathematical abstractions came when some old, obscure observations were reexamined by an unknown Swiss patent clerk named Albert Einstein. A botanist named Brown, using a microscope that was state of the art in 1827, observed tiny grains of pollen in a drop of water on a microscope slide, and found that they jumped around randomly for no apparent reason. Wondering at first if the pollen he'd assumed to be dead was actually alive, he tried looking at particles of soot, and found that the soot particles also moved around. The same results would occur with any small grain or particle suspended in a liquid. The phenomenon came to be referred to as Brownian motion, and its existence was filed away as a quaint and thoroughly unimportant fact, really just a nuisance for the microscopist.

It wasn't until 1906 that Einstein found the correct interpretation for Brown's observation: the water molecules were in continuous random motion, and were colliding with the particle all the time, kicking it in random directions. After all the millennia of speculation about atoms, at last there was solid proof. Einstein's calculations dispelled all doubt, since he was able to make accurate predictions of things like the average distance traveled by the particle in a certain amount of time. (Einstein received the Nobel Prize not for his theory of relativity but for his papers on Brownian motion and the photoelectric effect.)

Discussion Questions

◊ How could knowledge of the size of an individual aluminum atom be used to infer an estimate of its mass, or vice versa?

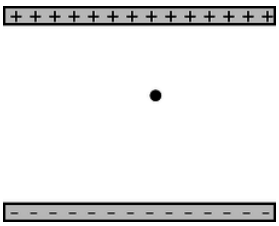
◊ How could one test Einstein's interpretation of Brownian motion by observing it at different temperatures?

8.1.4 Quantization of charge

Proving that atoms actually existed was a big accomplishment, but demonstrating their existence was different from understanding their properties. Note that the Brown-Einstein observations had nothing at all to do with electricity, and yet we know that matter is inherently electrical, and we have been successful in interpreting certain electrical phenomena in terms of mobile positively and negatively charged particles. Are these particles atoms? Parts of atoms? Particles that are entirely separate from atoms? It is perhaps premature to attempt to answer these questions without any conclusive evidence in favor of the charged-particle model of electricity.

f / A young Robert Millikan. (Contemporary)

Strong support for the charged-particle model came from a 1911 experiment by physicist Robert Millikan at the University of Chicago. Consider a jet of droplets of perfume or some other liquid made by blowing it through a tiny pinhole. The droplets emerging from the pinhole must be smaller than the pinhole, and in fact most of them are even more microscopic than that, since the turbulent flow of air tends to break them up. Millikan reasoned that the droplets would acquire a little bit of electric charge as they rubbed against the channel through which they emerged, and if the charged-particle model of electricity was right, the charge might be split up among so many minuscule liquid drops that a single drop might have a total charge amounting to an excess of only a few charged particles --- perhaps an excess of one positive particle on a certain drop, or an excess of two negative ones on another.



g / A simplified diagram of Millikan's apparatus.

Millikan's ingenious apparatus, g, consisted of two metal plates, which could be electrically charged as needed. He sprayed a cloud of oil droplets into the space between the plates, and selected one drop through a microscope for study. First, with no charge on the plates, he would determine the drop's mass by letting it fall through the air and measuring its terminal velocity, i.e., the velocity at which the force of air friction canceled out the force of gravity. The force of air drag on a slowly moving sphere had already been found by experiment to be bvr^2, where b was a constant. Setting the total force equal to zero when the drop is at terminal velocity gives

$$bvr^2 - mg = 0,$$

and setting the known density of oil equal to the drop's mass divided by its volume gives a second equation,

$$\rho = \frac{m}{\frac{4}{3}\pi r^3}.$$

Everything in these equations can be measured directly except for m and r, so these are two equations in two unknowns, which can be solved in order to determine how big the drop is.

Next Millikan charged the metal plates, adjusting the amount of charge so as to exactly counteract gravity and levitate the drop. If, for instance, the drop being examined happened to have a total charge that was negative, then positive charge put on the top plate would attract it, pulling it up, and negative charge on the bottom plate would repel it, pushing it up. (Theoretically only one plate would be necessary, but in practice a two-plate arrangement like this gave electrical forces that were more uniform in strength throughout the space where the oil drops were.) The amount of charge on the plates required to levitate the charged drop gave Millikan a handle on the amount of charge the drop carried. The more charge the drop had, the stronger the electrical forces on it would be, and the less charge would have to be put on the plates to do the trick. Unfortunately, expressing this relationship using Coulomb's law would have been impractical, because it would require a perfect knowledge of how the charge was distributed on each plate, plus the ability to perform vector addition of all the forces being exerted on the drop by all the charges on the plate. Instead, Millikan made use of the fact that the electrical force experienced by a pointlike charged object at a certain point in space is proportional to its charge,

$$\frac{F}{q} = \text{constant}.$$

With a given amount of charge on the plates, this constant could be determined for instance by discarding the oil drop, inserting between the plates a larger and more easily handled object with a known charge on it, and measuring the force with conventional methods. (Millikan actually used a slightly different set of techniques for determining the constant, but the concept is the same.) The amount of force on the actual oil drop had to equal mg, since it was just enough to levitate it, and once the calibration constant had been determined, the charge of the drop could then be found based on its previously determined mass.

q (C)	*q / (1.64 X 10^{-19} C)*

-1.970×10^{-18}	−12.02
-0.987×10^{-18}	−6.02
-2.773×10^{-18}	−16.93

A few samples of Millikan's data.

The table above shows a few of the results from Millikan's 1911 paper. (Millikan took data on both negatively and positively charged drops, but in his paper he gave only a sample of his data on negatively charged drops, so these numbers are all negative.) Even a quick look at the data leads to the suspicion that the charges are not simply a series of random numbers. For instance, the second charge is almost exactly equal to half the first one. Millikan explained the observed charges as all being integer multiples of a single number, 1.64×10^{-19} C. In the second column, dividing by this constant gives numbers that are essentially integers, allowing for the random errors present in the experiment. Millikan states in his paper that these results were a

... direct and tangible demonstration ... of the correctness of the view advanced many years ago and supported by evidence from many sources that all electrical charges, however produced, are exact multiples of one definite, elementary electrical charge, or in other words, that an electrical charge instead of being spread uniformly over the charged surface has a definite granular structure, consisting, in fact, of ... specks, or atoms of electricity, all precisely alike, peppered over the surface of the charged body.

In other words, he had provided direct evidence for the charged-particle model of electricity and against models in which electricity was described as some sort of fluid. The basic charge is notated e, and the modern value is $e=1.60\times10^{-19}$ C. The word "*quantized*" is used in physics to describe a quantity that can only have certain numerical values, and cannot have any of the values between those. In this language, we would say that Millikan discovered that charge is quantized. The charge e is referred to as the quantum of charge.

A historical note on Millikan's fraud

Very few undergraduate physics textbooks mention the well-documented fact that although Millikan's conclusions were correct, he was guilty of scientific fraud. His technique was difficult and painstaking to perform, and his original notebooks, which have been preserved, show that the data were far less perfect than he claimed in his published scientific papers. In his publications, he stated categorically that every single oil drop observed had had a charge that was a multiple of e, with no exceptions or omissions. But his notebooks are replete with notations such as "beautiful data, keep," and "bad run, throw out." Millikan, then, appears to have earned his Nobel Prize by advocating a correct position with dishonest descriptions of his data.

Why do textbook authors fail to mention Millikan's fraud? It may be that they think students are too unsophisticated to correctly evaluate the implications of the fact that scientific fraud has sometimes existed and even been rewarded by the scientific establishment. Maybe they are afraid students will reason that fudging data is OK, since Millikan got the Nobel Prize for it. But falsifying history in the name of encouraging truthfulness is more than a little ironic. English teachers don't edit Shakespeare's tragedies so that the bad characters are always punished and the good ones never suffer!

self-check:

Is money quantized? What is the quantum of money?

(answer in the back of the PDF version of the book)

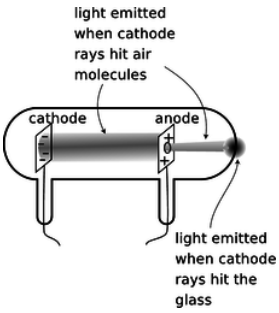
8.1.5 The electron

Cathode Rays

Nineteenth-century physicists spent a lot of time trying to come up with wild, random ways to play with electricity. The best experiments of this kind were the ones that made big sparks or pretty colors of light.

One such parlor trick was the cathode ray. To produce it, you first had to hire a good glassblower and find a good vacuum pump. The glassblower would create a hollow tube and embed two pieces of metal in it, called the electrodes, which were connected to the outside via metal wires passing through the glass. Before letting him seal up the whole tube, you would hook it up to a vacuum pump, and spend several hours huffing and puffing away at the pump's hand crank to get a good vacuum inside. Then, while you were still pumping on the tube, the glassblower would melt the glass and seal the whole thing shut. Finally, you would put a large amount of positive charge on one wire and a large amount of negative charge on the other. Metals have the property of letting

charge move through them easily, so the charge deposited on one of the wires would quickly spread out because of the repulsion of each part of it for every other part. This spreading-out process would result in nearly all the charge ending up in the electrodes, where there is more room to spread out than there is in the wire. For obscure historical reasons a negative electrode is called a cathode and a positive one is an anode.



i / Cathode rays observed in a vacuum tube.

Figure i shows the light-emitting stream that was observed. If, as shown in this figure, a hole was made in the anode, the beam would extend on through the hole until it hit the glass. Drilling a hole in the cathode, however would not result in any beam coming out on the left side, and this indicated that the stuff, whatever it was, was coming from the cathode. The rays were therefore christened “cathode rays.” (The terminology is still used today in the term “cathode ray tube” or “CRT” for the picture tube of a TV or computer monitor.)

Were cathode rays a form of light, or of matter?

Were cathode rays a form of light, or matter? At first no one really cared what they were, but as their scientific importance became more apparent, the light-versus-matter issue turned into a controversy along nationalistic lines, with the Germans advocating light and the English holding out for matter. The supporters of the material interpretation imagined the rays as consisting of a stream of atoms ripped from the substance of the cathode.

One of our defining characteristics of matter is that material objects cannot pass through each other. Experiments showed that cathode rays could penetrate at least some small thickness of matter, such as a metal foil a tenth of a millimeter thick, implying that they were a form of light.

Other experiments, however, pointed to the contrary conclusion. Light is a wave phenomenon, and one distinguishing property of waves is demonstrated by speaking into one end of a paper towel roll. The sound waves do not emerge from the other end of the tube as a focused beam. Instead, they begin spreading out in all directions as soon as they emerge. This shows that waves do not necessarily travel in straight lines. If a piece of metal foil in the shape of a star or a cross was placed in the way of the cathode ray, then a “shadow” of the same shape would appear on the glass, showing that the rays traveled in straight lines. This straight-line motion suggested that they were a stream of small particles of matter.

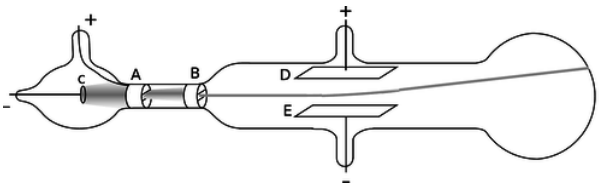
These observations were inconclusive, so what was really needed was a determination of whether the rays had mass and weight. The trouble was that cathode rays could not simply be collected in a cup and put on a scale. When the cathode ray tube is in operation, one does not observe any loss of material from the cathode, or any crust being deposited on the anode.

Nobody could think of a good way to weigh cathode rays, so the next most obvious way of settling the light/matter debate was to check whether the cathode rays possessed electrical charge. Light was known to be uncharged. If the cathode rays carried charge, they were definitely matter and not light, and they were presumably being made to jump the gap by the simultaneous repulsion of the negative charge in the cathode and attraction of the positive charge in the anode. The rays would overshoot the anode because of their momentum. (Although electrically charged particles do not normally leap across a gap of vacuum, very large amounts of charge were being used, so the forces were unusually intense.)

Thomson's experiments

j / J.J. Thomson in the lab.

Physicist J.J. Thomson at Cambridge carried out a series of definitive experiments on cathode rays around the year 1897. By turning them slightly off course with electrical forces, k, he showed that they were indeed electrically charged, which was strong evidence that they were material. Not only that, but he proved that they had mass, and measured the ratio of their mass to their charge, m/q. Since their mass was not zero, he concluded that they were a form of matter, and presumably made up of a stream of microscopic, negatively charged particles. When Millikan published his results fourteen years later, it was reasonable to assume that the charge of one such particle equaled minus one fundamental charge, $q = -e$, and from the combination of Thomson's and Millikan's results one could therefore determine the mass of a single cathode ray particle.



k / Thomson's experiment proving cathode rays had electric charge (redrawn from his original paper). The cathode, C, and anode, A, are as in any cathode ray tube. The rays pass through a slit in the anode, and a second slit, B, is interposed in order to make the beam thinner and eliminate rays that were not going straight. Charging plates D and E shows that cathode rays have charge: they are attracted toward the positive plate D and repelled by the negative plate E.

The basic technique for determining m/q was simply to measure the angle through which the charged plates bent the beam. The electric force acting on a cathode ray particle while it was between the plates would be proportional to its charge,

$$F_{elec} = (\text{known constant}) \cdot q.$$

Application of Newton's second law, $a = F/m$, would allow m/q to be determined:

$$\frac{m}{q} = \frac{\text{known constant}}{a}$$

There was just one catch. Thomson needed to know the cathode ray particles' velocity in order to figure out their acceleration. At that point, however, nobody had even an educated guess as to the speed of the cathode rays produced in a given vacuum tube. The beam appeared to leap across the vacuum tube practically instantaneously, so it was no simple matter of timing it with a stopwatch!

Thomson's clever solution was to observe the effect of both electric and magnetic forces on the beam. The magnetic force exerted by a particular magnet would depend on both the cathode ray's charge and its velocity:

$$F_{mag} = (\text{known constant } \#2) \cdot qv$$

Thomson played with the electric and magnetic forces until either one would produce an equal effect on the beam, allowing him to solve for the velocity,

$$v = \frac{(\text{known constant})}{(\text{known constant } \#2)}.$$

Knowing the velocity (which was on the order of 10% of the speed of light for his setup), he was able to find the acceleration and thus the mass-to-charge ratio m/q. Thomson's techniques were relatively crude (or perhaps more charitably we could say that they stretched the state of the art of the time), so with various methods he came up with m/q values that ranged over about a factor of

two, even for cathode rays extracted from a cathode made of a single material. The best modern value is $m/q = 5.69 \times 10^{-12}$ kg/C, which is consistent with the low end of Thomson's range.

The cathode ray as a subatomic particle: the indexelectronelectron

What was significant about Thomson's experiment was not the actual numerical value of m/q, however, so much as the fact that, combined with Millikan's value of the fundamental charge, it gave a mass for the cathode ray particles that was thousands of times smaller than the mass of even the lightest atoms. Even without Millikan's results, which were 14 years in the future, Thomson recognized that the cathode rays' m/q was thousands of times smaller than the m/q ratios that had been measured for electrically charged atoms in chemical solutions. He correctly interpreted this as evidence that the cathode rays were smaller building blocks --- he called them *electrons* --- out of which atoms themselves were formed. This was an extremely radical claim, coming at a time when atoms had not yet been proven to exist! Even those who used the word “atom” often considered them no more than mathematical abstractions, not literal objects. The idea of searching for structure inside of “unsplittable” atoms was seen by some as lunacy, but within ten years Thomson's ideas had been amply verified by many more detailed experiments.

Discussion Questions

◊ Thomson started to become convinced during his experiments that the “cathode rays” observed coming from the cathodes of vacuum tubes were building blocks of atoms --- what we now call electrons. He then carried out observations with cathodes made of a variety of metals, and found that m/q was roughly the same in every case, considering his limited accuracy. Given his suspicion, why did it make sense to try different metals? How would the consistent values of m/q serve to test his hypothesis?

◊ My students have frequently asked whether the m/q that Thomson measured was the value for a single electron, or for the whole beam. Can you answer this question?

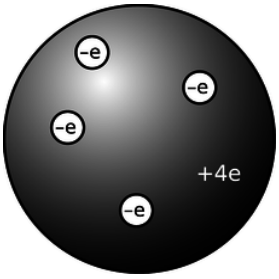
◊ Thomson found that the m/q of an electron was thousands of times smaller than that of charged atoms in chemical solutions. Would this imply that the electrons had more charge? Less mass? Would there be no way to tell? Explain. Remember that Millikan's results were still many years in the future, so q was unknown.

◊ Can you guess any practical reason why Thomson couldn't just let one electron fly across the gap before disconnecting the battery and turning off the beam, and then measure the amount of charge deposited on the anode, thus allowing him to measure the charge of a single electron directly?

◊ Why is it not possible to determine m and q themselves, rather than just their ratio, by observing electrons' motion in electric and magnetic fields?

8.1.6 The raisin cookie model of the atom

Based on his experiments, Thomson proposed a picture of the atom which became known as the raisin cookie model. In the neutral atom, l, there are four electrons with a total charge of $-4e$, sitting in a sphere (the “cookie”) with a charge of $+4e$ spread throughout it. It was known that chemical reactions could not change one element into another, so in Thomson's scenario, each element's cookie sphere had a permanently fixed radius, mass, and positive charge, different from those of other elements. The electrons, however, were not a permanent feature of the atom, and could be tacked on or pulled out to make charged ions. Although we now know, for instance, that a neutral atom with four electrons is the element beryllium, scientists at the time did not know how many electrons the various neutral atoms possessed.



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

This model is clearly different from the one you've learned in grade school or through popular culture, where the positive charge is concentrated in a tiny nucleus at the atom's center. An equally important change in ideas about the atom has been the realization that atoms and their constituent subatomic particles behave entirely differently from objects on the human scale. For instance, we'll

see later that an electron can be in more than one place at one time. The raisin cookie model was part of a long tradition of attempts to make mechanical models of phenomena, and Thomson and his contemporaries never questioned the appropriateness of building a mental model of an atom as a machine with little parts inside. Today, mechanical models of atoms are still used (for instance the tinker-toy-style molecular modeling kits like the ones used by Watson and Crick to figure out the double helix structure of DNA), but scientists realize that the physical objects are only aids to help our brains' symbolic and visual processes think about atoms.

Although there was no clear-cut experimental evidence for many of the details of the raisin cookie model, physicists went ahead and started working out its implications. For instance, suppose you had a four-electron atom. All four electrons would be repelling each other, but they would also all be attracted toward the center of the “cookie” sphere. The result should be some kind of stable, symmetric arrangement in which all the forces canceled out. People sufficiently clever with math soon showed that the electrons in a four-electron atom should settle down at the vertices of a pyramid with one less side than the Egyptian kind, i.e., a regular tetrahedron. This deduction turns out to be wrong because it was based on incorrect features of the model, but the model also had many successes, a few of which we will now discuss.

Example 3: Flow of electrical charge in wires

One of my former students was the son of an electrician, and had become an electrician himself. He related to me how his father had remained refused to believe all his life that electrons really flowed through wires. If they had, he reasoned, the metal would have gradually become more and more damaged, eventually crumbling to dust.
His opinion is not at all unreasonable based on the fact that electrons are material particles, and that matter cannot normally pass through matter without making a hole through it. Nineteenth-century physicists would have shared his objection to a charged-particle model of the flow of electrical charge. In the raisin-cookie model, however, the electrons are very low in mass, and therefore presumably very small in size as well. It is not surprising that they can slip between the atoms without damaging them.

Example 4: Flow of electrical charge across cell membranes

Your nervous system is based on signals carried by charge moving from nerve cell to nerve cell. Your body is essentially all liquid, and atoms in a liquid are mobile. This means that, unlike the case of charge flowing in a solid wire, entire charged atoms can flow in your nervous system

Example 5: Emission of electrons in a cathode ray tube

Why do electrons detach themselves from the cathode of a vacuum tube? Certainly they are encouraged to do so by the repulsion of the negative charge placed on the cathode and the attraction from the net positive charge of the anode, but these are not strong enough to rip electrons out of atoms by main force --- if they were, then the entire apparatus would have been instantly vaporized as every atom was simultaneously ripped apart!
The raisin cookie model leads to a simple explanation. We know that heat is the energy of random motion of atoms. The atoms in any object are therefore violently jostling each other all the time, and a few of these collisions are violent enough to knock electrons out of atoms. If this occurs near the surface of a solid object, the electron may can come loose. Ordinarily, however, this loss of electrons is a self-limiting process; the loss of electrons leaves the object with a net positive charge, which attracts the lost sheep home to the fold. (For objects immersed in air rather than vacuum, there will also be a balanced exchange of electrons between the air and the object.)
This interpretation explains the warm and friendly yellow glow of the vacuum tubes in an antique radio. To encourage the emission of electrons from the vacuum tubes' cathodes, the cathodes are intentionally warmed up with little heater coils.

Discussion Questions

◊ Today many people would define an ion as an atom (or molecule) with missing electrons or extra electrons added on. How would people have defined the word “ion” before the discovery of the electron?

◊ Since electrically neutral atoms were known to exist, there had to be positively charged subatomic stuff to cancel out the negatively charged electrons in an atom. Based on the state of knowledge immediately after the Millikan and Thomson experiments, was it possible that the positively charged stuff had an unquantized amount of charge? Could it be quantized in units of +e? In units of +2e? In units of +5/7e?

Contributor

9.2: The Nucleus

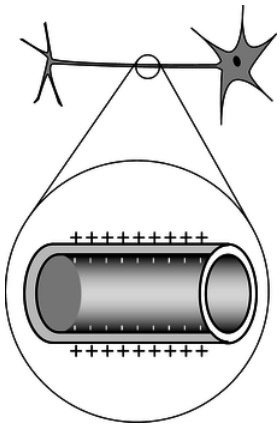
Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

9.3: Footnotes

1. Alpha decay is more common because an alpha particle happens to be a very stable arrangement of protons and neutrons.
2. The evidence for the big bang theory of the origin of the universe was discussed on p. 356.
3. For two opposing viewpoints, see Tubiana et al., “The Linear No-Threshold Relationship Is Inconsistent with Radiation Biologic and Experimental Data,” Radiology, 251 (2009) 13 and Little et al., “ Risks Associated with Low Doses and Low Dose Rates of Ionizing Radiation: Why Linearity May Be (Almost) the Best We Can Do,” Radiology, 251 (2009) 6.
4. Baker and Chesser, Env. Toxicology and Chem. 19 (1231) 2000. Similar effects have been seen at the Bikini Atoll, the site of a 1954 hydrogen bomb test. Although some species have disappeared from the area, the coral reef is in many ways healthier than similar reefs elsewhere, because humans have tended to stay away for fear of radiation (Richards et al., Marine Pollution Bulletin 56 (2008) 503).
5. The evidence for the big bang theory of the origin of the universe was discussed in book 3 of this series.

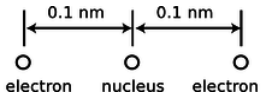
9.4: Problems



a / Problem 1. Top: A realistic picture of a neuron. Bottom: A simplified diagram of one segment of the tail (axon).

1. 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 does not the positive charge wander away from the outside surface?

2. The Earth and Moon are bound together by gravity. If, instead, the force of attraction were the result of each having a charge of the same magnitude but opposite in sign, find the quantity of charge that would have to be placed on each to produce the required force.(answer check available at lightandmatter.com)



b / Problem 3.

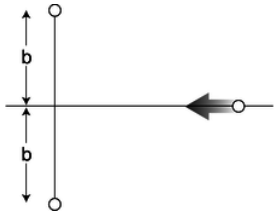
3. A helium atom finds itself momentarily in this arrangement. Find the direction and magnitude of the force acting on the right-hand electron. The two protons in the nucleus are so close together (~ 1 fm) that you can consider them as being right on top of each other.

(answer check available at lightandmatter.com)

4. $^{234}\mathrm{Pu}$ decays either by electron decay or by alpha decay. (A given $^{234}\mathrm{Pu}$ nucleus may do either one; it's random.) What are the isotopes created as products of these two modes of decay?

5. Suppose that a proton in a lead nucleus wanders out to the surface of the nucleus, and experiences a strong nuclear force of about 8 kN from the nearby neutrons and protons pulling it back in. Compare this numerically to the repulsive electrical force from the other protons, and verify that the net force is attractive. A lead nucleus is very nearly spherical, is about 6.5 fm in radius, and contains 82 protons, each with a charge of $+e$, where $e = 1.60 \times 10^{-19}$ C . (answer check available at lightandmatter.com)

6. The nuclear process of beta decay by electron capture is described parenthetically on page 493. The reaction is $\mathrm{p} + \mathrm{e}^{-} \rightarrow \mathrm{n} + \nu$.
(a) Show that charge is conserved in this reaction.
(b) Conversion between energy and mass is discussed in chapter 7. Based on these ideas, explain why electron capture does not occur in hydrogen atoms. (If it did, matter wouldn't exist!)

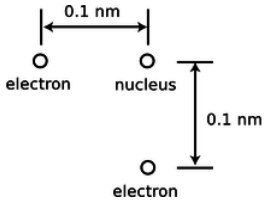


c / Problem 7.

7. 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 charges are fixed on the y axis at the locations shown in the figure. The two fixed charges are equal to each other, but not to the charge of the ball.
(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 you answer in terms of b. (Assume the ball has enough energy to continue through, so you don't have to worry about whether it actually gets as far as this particular value of x.) (Based on a problem by Halliday and Resnick.)

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.(answer check available at lightandmatter.com)

9. Use the nutritional information on some packaged food to make an order-of-magnitude estimate of the amount of chemical energy stored in one atom of food, in units of joules. Assume that a typical atom has a mass of 10^{-26} kg. This constitutes a rough estimate of the amounts of energy there are on the atomic scale. [See chapter 0 for help on how to do order-of-magnitude estimates. Note that a nutritional “calorie” is really a kilocalorie; see page 943.](answer check available at lightandmatter.com)



d / Problem 10.

10. The helium atom of problem 3 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.)(answer check available at lightandmatter.com)

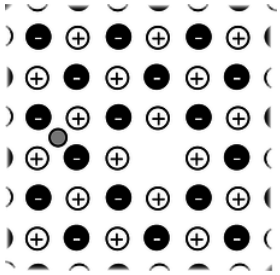
11. A neon light consists of a long glass tube full of neon, with metal caps on the ends. Positive charge is placed on one end of the tube, and negative charge on the other. The electric forces generated can be strong enough to strip electrons off of a certain number of neon atoms. Assume for simplicity that only one electron is ever stripped off of any neon atom. When an electron is stripped off of an atom, both the electron and the neon atom (now an ion) have electric charge, and they are accelerated by the forces exerted by the charged ends of the tube. (They do not feel any significant forces from the other ions and electrons within the tube, because only a tiny minority of neon atoms ever gets ionized.) Light is finally produced when ions are reunited with electrons. Give a numerical comparison of the magnitudes and directions of the accelerations of the electrons and ions. [You may need some data from appendix 5.] (answer check available at lightandmatter.com)

12. 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?(answer check available at lightandmatter.com)

13. (a) Recall that the gravitational energy of two gravitationally interacting spheres is given by $U_g = -Gm_1m_2/r$, where r is the center-to-center distance. What would be the analogous equation for two electrically interacting spheres? Justify your choice of a plus or minus sign on physical grounds, considering attraction and repulsion.(answer check available at lightandmatter.com)
(b) Use this expression to estimate the energy required to pull apart a raisin-cookie atom of the one-electron type, assuming a radius

of 10^{-10} m.(answer check available at lightandmatter.com)
(c) Compare this with the result of problem 9.

14. If you put two hydrogen atoms near each other, they will feel an attractive force, and they will pull together to form a molecule. (Molecules consisting of two hydrogen atoms are the normal form of hydrogen gas.) Why do they feel a force if they are near each other, since each is electrically neutral? Shouldn't the attractive and repulsive forces all cancel out exactly? Use the raisin cookie model. (Students who have taken chemistry often try to use fancier models to explain this, but if you cannot explain it using a simple model, you probably don't understand the fancy model as well as you thought you did!)



e / Problem 15.

15. 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 $-e$. The smaller circles are sodium ions, with charges of $+e$. The 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. \hwhint{hwhint:nacl}(answer check available at lightandmatter.com)

16. [See section 0.2 for help on how to do order-of-magnitude estimates.] Suppose you are holding your hands in front of you, 10 cm apart.
(a) Estimate the total number of electrons in each hand.
(b) Estimate the total repulsive force of all the electrons in one hand on all the electrons in the other.
(c) Why don't you feel your hands repelling each other?
(d) Estimate how much the charge of a proton could differ in magnitude from the charge of an electron without creating a noticeable force between your hands.

Contributors and Attributions

CHAPTER OVERVIEW

10: Circuits

Madam, what good is a baby? -- *Michael Faraday, when asked by Queen Victoria what the electrical devices in his lab were good for*

A few years ago, my wife and I bought a house with Character, Character being a survival mechanism that houses have evolved in order to convince humans to agree to much larger mortgage payments than they'd originally envisioned. Anyway, one of the features that gives our house Character is that it possesses, built into the wall of the family room, a set of three pachinko machines. These are Japanese gambling devices sort of like vertical pinball machines. (The legal papers we got from the sellers hastened to tell us that they were “for amusement purposes only.”) Unfortunately, only one of the three machines was working when we moved in, and it soon died on us. Having become a pachinko addict, I decided to fix it, but that was easier said than done. The inside is a veritable Rube Goldberg mechanism of levers, hooks, springs, and chutes. My hormonal pride, combined with my Ph.D. in physics, made me certain of success, and rendered my eventual utter failure all the more demoralizing.

Contemplating my defeat, I realized how few complex mechanical devices I used from day to day. Apart from our cars and my saxophone, every technological tool in our modern life-support system was electronic rather than mechanical.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

10.1: Current and Voltage

Contributors and Attributions

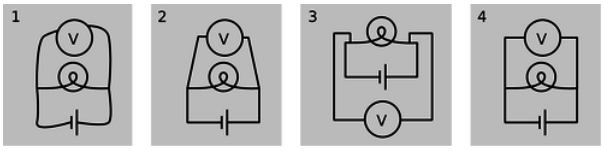
Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

10.2: Parallel and Series Circuits

In section 9.1, we limited ourselves to relatively simple circuits, essentially nothing more than a battery and a single lightbulb. The purpose of this chapter is to introduce you to more complex circuits, containing multiple resistors or voltage sources in series, in parallel, or both.

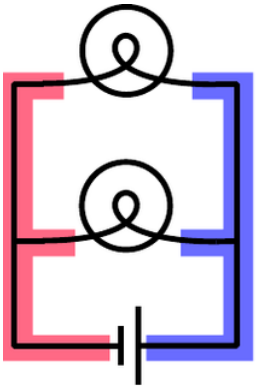
9.2.1 Schematics

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 voltage 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.



b / The two shaded areas shaped like the letter “E” are both regions of constant voltage.

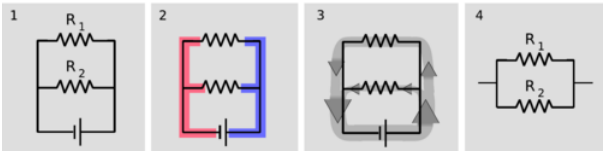
The most important first step in learning to read schematics is to learn to recognize contiguous pieces of wire which must have constant voltage throughout. In figure b, for example, the two shaded E-shaped pieces of wire must each have constant voltage. This focuses our attention on two of the main unknowns we'd like to be able to predict: the voltage of the left-hand E and the voltage of the one on the right.

9.2.2 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-voltage areas in the circuit, c/2, all three components have the same voltage difference across them. A battery normally succeeds in maintaining the voltage 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 voltage 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-voltage 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:

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.

Example 10: Two lamps on the same household circuit

▷ 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

$$\begin{aligned} R &= \left(\frac{1}{R_1} + \frac{1}{R_2}\right)^{-1} \\ &= \left(\frac{1}{1\,\Omega} + \frac{1}{1\,\Omega}\right)^{-1} \\ &= \left(1\,\Omega^{-1} + 1\,\Omega^{-1}\right)^{-1} \\ &= \left(2\,\Omega^{-1}\right)^{-1} \\ &= 0.5\,\Omega \end{aligned}$$

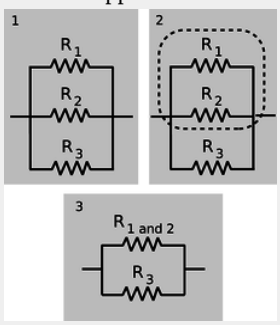
The voltage difference across the whole circuit is always the 110 V set by the electric company (it's alternating current, but that's irrelevant). The resistance of the whole circuit has been cut in half by turning on the second lamp, so a fixed amount of voltage will produce twice as much current. Twice the current flowing across the same voltage 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.

Example 11: Three resistors in parallel

▷ What happens if we have three or more resistors in parallel?



d / Three 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 them simplifying those parts. In the circuit d/1, with three resistors in parallel, we can think of two of the resistors as forming a single big resistor, d/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 d/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 12: An arbitrary number of identical resistors in parallel

▷ What is the resistance of N identical resistors in parallel?

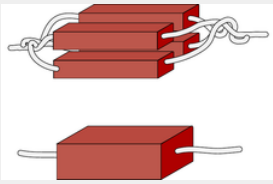
▷ Generalizing the results for two and three resistors, we have

$$R_N = \left(\frac{1}{R_1} + \frac{1}{R_2} + \ldots\right)^{-1},$$

where "..." means that the sum includes all the resistors. If all the resistors are identical, this becomes

$$\begin{aligned} R_N &= \left(\frac{N}{R}\right)^{-1} \\ &= \frac{R}{N} \end{aligned}$$

Example 13: Dependence of resistance on cross-sectional area



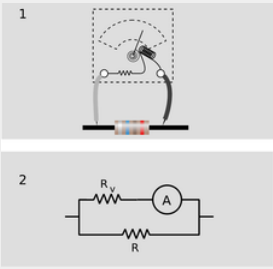
e / Uniting four resistors in parallel is equivalent to making a single resistor with the same length but four times the cross-sectional area. The result is to make a resistor with one quarter the resistance.

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 e, increasing a resistors'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 μm^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.

f / 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.

Example 14: Incorrect readings from a voltmeter



g / 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.

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 voltage 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 what how current flows in the circuit, since it is essentially just some coiled-up wire with a very low resistance.

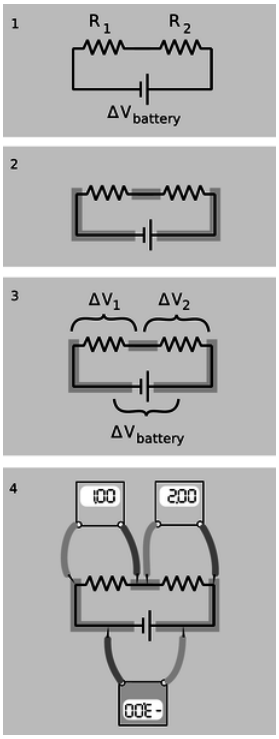
Now if we are carrying out this measurement on a resistor that is part of a larger circuit, we have changed the behavior of the circuit through our act of measuring. It is as though we had modified the circuit by replacing the resistance R with the smaller equivalent resistance of R and R_v in parallel. 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 $M\Omega$ to measure the voltage drop across a one-ohm resistor, the equivalent resistance is 0.999999 Ω, 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 $2-M\Omega$ 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 voltage 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-

voltage 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 voltage 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.

9.2.3 Series resistances



h / 1. A battery drives current through two resistors in series. 2. There are three constant-voltage 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.

The two basic circuit layouts are parallel and series, so a pair of resistors in series, h/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 h/2 shows the three constant-voltage areas. Voltage differences are more physically significant than voltages, so we define symbols for the voltage differences across the two resistors in figure h/3.

We have three constant-voltage areas, with symbols for the difference in voltage between every possible pair of them. These three voltage 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 voltage 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 h/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:

Assuming the standard convention for plus and minus signs, the sum of the voltage drops around any closed loop in a circuit must be zero.

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!

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:

$$I_1 R_1 + I_2 R_2 + \Delta V_{battery} = 0.$$

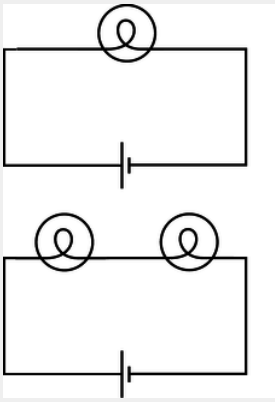
The currents are the same, so we can factor them out:

$$I\,(R_1 + R_2) + \Delta V_{battery} = 0,$$

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.

Example 15: Two lightbulbs in series

▷ If two identical lightbulbs are placed in series, how do their brightnesses compare with the brightness of a single bulb?



i / Example 15.

▷ 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.

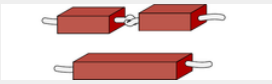
Roughly speaking, we might expect this to result in one quarter the light being produced by each bulb, but in reality lightbulbs waste quite a high percentage of their power in the form of heat and wavelengths of light that are not visible (infrared and ultraviolet). Less light will be produced, but it's hard to predict exactly how much less, since the efficiency of the bulbs will be changed by operating them under different conditions.

Example 16: More than two equal resistances in series

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.

Example 17: Dependence of resistance on length

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.



j / Doubling the length of a resistor is like putting two resistors in series. The resistance is doubled.

Putting the two arguments together, 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.

Example 18: Choice of high voltage for power lines

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.)

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

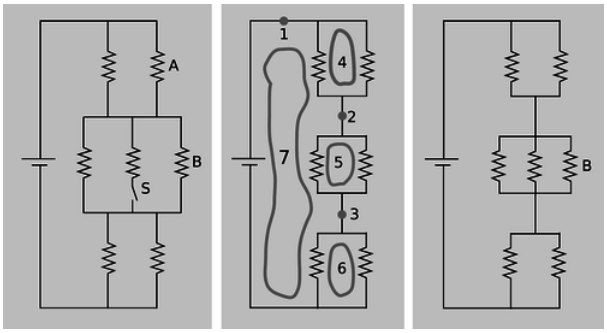
$$efficiency = \frac{1}{1 + \frac{I^2 R_T}{P_L}}$$

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.

Example 19: Getting killed by your ammeter

As with a voltmeter, an ammeter can give erroneous readings if it is used in such a way that it changes the behavior 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.



k / Example 20.

Example 20: A complicated circuit

▷ All seven resistors in the left-hand panel of figure k are identical. Initially, the switch S is open as shown in the figure, and the current through resistor A is I_o. The switch is then closed. Find the current through resistor B, after the switch is closed, in terms of I_o.

▷ 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 $I_o = V/3R$.

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 $(1/3)(3V/4R) = V/4R = 3I_o/4$.

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 the first, and therefore the current through resistor B is lessened over all.

Discussion Question

◊ 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?

Contributors

10.E: Circuits (Exercises)

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. (solution in the pdf version of the book)

a / Problem 2.

2. Referring back to our old friend the neuron from problem 1 on page 504, let's now consider what happens when the nerve is stimulated to transmit information. When the blob at the top (the cell body) is stimulated, it causes Na^+ ions to rush into the top of the tail (axon). This electrical pulse will then travel down the axon, like a flame burning down from the end of a fuse, with the Na^+ ions at each point first going out and then coming back in. If 10^{10} Na^+ ions cross the cell membrane in 0.5 ms, what amount of current is created?

(answer check available at lightandmatter.com)

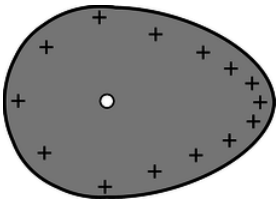
3. If a typical light bulb draws about 900 mA from a 110-V household circuit, what is its resistance? (Don't worry about the fact that it's alternating current.) (answer check available at lightandmatter.com)

4. (a) Express the power dissipated by a resistor in terms of R and ΔV only, eliminating I. (answer check available at lightandmatter.com)
(b) Electrical receptacles in your home are mostly 110 V, but circuits for electric stoves, air conditioners, and washers and driers 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? (answer check available at lightandmatter.com)

5. 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. (answer check available at lightandmatter.com)

6. 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. (answer check available at lightandmatter.com)
(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.

7. 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. (answer check available at lightandmatter.com)



b / Problem 8.

8. Problem 8 has been deleted.

9. Use the result of problem 38 on page 550 to find an equation for the voltage at a point in space at a distance r from a point charge Q. (Take your $V = 0$ distance to be anywhere you like.) (answer check available at lightandmatter.com)

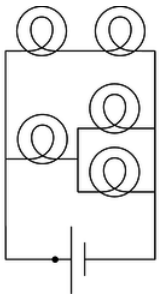
10. Today, even a big luxury car like a Cadillac can have an electrical system that is relatively low in power, since it doesn't need to do much more than run headlights, power windows, etc. In the near future, however, manufacturers plan to start making cars with electrical systems about five times more powerful. This will allow certain energy-wasting parts like the water pump to be run on

electrical motors and turned off when they're not needed --- currently they're run directly on shafts from the motor, so they can't be shut off. It may even be possible to make an engine that can shut off at a stoplight and then turn back on again without cranking, since the valves can be electrically powered. Current cars' electrical systems have 12-volt batteries (with 14-volt chargers), but the new systems will have 36-volt batteries (with 42-volt chargers).
(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 36-volt batteries rather than increasing the power without increasing the voltage? (answer check available at lightandmatter.com)

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. What resistance values can be created by combining a 1 $\text{k}\Omega$ resistor and a 10 $\text{k}\Omega$ resistor? (solution in the pdf version of the book)

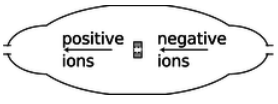
13. The figure shows a circuit containing five lightbulbs connected to a battery. Suppose you're going to connect one probe of a voltmeter to the circuit at the point marked with a dot. How many unique, nonzero voltage differences could you measure by connecting the other probe to other wires in the circuit?



c / Problems 13 and 14.

14. 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.

15. (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? (answer check available at lightandmatter.com)
(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? (answer check available at lightandmatter.com)



d / Problem 16.

16. 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% of the speed of light. 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

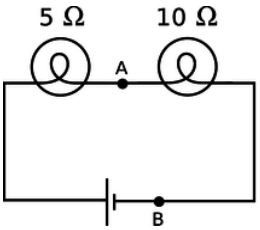
physically penetrate. Passing through the foil strips off some number of electrons, changing the atom into a positive ion, with a charge of $+n$ times the fundamental charge. Now that the atom is positive, it is repelled by the terminal, and accelerates some more on its way out of the accelerator.

(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.

(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? (answer check available at lightandmatter.com)

(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 hydrogen. 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. (answer check available at lightandmatter.com)

17. 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? (answer check available at lightandmatter.com)



e / Problem 18.

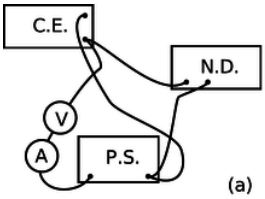
18. In the figure, the battery is 9 V.

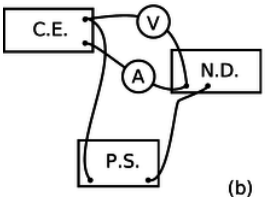
(a) What are the voltage differences across each light bulb?(answer check available at lightandmatter.com)

(b) What current flows through each of the three components of the circuit? (answer check available at lightandmatter.com)

(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?



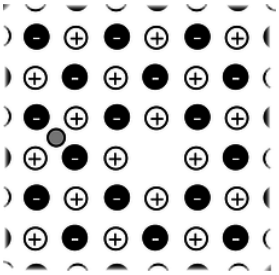


f / Problem 19.

19. 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 1 into a standard-style schematic. What is correct and incorrect about this group's setup?

(b) Do the same for diagram 2.



g / Problem 20.

20. Referring back to problem 15 on page 506 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: The approach is similar to the one used for the other problem, but you want to work with voltage and electrical energy rather than force.] (answer check available at lightandmatter.com)

21. A 1.0 Ω toaster and a 2.0 Ω 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. (answer check available at lightandmatter.com)
(c) Suppose they were instead hooked up in series. Draw a schematic and calculate the same things.(answer check available at lightandmatter.com)

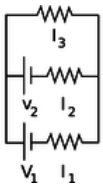
22. The heating element of an electric stove is connected in series with a switch that opens and closes many times per second. When you turn the knob up for more power, the fraction of the time that the switch is closed increases. Suppose someone suggests a simpler alternative for controlling the power 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?

23. 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?

24. How many different resistance values can be created by combining three unequal resistors? (Don't count possibilities in which not all the resistors are used, i.e., ones in which there is zero current in one or more of them.)

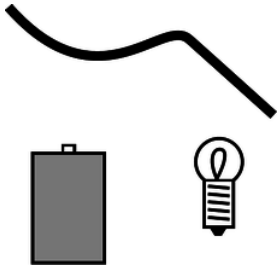
25. Suppose six identical resistors, each with resistance R, are connected so that they form the edges of a tetrahedron (a pyramid with three sides in addition to the base, i.e., one less side than an Egyptian pyramid). What resistance value or values can be obtained by making connections onto any two points on this arrangement? (solution in the pdf version of the book)

26. 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.



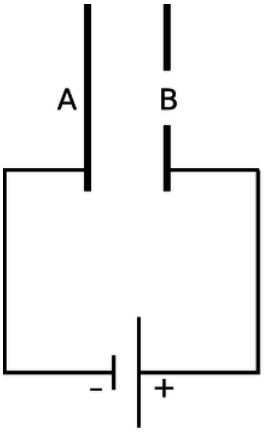
h / Problem 27.

27. All three resistors have the same resistance, R. Find the three unknown currents in terms of V_1, V_2, and R. (answer check available at lightandmatter.com)



i / Problem 28.

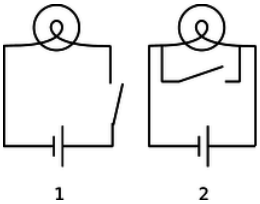
28. You are given a battery, a flashlight bulb, and a single piece of wire. Draw at least two configurations of these items that would result in lighting up the bulb, and at least two that would not light it. (Don't draw schematics.) If you're not sure what's going on, borrow the materials from your instructor and try it. 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). [Problem by Arnold Arons.]



j / Problem 29.

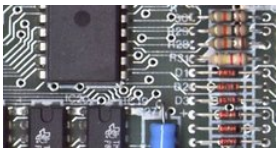
29. The figure shows a simplified diagram of an electron gun such as the one that creates the electron beam in a TV tube. Electrons that spontaneously emerge from the negative electrode (cathode) are then accelerated to the positive electrode, which has a hole in it. (Once they emerge through the hole, 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, and that the velocity is nonrelativistic. (If you haven't read ch. 7 yet, don't worry about the remark about relativity.) (answer check available at lightandmatter.com)
(b) Evaluate your expression numerically for the case where ΔV=10 kV, and compare to the speed of light. If you've read ch. 7 already, comment on whether the assumption of nonrelativistic motion was justified. (solution in the pdf version of the book) (answer check available at lightandmatter.com)

30. (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?



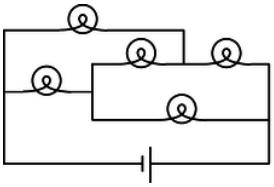
k / Problem 31.

31. 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?



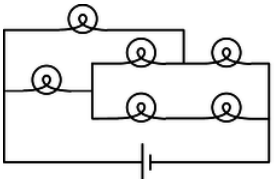
l / A printed circuit board, like the kind referred to in problem 32.

32. 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 in the pdf version of the book)



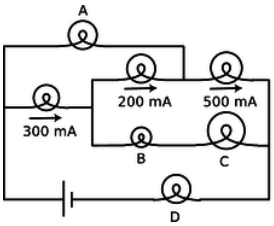
m / Problem 33.

33. The bulbs are all identical. Which one doesn't light up?



n / Problem 34.

34. Each bulb has a resistance of one ohm. How much power is drawn from the one-volt battery? (answer check available at lightandmatter.com)



o / Problem 35.

35. The bulbs all have unequal resistances. Given the three currents shown in the figure, find the currents through bulbs A, B, C, and D.

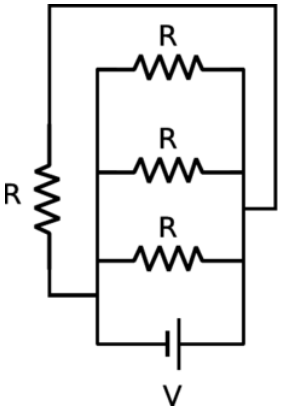
36. A silk thread is uniformly charged by rubbing it with llama fur. The thread is then dangled vertically above a metal plate and released. As each part of the thread makes contact with the conducting plate, its charge is deposited onto the plate. Since the thread is accelerating due to gravity, the 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.(answer check available at lightandmatter.com)
(b) Suppose that the charge is immediately carried away through a resistance R. Find the power dissipated as heat.(answer check available at lightandmatter.com)

37. In example 9 on p. 524, suppose that the larger sphere has radius a, the smaller one b. (a) Use the result of problem 9 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.

38. (a) Recall that the gravitational energy of two gravitationally interacting spheres is given by $PE = -Gm_1m_2/r$, where r is the center-to-center distance. Sketch a graph of PE as a function of r, making sure that your graph behaves properly at small

values of r, where you're dividing by a small number, and at large ones, where you're dividing by a large one. Check that your graph behaves properly when a rock is dropped from a larger r to a smaller one; the rock should *lose* potential energy as it gains kinetic energy.

(b) Electrical forces are closely analogous to gravitational ones, since both depend on $1/r^2$. Since the forces are analogous, the potential energies should also behave analogously. Using this analogy, write down the expression for the electrical potential energy of two interacting charged particles. The main uncertainty here is the sign out in front. Like masses attract, but like charges repel. To figure out whether you have the right sign in your equation, sketch graphs in the case where both charges are positive, and also in the case where one is positive and one negative; make sure that in both cases, when the charges are released near one another, their motion causes them to lose PE while gaining KE.(answer check available at lightandmatter.com)



p / Problem 39.

39. Find the current drawn from the battery. (answer check available at lightandmatter.com)

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

CHAPTER OVERVIEW

11: Fields

> “Okay. Your duties are as follows: Get Breen. I don't care how you get him, but get him soon. That faker! He posed for twenty years as a scientist without ever being apprehended. Well, I'm going to do some apprehending that'll make all previous apprehending look like no apprehension at all. You with me?”
>
> “Yes,” said Battle, very much confused. “What's that thing you have?”
>
> “Piggy-back heat-ray. You transpose the air in its path into an unstable isotope which tends to carry all energy as heat. Then you shoot your juice light, or whatever along the isotopic path and you burn whatever's on the receiving end. You want a few?”
>
> “No,” said Battle. “I have my gats. What else have you got for offense and defense?” Underbottam opened a cabinet and proudly waved an arm. “Everything,” he said.
>
> “Disintegraters, heat-rays, bombs of every type. And impenetrable shields of energy, massive and portable. What more do I need?”

From THE REVERSIBLE REVOLUTIONS by Cecil Corwin, Cosmic Stories, March 1941. Art by Morey, Bok, Kyle, Hunt, Forte. Copyright expired.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

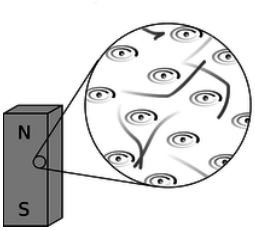
11.1: Fields of Force

Cutting-edge science readily infiltrates popular culture, though sometimes in garbled form. The Newtonian imagination populated the universe mostly with that nice solid stuff called matter, which was made of little hard balls called atoms. In the early twentieth century, consumers of pulp fiction and popularized science began to hear of a new image of the universe, full of x-rays, N-rays, and Hertzian waves. What they were beginning to soak up through their skins was a drastic revision of Newton's concept of a universe made of chunks of matter which happened to interact via forces. In the newly emerging picture, the universe was *made* of force, or, to be more technically accurate, of ripples in universal fields of force. Unlike the average reader of Cosmic Stories in 1941, you now possess enough technical background to understand what a “force field” really is.

10.1.1 Why fields?

Time delays in forces exerted at a distance

What convinced physicists that they needed this new concept of a field of force? Although we have been dealing mostly with electrical forces, let's start with a magnetic example. (In fact the main reason I've delayed a detailed discussion of magnetism for so long is that mathematical calculations of magnetic effects are handled much more easily with the concept of a field of force.) First a little background leading up to our example.

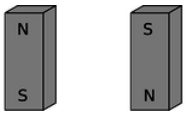


a / A bar magnet's atoms are (partially) aligned.

A bar magnet, a, has an axis about which many of the electrons' orbits are oriented. The earth itself is also a magnet, although not a bar-shaped one.

b / A bar magnet interacts with our magnetic planet.

The interaction between the earth-magnet and the bar magnet, b, makes them want to line up their axes in opposing directions (in other words such that their electrons rotate in parallel planes, but with one set rotating clockwise and the other counterclockwise as seen looking along the axes).



c / Magnets aligned north-south.

On a smaller scale, any two bar magnets placed near each other will try to align themselves head-to-tail, c.

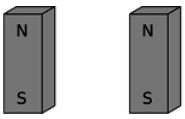
Now we get to the relevant example. It is clear that two people separated by a paper-thin wall could use a pair of bar magnets to signal to each other. Each person would feel her own magnet trying to twist around in response to any rotation performed by the other person's magnet. The practical range of communication would be very short for this setup, but a sensitive electrical apparatus could pick up magnetic signals from much farther away. In fact, this is not so different from what a radio does: the electrons racing up and down the transmitting antenna create forces on the electrons in the distant receiving antenna. (Both magnetic and electric forces are involved in real radio signals, but we don't need to worry about that yet.)

A question now naturally arises as to whether there is any time delay in this kind of communication via magnetic (and electric) forces. Newton would have thought not, since he conceived of physics in terms of instantaneous action at a distance. We now know, however, that there is such a time delay. If you make a long-distance phone call that is routed through a communications satellite, you should easily be able to detect a delay of about half a second over the signal's round trip of 50,000 miles. Modern measurements have shown that electric, magnetic, and gravitational forces all travel at the speed of light, 3×10^8 m/s. (In fact, we will soon discuss how light itself is made of electricity and magnetism.)

If it takes some time for forces to be transmitted through space, then apparently there is some *thing* that travels *through* space. The fact that the phenomenon travels outward at the same speed in all directions strongly evokes wave metaphors such as ripples on a pond.

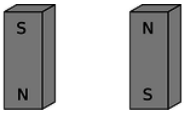
More evidence that fields of force are real: they carry energy.

The smoking-gun argument for this strange notion of traveling force ripples comes from the fact that they carry energy.



d / The second magnet is reversed.

First suppose that the person holding the bar magnet on the right decides to reverse hers, resulting in configuration d. She had to do mechanical work to twist it, and if she releases the magnet, energy will be released as it flips back to c. She has apparently stored energy by going from c to d. So far everything is easily explained without the concept of a field of force.



e / Both magnets are reversed.

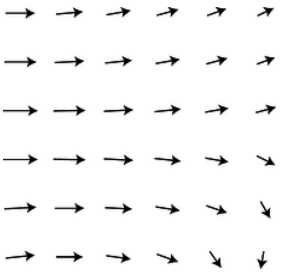
But now imagine that the two people start in position c and then simultaneously flip their magnets extremely quickly to position e, keeping them lined up with each other the whole time. Imagine, for the sake of argument, that they can do this so quickly that each magnet is reversed while the force signal from the other is still in transit. (For a more realistic example, we'd have to have two radio antennas, not two magnets, but the magnets are easier to visualize.) During the flipping, each magnet is still feeling the forces arising from the way the other magnet *used* to be oriented. Even though the two magnets stay aligned during the flip, the time delay causes each person to feel resistance as she twists her magnet around. How can this be? Both of them are apparently doing mechanical work, so they must be storing magnetic energy somehow. But in the traditional Newtonian conception of matter interacting via instantaneous forces at a distance, interaction energy arises from the relative positions of objects that are interacting via forces. If the magnets never changed their orientations relative to each other, how can any magnetic energy have been stored?

The only possible answer is that the energy must have gone into the magnetic force ripples crisscrossing the space between the magnets. Fields of force apparently carry energy across space, which is strong evidence that they are real things.

This is perhaps not as radical an idea to us as it was to our ancestors. We are used to the idea that a radio transmitting antenna consumes a great deal of power, and somehow spews it out into the universe. A person working around such an antenna needs to be careful not to get too close to it, since all that energy can easily cook flesh (a painful phenomenon known as an “RF burn”).

10.1.2 The gravitational field

Given that fields of force are real, how do we define, measure, and calculate them? A fruitful metaphor will be the wind patterns experienced by a sailing ship. Wherever the ship goes, it will feel a certain amount of force from the wind, and that force will be in a certain direction. The weather is ever-changing, of course, but for now let's just imagine steady wind patterns. Definitions in physics are operational, i.e., they describe how to measure the thing being defined. The ship's captain can measure the wind's “field of force” by going to the location of interest and determining both the direction of the wind and the strength with which it is blowing. Charting all these measurements on a map leads to a depiction of the field of wind force like the one shown in the figure. This is known as the “sea of arrows” method of visualizing a field.



f / The wind patterns in a certain area of the ocean could be charted in a “sea of arrows” representation like this. Each arrow represents both the wind's strength and its direction at a certain location.

Now let's see how these concepts are applied to the fundamental force fields of the universe. We'll start with the gravitational field, which is the easiest to understand. As with the wind patterns, we'll start by imagining gravity as a static field, even though the existence of the tides proves that there are continual changes in the gravity field in our region of space. When the gravitational field was introduced in chapter 2, I avoided discussing its direction explicitly, but defining it is easy enough: we simply go to the location of interest and measure the direction of the gravitational force on an object, such as a weight tied to the end of a string.

In chapter 2, I defined the gravitational field in terms of the energy required to raise a unit mass through a unit distance. However, I'm going to give a different definition now, using an approach that will be more easily adapted to electric and magnetic fields. This approach is based on force rather than energy. We couldn't carry out the energy-based definition without dividing by the mass of the object involved, and the same is true for the force-based definition. For example, gravitational forces are weaker on the moon than on the earth, but we cannot specify the strength of gravity simply by giving a certain number of newtons. The number of newtons of gravitational force depends not just on the strength of the local gravitational field but also on the mass of the object on which we're testing gravity, our “test mass.” A boulder on the moon feels a stronger gravitational force than a pebble on the earth. We can get around this problem by defining the strength of the gravitational field as the force acting on an object, *divided by the object's mass:*

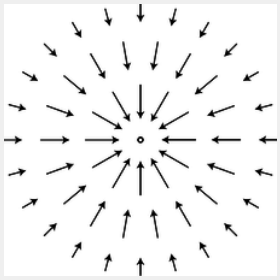
The gravitational field vector, $\mathbf{g}$, at any location in space is found by placing a test mass m_t at that point. The field vector is then given by $\mathbf{g} = \mathbf{F}/m_t$, where $\mathbf{F}$ is the gravitational force on the test mass.

We now have three ways of representing a gravitational field. The magnitude of the gravitational field near the surface of the earth, for instance, could be written as 9.8 N/kg, 9.8 $\mathrm{J/kg \cdot m}$, or 9.8 $\mathrm{m/s^2}$. If we already had two names for it, why invent a third? The main reason is that it prepares us with the right approach for defining other fields.

The most subtle point about all this is that the gravitational field tells us about what forces *would* be exerted on a test mass by the earth, sun, moon, and the rest of the universe, *if* we inserted a test mass at the point in question. The field still exists at all the places where we didn't measure it.

Example 1: Gravitational field of the earth

▷ What is the magnitude of the earth's gravitational field, in terms of its mass, M, and the distance r from its center?



g / The gravitational field surrounding a clump of mass such as the earth.

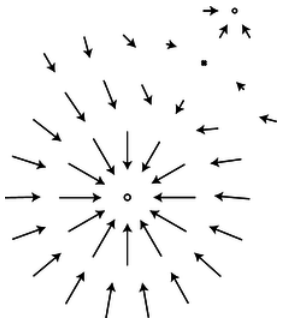
▷ Substituting $|\mathbf{F}| = GMm_t/r^2$ into the definition of the gravitational field, we find $|\mathbf{g}| = GM/r^2$. This expression could be used for the field of any spherically symmetric mass distribution, since the equation we assumed for the gravitational force would apply in any such case.

Sources and sinks

If we make a sea-of-arrows picture of the gravitational fields surrounding the earth, g, the result is evocative of water going down a drain. For this reason, anything that creates an inward-pointing field around itself is called a sink. The earth is a gravitational sink. The term “source” can refer specifically to things that make outward fields, or it can be used as a more general term for both “outies” and “innies.” However confusing the terminology, we know that gravitational fields are only attractive, so we will never find a region of space with an outward-pointing field pattern.

Knowledge of the field is interchangeable with knowledge of its sources (at least in the case of a static, unchanging field). If aliens saw the earth's gravitational field pattern they could immediately infer the existence of the planet, and conversely if they knew the mass of the earth they could predict its influence on the surrounding gravitational field.

Superposition of indexsuperposition of fieldsindexfieldssuperposition offields



h / The gravitational fields of the earth and moon superpose. Note how the fields cancel at one point, and how there is no boundary between the interpenetrating fields surrounding the two bodies.

A very important fact about all fields of force is that when there is more than one source (or sink), the fields add according to the rules of vector addition. The gravitational field certainly will have this property, since it is defined in terms of the force on a test mass, and forces add like vectors. Superposition is an important characteristics of waves, so the superposition property of fields is consistent with the idea that disturbances can propagate outward as waves in a field.

Example 2: Reduction in gravity on Io due to Jupiter's gravity

▷ The average gravitational field on Jupiter's moon Io is 1.81 N/kg. By how much is this reduced when Jupiter is directly overhead? Io's orbit has a radius of 4.22×10^8 m, and Jupiter's mass is 1.899×10^{27} kg.

▷ By the shell theorem, we can treat the Jupiter as if its mass was all concentrated at its center, and likewise for Io. If we visit Io and land at the point where Jupiter is overhead, we are on the same line as these two centers, so the whole problem can be treated one-dimensionally, and vector addition is just like scalar addition. Let's use positive numbers for downward fields (toward the center of Io) and negative for upward ones. Plugging the appropriate data into the expression derived in example 1, we find that the Jupiter's contribution to the field is -0.71 N/kg. Superposition says that we can find the actual gravitational field by adding up the fields created by Io and Jupiter: $1.81 - 0.71$ N/kg = 1.1 N/kg. You might think that this reduction would create some spectacular effects, and make Io an exciting tourist destination. Actually you would not detect any difference if you flew from one side of Io to the other. This is because your body and Io both experience Jupiter's gravity, so you follow the same orbital curve through the space around Jupiter.

i / The part of the LIGO gravity wave detector at Hanford Nuclear Reservation, near Richland, Washington. The other half of the detector is in Louisiana.

Gravitational waves

A source that sits still will create a static field pattern, like a steel ball sitting peacefully on a sheet of rubber. A moving source will create a spreading wave pattern in the field, like a bug thrashing on the surface of a pond. Although we have started with the gravitational field as the simplest example of a static field, stars and planets do more stately gliding than thrashing, so gravitational waves are not easy to detect. Newton's theory of gravity does not describe gravitational waves, but they are predicted by Einstein's general theory of relativity. J.H. Taylor and R.A. Hulse were awarded the Nobel Prize in 1993 for giving indirect evidence that Einstein's waves actually exist. They discovered a pair of exotic, ultra-dense stars called neutron stars orbiting one another very closely, and showed that they were losing orbital energy at the rate predicted by Einstein's theory.

A Caltech-MIT collaboration has built a pair of gravitational wave detectors called LIGO to search for more direct evidence of gravitational waves. Since they are essentially the most sensitive vibration detectors ever made, they are located in quiet rural areas, and signals will be compared between them to make sure that they were not due to passing trucks. The project began operating at full sensitivity in 2005, and is now able to detect a vibration that causes a change of 10^{-18} m in the distance between the mirrors at the ends of the 4-km vacuum tunnels. This is a thousand times less than the size of an atomic nucleus! There is only enough funding to keep the detectors operating for a few more years, so the physicists can only hope that during that time, somewhere in the universe, a sufficiently violent cataclysm will occur to make a detectable gravitational wave. (More accurately, they want the wave to arrive in our solar system during that time, although it will have been produced millions of years before.)

10.1.3 The electric field

Definition

The definition of the electric field is directly analogous to, and has the same motivation as, the definition of the gravitational field:

The electric field vector, $\mathbf{E}$, at any location in space is found by placing a test charge q_t at that point. The electric field vector is then given by $\mathbf{E} = \mathbf{F}/q_t$, where $\mathbf{F}$ is the electric force on the test charge.

Charges are what create electric fields. Unlike gravity, which is always attractive, electricity displays both attraction and repulsion. A positive charge is a source of electric fields, and a negative one is a sink.

The most difficult point about the definition of the electric field is that the force on a negative charge is in the opposite direction compared to the field. This follows from the definition, since dividing a vector by a negative number reverses its direction. It's as though we had some objects that fell upward instead of down.

self-check:

Find an equation for the magnitude of the field of a single point charge Q.

(answer in the back of the PDF version of the book)

Example 3: Superposition of electric fields

▷ Charges q and $-q$ are at a distance b from each other, as shown in the figure. What is the electric field at the point P, which lies at a third corner of the square?

j / Example 3.

▷ The field at P is the vector sum of the fields that would have been created by the two charges independently. Let positive x be to the right and let positive y be up.

Negative charges have fields that point at them, so the charge $-q$ makes a field that points to the right, i.e., has a positive x component. Using the answer to the self-check, we have

$$E_{-q,x} = \frac{kq}{b^2}$$
$$E_{-q,y} = 0.$$

Note that if we had blindly ignored the absolute value signs and plugged in $-q$ to the equation, we would have incorrectly concluded that the field went to the left.

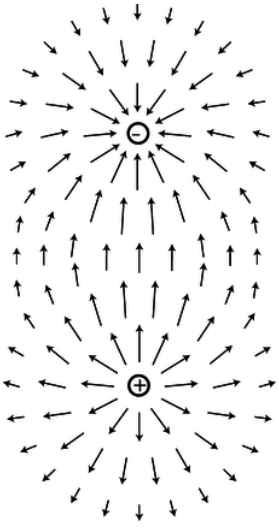
By the Pythagorean theorem, the positive charge is at a distance $\sqrt{2}b$ from P, so the magnitude of its contribution to the field is $E = kq/2b^2$. Positive charges have fields that point away from them, so the field vector is at an angle of 135° counterclockwise from the x axis.

$$\begin{aligned} E_{q,x} &= \frac{kq}{2b^2}\cos 135^\circ \\ &= -\frac{kq}{2^{3/2}b^2} \\ E_{q,y} &= \frac{kq}{2b^2}\sin 135^\circ \\ &= \frac{kq}{2^{3/2}b^2} \end{aligned}$$

The total field is

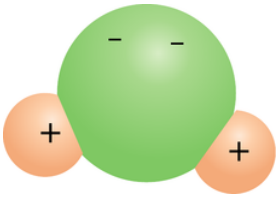
$$E_x = \left(1 - 2^{-3/2}\right)\frac{kq}{b^2}$$
$$E_y = \frac{kq}{2^{3/2}b^2}$$

Dipoles



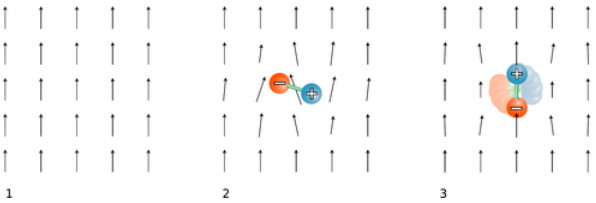
k / A dipole field. Electric fields diverge from a positive charge and converge on a negative charge.

The simplest set of sources that can occur with electricity but not with gravity is the *dipole*, consisting of a positive charge and a negative charge with equal magnitudes. More generally, an electric dipole can be any object with an imbalance of positive charge on one side and negative on the other.



l / A water molecule is a dipole.

A water molecule, l, is a dipole because the electrons tend to shift away from the hydrogen atoms and onto the oxygen atom.



m / 1. A uniform electric field created by some charges “off-stage.” 2. A dipole is placed in the field. 3. The dipole aligns with the field.

Your microwave oven acts on water molecules with electric fields. Let us imagine what happens if we start with a uniform electric field, m/1, made by some external charges, and then insert a dipole, m/2, consisting of two charges connected by a rigid rod. The dipole disturbs the field pattern, but more important for our present purposes is that it experiences a torque. In this example, the positive charge feels an upward force, but the negative charge is pulled down. The result is that the dipole wants to align itself with the field, m/3. The microwave oven heats food with electrical (and magnetic) waves. The alternation of the torque causes the molecules to wiggle and increase the amount of random motion. The slightly vague definition of a dipole given above can be improved by saying that a dipole is any object that experiences a torque in an electric field.

What determines the torque on a dipole placed in an externally created field? Torque depends on the force, the distance from the axis at which the force is applied, and the angle between the force and the line from the axis to the point of application. Let a dipole consisting of charges $+q$ and $-q$ separated by a distance ℓ be placed in an external field of magnitude $|\mathbf{E}|$, at an angle θ with respect to the field. The total torque on the dipole is

$$\begin{aligned}\tau &= \frac{\ell}{2}q|\mathbf{E}|\sin\theta + \frac{\ell}{2}q|\mathbf{E}|\sin\theta \\ &= \ell q|\mathbf{E}|\sin\theta.\end{aligned}$$

(Note that even though the two forces are in opposite directions, the torques do not cancel, because they are both trying to twist the dipole in the same direction.) The quantity is called the dipole moment, notated D. (More complex dipoles can also be assigned a dipole moment --- they are defined as having the same dipole moment as the two-charge dipole that would experience the same torque.)

Employing a little more mathematical elegance, we can define a dipole moment *vector*,

$$\mathbf{D} = \sum q_i \mathbf{r}_i,$$

where $\mathbf{r}_i$ is the position vector of the charge labeled by the index i. We can then write the torque in terms of a vector cross product (page 281),

$$\boldsymbol{\tau} = \mathbf{D} \times \mathbf{E}.$$

No matter how we notate it, the definition of the dipole moment requires that we choose point from which we measure all the position vectors of the charges. However, in the commonly encountered special case where the total charge of the object is zero, the dipole moment is the same regardless of this choice.

Example 4: Dipole moment of a molecule of NaCl gas

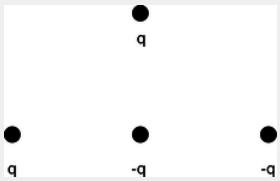
▷ In a molecule of NaCl gas, the center-to-center distance between the two atoms is about 0.6 nm. Assuming that the chlorine completely steals one of the sodium's electrons, compute the magnitude of this molecule's dipole moment.

▷ 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

$$\begin{aligned} D &= (6\times 10^{-10}\text{ m})(e) \\ &= (6\times 10^{-10}\text{ m})(1.6\times 10^{-19}\text{ C}) \\ &= 1\times 10^{-28}\text{ C}\cdot\text{m} \end{aligned}$$

Example 5: Dipole moments as vectors

▷ The horizontal and vertical spacing between the charges in the figure is b. Find the dipole moment.



n / Example 5.

▷ Let the origin of the coordinate system be at the leftmost charge.

$$\begin{aligned} \mathbf{D} &= \sum q_i\mathbf{r}_i \\ &= (q)(0)+(-q)(b\hat{\mathbf{x}})+(q)(b\hat{\mathbf{x}}+b\hat{\mathbf{y}})+(-q)(2b\hat{\mathbf{x}}) \\ &= -2bq\hat{\mathbf{x}}+bq\hat{\mathbf{y}} \end{aligned}$$

Alternative definition of the electric field

The behavior of a dipole in an externally created field leads us to an alternative definition of the electric field:

The electric field vector, E, at any location in space is defined by observing the torque exerted on a test dipole D_t placed there. The direction of the field is the direction in which the field tends to align a dipole (from $-$ to +), and the field's magnitude is $|\mathbf{E}| = \tau/D_t \sin\theta$. In other words, the field vector is the vector that satisfies the equation $\boldsymbol{\tau} = \mathbf{D}_t \times \mathbf{E}$ for any test dipole $\mathbf{D}_t$ placed at that point in space.

The main reason for introducing a second definition for the same concept is that the magnetic field is most easily defined using a similar approach.

Discussion Questions

◊ In the definition of the electric field, does the test charge need to be 1 coulomb? Does it need to be positive?

◊ Does a charged particle such as an electron or proton feel a force from its own electric field?

◊ Is there an electric field surrounding a wall socket that has nothing plugged into it, or a battery that is just sitting on a table?

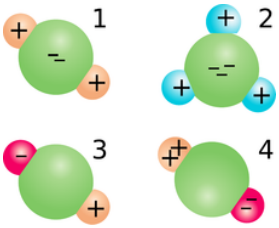
◊ In a flashlight powered by a battery, which way do the electric fields point? What would the fields be like inside the wires? Inside the filament of the bulb?

◊ Criticize the following statement: “An electric field can be represented by a sea of arrows showing how current is flowing.”

◊ The field of a point charge, $|\mathbf{E}| = kQ/r^2$, was derived in a self-check. How would the field pattern of a uniformly charged sphere compare with the field of a point charge?

◊ The interior of a perfect electrical conductor in equilibrium must have zero electric field, since otherwise the free charges within it would be drifting in response to the field, and it would not be in equilibrium. What about the field right at the surface of a perfect conductor? Consider the possibility of a field perpendicular to the surface or parallel to it.

◊ Compare the dipole moments of the molecules and molecular ions shown in the figure.



o / Discussion question H.

◊ Small pieces of paper that have not been electrically prepared in any way can be picked up with a charged object such as a charged piece of tape. In our new terminology, we could describe the tape's charge as inducing a dipole moment in the paper. Can a similar technique be used to induce not just a dipole moment but a charge?

Contributors

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

10.2.1 One dimension

Voltage is electrical energy per unit charge, and electric field is force per unit charge. For a particle moving in one dimension, along the x axis, we can therefore relate voltage and field if we start from the relationship between interaction energy and force,

$$dU = -F_x dx,$$

and divide by charge,

$$\frac{dU}{q} = -\frac{F_x}{q} dx,$$

giving

$$dV = -E_x dx,$$

or

$$\frac{dV}{dx} = -E_x.$$

The interpretation is that a strong electric field occurs in a region of space where the voltage is rapidly changing. By analogy, a steep hillside is a place on the map where the altitude is rapidly changing.

Example 6: Field generated by an electric eel

▷ 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. Subject to the possibly inaccurate assumption of a constant field parallel to the eel's body, we have

$$\begin{aligned} |\mathbf{E}| &= \frac{dV}{dx} \\ &\approx \frac{\Delta V}{\Delta x} [\text{assumption of constant field}] \\ &= 1000 \text{ V/m}. \end{aligned}$$

Example 7: Relating the units of electric field and voltage

From our original definition of the electric field, we expect it to have units of newtons per coulomb, N/C. The example above, however, came out in volts per meter, V/m. Are these inconsistent? Let's reassure ourselves that this all works. In this kind of situation, the best strategy is usually to simplify the more complex units so that they involve only mks units and coulombs. Since voltage is defined as electrical energy per unit charge, it has units of J/C:

$$\begin{aligned} \frac{\text{V}}{\text{m}} &= \frac{\text{J/C}}{\text{m}} \\ &= \frac{\text{J}}{\text{C} \cdot \text{m}}. \end{aligned}$$

To connect joules to newtons, we recall that work equals force times distance, so $\text{J} = \text{N} \cdot \text{m}$, so

$$\begin{aligned} \frac{\text{V}}{\text{m}} &= \frac{\text{N} \cdot \text{m}}{\text{C} \cdot \text{m}} \\ &= \frac{\text{N}}{\text{C}} \end{aligned}$$

As with other such difficulties with electrical units, one quickly begins to recognize frequently occurring combinations.

Example 8: Voltage associated with a point charge

▷ What is the voltage associated with a point charge?

▷ As derived previously in self-check A on page 563, the field is

$$|\mathbf{E}| = \frac{kQ}{r^2}$$

The difference in voltage between two points on the same radius line is

$$\begin{aligned}\Delta V &= -\int dV \\ &= -\int E_x dx\end{aligned}$$

In the general discussion above, x was just a generic name for distance traveled along the line from one point to the other, so in this case x really means r.

$$\begin{aligned}\Delta V &= -\int_{r_1}^{r_2} E_r dr \\ &= -\int_{r_1}^{r_2} \frac{kQ}{r^2} dr \\ &= \frac{kQ}{r}\Big]_{r_1}^{r_2} \qquad\qquad = \frac{kQ}{r_2} - \frac{kQ}{r_1}.\end{aligned}$$

The standard convention is to use $r_1 = \infty$ as a reference point, so that the voltage at any distance r from the charge is

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

The interpretation is that if you bring a positive test charge closer to a positive charge, its electrical energy is increased; if it was released, it would spring away, releasing this as kinetic energy.

self-check:

Show that you can recover the expression for the field of a point charge by evaluating the derivative $E_x = -dV/dx$.

(answer in the back of the PDF version of the book)

10.2.2 Two or three dimensions

a / A topographical map of Shelburne Falls, Mass. (USGS).

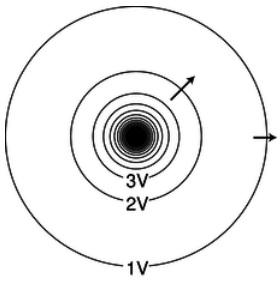
The topographical map in figure a suggests a good way to visualize the relationship between field and voltage in two dimensions. Each contour on the map is a line of constant height; some of these are labeled with their elevations in units of feet. Height is related to gravitational energy, so in a gravitational analogy, we can think of height as representing voltage. 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 (voltage) is a function of the east-west coordinate x. Using the usual mathematical definition of the slope, and writing V for the height in order to remind us of the electrical analogy,

the slope along such a line is dV/dx (the rise over the run).

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 voltages in the same way, as shown in figure b. The electric field is strongest where the constant-voltage curves are closest together, and the electric field vectors always point perpendicular to the constant-voltage curves.



b / The constant-voltage curves 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.

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

$$E_x = -\frac{dV}{dx}$$
$$E_y = -\frac{dV}{dy}$$
$$E_z = -\frac{dV}{dz}$$

This can be notated as a gradient (page 215),

$$\mathbf{E} = \nabla V,$$

and if we know the field and want to find the voltage, we can use a line integral,

$$\Delta V = \int_C \mathbf{E} \cdot d\mathbf{r},$$

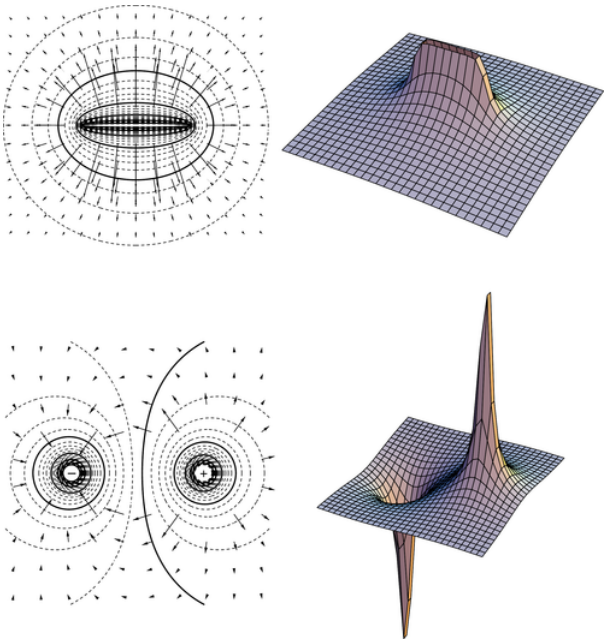
where the quantity inside the integral is a vector dot product.

self-check:

Imagine that figure a represents voltage rather than height. (a) Consider the stream the starts near the center of the map. Determine the positive and negative signs of dV/dx and dV/dy, and relate these to the direction of the force that is pushing the current forward against the resistance of friction. (b) If you wanted to find a lot of electric charge on this map, where would you look?

(answer in the back of the PDF version of the book)

Figure c shows some examples of ways to visualize field and voltage patterns.



c / Two-dimensional field and voltage patterns. Top: A uniformly charged rod. Bottom: A dipole. In each case, the diagram on the left shows the field vectors and constant-voltage curves, while the one on the right shows the voltage (up-down coordinate) as a function of x and y. Interpreting the field diagrams: Each arrow represents the field at the point where its tail has been positioned. For clarity, some of the arrows in regions of very strong field strength are not shown --- they would be too long to show. Interpreting the constant-voltage curves: In regions of very strong fields, the curves are not shown because they would merge together to make solid black regions. Interpreting the perspective plots: Keep in mind that even though we're visualizing things in three dimensions, these are really two-dimensional voltage patterns being represented. The third (up-down) dimension represents voltage, not position.

Contributors

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

11.3: Fields by Superposition

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

11.4: Energy In Fields

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

11.5: LRC Circuits

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 (or some more recent guitar hero) 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. You have already studied negative feedback extensively in section 3.3 in the case of a mechanical system, although we didn't call it that.

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. 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.

Capacitors

A capacitor's energy exists in its surrounding electric fields. It 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.

a / The symbol for a capacitor.

b / Some capacitors.

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 millifarad and farad values became available in practical physical sizes. Figure a shows the symbol used in schematics to represent a capacitor.

Example 11.5.1: A parallel-plate capacitor

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?

Solution

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. The result of example 14 then applies, and for the region between the plates, we have

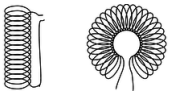
$$E = 4\pi k\sigma = 4\pi kq/A \tag{11.5.1}$$

and

$$U_e = (1/8\pi k)E^2Ah. \tag{11.5.2}$$

Substituting the first expression into the second, we find $U_e = 2\pi kq^2h/A$. Comparing this to the definition of capacitance, we end up with $C = A/4\pi kh$.

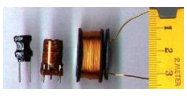
Inductors



c / Two common geometries for inductors. The cylindrical shape on the left is called a solenoid.

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, c, and a stylized solenoid, d, is the symbol used to represent an inductor in a circuit regardless of its actual geometry.

d / The symbol for an inductor.



e / Some inductors.

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

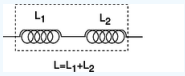
$$U_L = \frac{L}{2}I^2.$$

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 henry = 1 $\mathrm{J/A^2}$. 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!"

There is a lumped circuit approximation for inductors, just like the one for capacitors. 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.

Example 11.5.2: Identical inductances in series

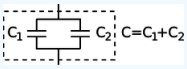
If two inductors are placed in series, any current that passes through the combined double inductor must pass through both its parts. If we assume the lumped circuit approximation, the two inductors' fields don't interfere with each other, so the energy is doubled for a given current. Thus by the definition 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.)



f / Inductances in series add.

Example 11.5.3: Identical capacitances in parallel

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.



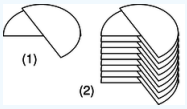
g / Capacitances in parallel add.

This is consistent with the result of example 18, 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 homework problems 36 and 33.

Example 11.5.4: A variable capacitor

Figure h/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.

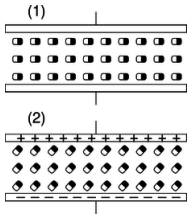


h / A variable capacitor.

Discussion Questions

◊ 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?

◊ 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.



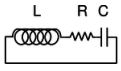
i / Discussion question B.

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?

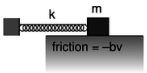
Now redo the analysis in terms of the mechanical work needed in order to charge up the plates.

10.5.2 Oscillations



j / A series LRC circuit.

Figure j 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.



k / A mechanical analogy for the LRC circuit.

We can analyze it without any sweat or tears whatsoever, simply by constructing an analogy with a mechanical system. In a mechanical oscillator, k, we have two forms of stored energy,

$$U_{spring} = \frac{1}{2}kx^2 \qquad (1)$$
$$K = \frac{1}{2}mv^2. \qquad (2)$$

In the case of a mechanical oscillator, we have usually assumed a friction force of the form that turns out to give the nicest mathematical results, $F = -bv$. 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, $P = W/dt = Fdx/dt = Fv = -bv^2$, so

$$\text{rate of heat dissipation} = -bv^2. \quad (3)$$

self-check:

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 in the back of the PDF version of the book)

In the circuit, the stored forms of energy are

$$U_C = \frac{1}{2C}q^2 \qquad (1')$$
$$U_L = \frac{1}{2}LI^2, \qquad (2')$$

and the rate of heat dissipation in the resistor is

$$\text{rate of heat dissipation} = -RI^2. \quad (3')$$

Comparing the two sets of equations, we first form analogies between quantities that represent the state of the system at some moment in time:

$$x \leftrightarrow q$$
$$v \leftrightarrow I$$

self-check:

How is v related mathematically to x? How is I connected to q? Are the two relationships analogous? (answer in the back of the PDF version of the book)

Next we relate the ones 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[3] $\omega \approx \sqrt{k/m}$, we can immediately solve the electrical version by analogy, giving

$$\omega \approx \frac{1}{\sqrt{LC}}.$$

Since the resistance R is analogous to 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.

Example 22: Tuning a radio receiver

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.

Example 23: A numerical calculation

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).
This is actually quite a large inductance value, and would require a big, heavy, expensive coil. In fact, there is a trick for making this kind of circuit small and cheap. There is a kind of silicon chip called an op-amp, which, among other things, can be used to simulate the behavior of an inductor. The main limitation of the op-amp is that it is restricted to low-power applications.

10.5.3 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 C: 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.

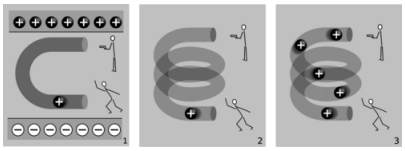
l / The inductor releases energy and gives it to the black box.

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 l, 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, and likewise the loop rule tells us $V_{inductor} + V_{black\ box} = 0$, so the two voltage drops have the same absolute value, which we can refer to as V. Whatever the black box is, the rate at which it is taking energy from the inductor is given by $|P| = |IV|$, so

$$\begin{aligned} |IV| &= \left|\frac{dU_L}{dt}\right| \\ &= \left|\frac{d}{dt}\left(\frac{1}{2}LI^2\right)\right| \\ &= \left|LI\frac{dI}{dt}\right|, \\ \text{or} |V| &= \left|L\frac{dI}{dt}\right|, \end{aligned}$$

which in many books is taken to be the definition of inductance. The direction of the voltage drop (plus or minus sign) is such that the inductor resists the change in current.

There's one very intriguing thing about this result. Suppose, for concreteness, that the black box in figure l 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! What we've discovered is one special case of a more general principle, the principle of induction: a changing magnetic field creates an electric field, which is in addition to any electric field created by charges. (The reverse is also true: any electric field that changes over time creates a magnetic field.) Induction forms the basis for such technologies as the generator and the transformer, and ultimately it leads to the existence of light, which is a wave pattern in the electric and magnetic fields. These are all topics for chapter 11, but it's truly remarkable that we could come to this conclusion without yet having learned any details about magnetism.



m / Electric fields made by charges, 1, and by changing magnetic fields, 2 and 3.

The cartoons in figure m compares electric fields made by charges, 1, to electric fields made by changing magnetic fields, 2-3. In m/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 m/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, m/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.

Example 24: Ballasts

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.

n / 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.

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 fluroescent light's ballast to be 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. These days, the trend is toward using a solid-state circuit that mimics the behavior of an inductor, but at a frequency in the kilohertz range, eliminating the flicker and hum. Modern compact fluorescent bulbs electronic have ballasts built into their bases, so they can be used as plug-in replacements for incandescent bulbs. A compact fluorescent bulb uses about 1/4 the electricity of an incandescent bulb, lasts ten times longer, and saves $30 worth of electricity over its lifetime.

Discussion Question

◊ What happens when the physicist at the bottom in figure m/3 starts getting tired, and decreases the current?

10.5.4 Decay

Up until now I've soft-pedaled 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. For friction proportional to v, as we've been assuming, the result is that the velocity approaches zero, but never actually reaches zero. This is unrealistic for the mechanical oscillator, which will not have vanishing friction at low velocities, but it is quite realistic in the case of an electrical circuit, for which the voltage drop across the resistor really does approach zero as the current approaches zero.

We do not even have to reduce k to exactly zero in order to get non-oscillatory behavior. There is actually a finite, critical value below which the behavior changes, so that the mass never even makes it through one cycle. This is the case of overdamping, discussed on page 186.

Electrical circuits can exhibit all the same behavior. For simplicity we will analyze only the cases of LRC circuits with $L = 0$ or $C = 0$.

The RC circuit

o / An RC circuit.

We first analyze the RC circuit, o. In reality one would have to “kick” the circuit, for example by briefly inserting a battery, in order to get any interesting behavior. We start with Ohm's law and the equation for the voltage across a capacitor:

$$V_R = IR$$
$$V_C = q/C$$

The loop rule tells us

$$V_R + V_C = 0,$$

and combining the three equations results in a relationship between q and I:

$$I = -\frac{1}{RC}q$$

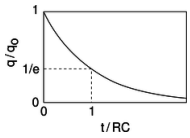
The negative sign tells us that the current tends to reduce the charge on the capacitor, i.e., to discharge it. It makes sense that the current is proportional to q : if q is large, then the attractive forces between the $+q$ and $-q$ charges on the plates of the capacitor are large, and charges will 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

$$\frac{dq}{dt} = -\frac{1}{RC}q.$$

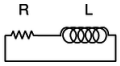
We need a function $q(t)$ whose derivative equals itself, but multiplied by a negative constant. A function of the form ae^t, where $e = 2.718\ldots$ is the base of natural logarithms, is the only one that has its derivative equal to itself, and ae^{bt} has its derivative equal to itself multiplied by b. Thus our solution is

$$q = q_o \exp\left(-\frac{t}{RC}\right).$$



p / Over a time interval RC, the charge on the capacitor is reduced by a factor of e.

The RL circuit



q / An RL circuit.

The RL circuit, q, can be attacked by similar methods, and it can easily be shown that it gives

$$I = I_o \exp\left(-\frac{R}{L}t\right).$$

The RL time constant equals L/R.

Example 25 Death by solenoid; spark plugs

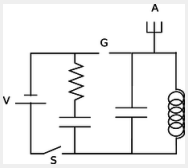
When we suddenly break an RL circuit, what will happen? It might seem that we're faced with a paradox, since we only have two forms of energy, magnetic energy and heat, and if the current stops suddenly, the magnetic field must collapse suddenly. But where does the lost magnetic energy go? It can't go 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.

Example 26: A spark-gap radio transmitter



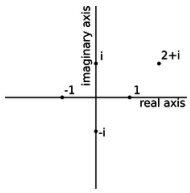
r / Example 26.

Figure r shows a primitive type of radio transmitter, called a spark gap transmitter, used to send Morse code around the turn of the 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 (chapter 11).

Discussion Questions

◊ 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?

10.5.5 Review of complex numbers

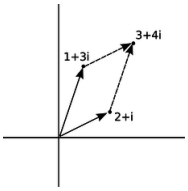


s / Visualizing complex numbers as points in a plane.

For a more detailed treatment of complex numbers, see ch. 3 of James Nearing's free book at http://www.physics.miami.edu/nearing/mathmethods/.

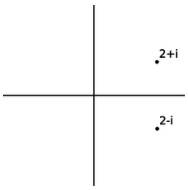
We assume there is a number, i, such that $i^2 = -1$. The square roots of -1 are then i and $-i$. (In electrical engineering work, where i stands for current, j is sometimes used instead.) This gives rise to a number system, called the complex numbers, containing the real numbers as a subset. Any complex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



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

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 notated as $\bar{z}$ or z^*.



u / A complex number and its conjugate.

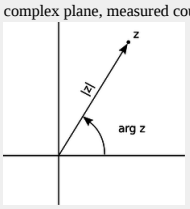
The complex numbers obey all the same rules of arithmetic as the reals, except that they can't be ordered along a single line. That is, it's not possible to say whether one complex number is greater than another. We can compare them in terms of their magnitudes (their distances from the origin), but 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.

Example 27: A square root of i

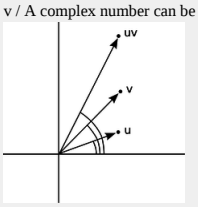
▷ Prove that $1/\sqrt{2}+i/\sqrt{2}$ is a square root of i.
▷ Our proof can use any ordinary rules of arithmetic, except for ordering.

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

Example 27 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.



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



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

self-check:

Using this interpretation of multiplication, how could you find the square roots of a complex number? (answer in the back of the PDF version of the book)

Example 28: An identity

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. [4]

This whole system was built up in order to make every number have square roots. What about cube roots, fourth roots, and so on? Does it get even more weird when you want to do those as well? No. The complex number system we've already discussed is sufficient to handle all of them. The nicest way of thinking about it is in terms of roots of polynomials. In the real number system, the polynomial x^2-1 has two roots, i.e., two values of x (plus and minus one) that we can plug in to the polynomial and get zero. Because it has these two real roots, we can rewrite the polynomial as $(x-1)(x+1)$. However, the polynomial x^2+1 has no real roots. It's ugly that in the real number system, some second-order polynomials have two roots, and can be factored, while others can't. In the complex number system, they all can. For instance, x^2+1 has roots i and $-i$, and can be factored as $(x-i)(x+i)$. In general, the fundamental theorem of algebra states that in the complex number system, any nth-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

◊ Find $\arg i$, $\arg(-i)$, and $\arg 37$, where $\arg z$ denotes the argument of the complex number z.

◊ Visualize the following multiplications in the complex plane using the interpretation of multiplication in terms of multiplying magnitudes and adding arguments: $(i)(i)=-1$, $(i)(-i)=1$, $(-i)(-i)=-1$.

◊ 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 $\sqrt{z}$.

◊ Find four different complex numbers z such that $z^4 = 1$.

◊ Compute the following. Use the magnitude and argument, not the real and imaginary parts.

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

Based on the results above, compute the real and imaginary parts of $1/(1+i)$.

10.5.6 Euler's formula

y / Leonhard Euler (1707-1783).

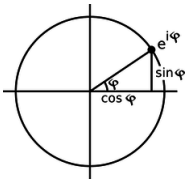
Having expanded our horizons to include the complex numbers, it's natural to want to extend functions we knew and loved from the world of real numbers so that they can also operate on complex numbers. The only really natural way to do this in general is to use Taylor series. A particularly beautiful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

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

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 x.



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

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:

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

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 nth power multiplies its argument by n, giving a number with an argument of ϕ.

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

Example 29: Trig functions in terms of complex exponentials

▷ 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 $\sin(-\theta) = -\sin\theta$.

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

Example 30: A hard integral made easy

▷ 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.[5]

10.5.7 Impedance

So far we have been thinking in terms of the free oscillations of a circuit. This is like a mechanical oscillator that has been kicked but then left to oscillate on its own without any external force to keep the vibrations from dying out. Suppose an LRC circuit is driven with a sinusoidally varying voltage, such as will occur when a radio tuner is hooked up to a receiving antenna. We know that a current will flow in the circuit, and we know that there will be resonant behavior, but it is not necessarily simple to relate current to voltage in the most general case. Let's start instead with the special cases of LRC circuits consisting of only a resistance, only a capacitance, or only an inductance. We are interested only in the steady-state response.

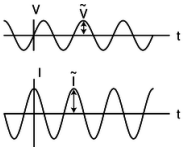
The purely resistive case is easy. Ohm's law gives

$$I = \frac{V}{R}.$$

In the purely capacitive case, the relation $V = q/C$ lets us calculate

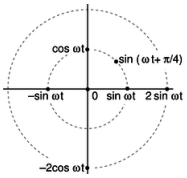
$$\begin{aligned} I &= \frac{dq}{dt} \\ &= C\frac{dV}{dt}. \end{aligned}$$

This is partly analogous to Ohm's law. For example, if we double the amplitude of a sinusoidally varying AC voltage, the derivative dV/dt will also double, and the amplitude of the sinusoidally varying current will also double. However, it is not true that $I = V/R$, because taking the derivative of a sinusoidal function shifts its phase by 90 degrees. If the voltage varies as, for example, $V(t) = V_o \sin(\omega t)$, then the current will be $I(t) = \omega C V_o \cos(\omega t)$. The amplitude of the current is $\omega C V_o$, which is proportional to V_o, but it's not true that $I(t) = V(t)/R$ for some constant R.



z / In a capacitor, the current is 90° ahead of the voltage in phase.

A second problem that crops up is that our entire analysis of DC resistive circuits was built on the foundation of the loop rule and the junction rule, both of which are statements about sums. To apply the junction rule to an AC circuit, for exampe, we would say that the sum of the sine waves describing the currents coming into the junction is equal (at every moment in time) to the sum of the sine waves going out. Now sinusoidal functions have a remarkable property, which is that if you add two different sinusoidal functions having the same frequency, the result is also a sinusoid with that frequency. For example, $\cos\omega t + \sin\omega t = \sqrt{2}\sin(\omega t + \pi/4)$, which can be proved using trig identities. The trig identities can get very cumbersome, however, and there is a much easier technique involving complex numbers.



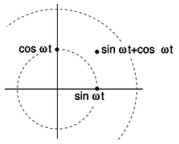
aa / Representing functions with points in polar coordinates.

Figure aa shows a useful way to visualize what's going on. When a circuit is oscillating at a frequency ω, we use points in the plane to represent sinusoidal functions with various phases and amplitudes.

self-check:

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

The simplest examples of how to visualize this in polar coordinates are ones like $\cos\omega t + \cos\omega t = 2\cos\omega t$, where everything has the same phase, so all the points lie along a single line in the polar plot, and addition is just like adding numbers on the number line. The less trivial example $\cos\omega t + \sin\omega t = \sqrt{2}\sin(\omega t + \pi/4)$, can be visualized as in figure ab.



ab / Adding two sinusoidal functions.

Figure ab suggests that all of this can be tied together nicely if we identify our plane with the plane of complex numbers. For example, the complex numbers 1 and i represent the functions $\sin\omega t$ and $\cos\omega t$. In figure z, for example, the voltage across the capacitor 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 a sine wave, 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

$$Z_C = -\frac{i}{\omega C}, \text{[impedance of a capacitor]}$$

having units of ohms, is called 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.

Note that we have two types of complex numbers: those that represent sinusoidal functions of time, and those that represent impedances. The ones that represent sinusoidal functions have tildes on top, which look like little sine waves.

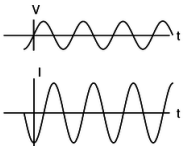
self-check:

Why can't a capacitor have its impedance printed on it along with its capacitance? (answer in the back of the PDF version of the book)

Similar math (but this time with an integral instead of a derivative) gives

$$Z_L = i\omega L \text{[impedance of an inductor]}$$

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.



ac / The current through an inductor lags behind the voltage by a phase angle of 90°.

The phase relationships shown in figures z and ac can be remembered using my own 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. A more traditional mnemonic is "ELI the ICE man," which uses the notation E for emf, a concept closely related to voltage (see p. 686).

Summarizing, the impedances of resistors, capacitors, and inductors are

$$Z_R = R$$
$$Z_C = -\frac{i}{\omega C}$$
$$Z_L = i\omega L.$$

Example 31: Low-pass and high-pass filters

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

◊ Figure z on page 607 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?

◊ Figure ac on page 609 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.

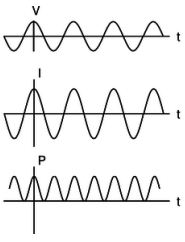
◊ Relate the features of the graph in figure ac on page 609 to the story told in cartoons in figure m/2-3 on page 598.

10.5.8 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.

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 a glowing lightbulb, which is driven by an oscillating voltage with amplitude V_o. In the special case of a resistor, we know that I and V are in phase. For example, if V varies as $V_o \cos \omega t$, then I will be a cosine as well, $I_o \cos \omega t$. The power is then $I_o V_o \cos^2 \omega t$, which is always positive,[6] and varies between 0 and $I_o V_o$. Even if the time variation was $\cos \omega t$ or $\sin(\omega t + \pi/4)$, we would still have a maximum power of $I_o V_o$, 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.[7] Since the circuit oscillates at a frequency[8] 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.



ad / Power in a resistor: the rate at which electrical energy is being converted into heat.

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

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

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

$$P_{av} = \frac{I_o V_o}{2} [\text{average power in a resistor}]$$

for a resistor.

RMS quantities

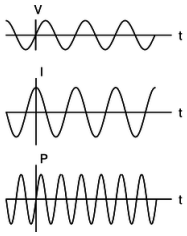
Suppose one day the electric company decided to start supplying your electricity as DC rather than AC. How would the DC voltage have to be related to the amplitude V_o of the AC voltage previously used if they wanted your lightbulbs to have the same brightness as before? The resistance of the bulb, R, 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.

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 $I = CdV/dt$ 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 ae, 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!



ae / Power in a capacitor: the rate at which energy is being stored in (+) or removed from (-) the electric field.

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,

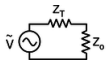
$$P_{av} = 0 [\text{average power in a capacitor}]$$

Notice that although the average power is zero, the power at any given instant is *not* typically zero, as shown in figure ae. The capacitor *does* transfer energy: it's just that after borrowing some energy, it always pays it back in the next quarter-cycle.

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.

10.5.9 Impedance matching



af / We wish to maximize the power delivered to the load, Z_o, by adjusting its impedance.

Figure af 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 is displays both inductance and resistance. (For a purely inductive or capacitive load, P_{av} equals zero, so the problem isn't very interesting!)

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

$$\begin{aligned} I_o &= \frac{V_o}{Z_o+Z_T}, \\ \text{so} P_{av} &= I_o V_o/2 \\ &= I_o^2 Z_o/2 \\ &= \frac{V_o^2 Z_o}{(Z_o+Z_T)^2}/2. \text{ The maximum of this expression occurs where the derivative is zero,} 0 \qquad = \frac{1}{2}\frac{d}{dZ_o}\left[\frac{V_o^2 Z_o}{(Z_o+Z_T)^2}\right] \\ 0 &= \frac{1}{2}\frac{d}{dZ_o}\left[\frac{Z_o}{(Z_o+Z_T)^2}\right] \\ 0 &= (Z_o+Z_T)^{-2} - 2Z_o(Z_o+Z_T)^{-3} \\ 0 &= (Z_o+Z_T) - 2Z_o \\ Z_o &= Z_T \end{aligned}$$

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 two watts were being dissipated as heat inside the amp for every amp 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.

10.5.10 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 as resistances.

Example 32: Series impedance

▷ 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

$$\begin{aligned} Z &= Z_L + Z_C \\ &= i\omega L - \frac{i}{\omega C}. \end{aligned}$$

$$\text{The current is then} \tilde{I} = \frac{\tilde{V}}{i\omega L - i/\omega C}.$$

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 (cf. figure j on page 181).

Example 33: Resonance with damping

▷ What is the amplitude of the current in a series LRC circuit?

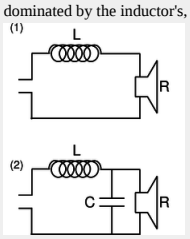
▷ Generalizing from example 32, we add a third, real impedance:

$$\begin{aligned} |\tilde{I}| &= \frac{|\tilde{V}|}{|Z|} \\ &= \frac{|\tilde{V}|}{|R + i\omega L - i/\omega C|} \\ &= \frac{|\tilde{V}|}{\sqrt{R^2 + (\omega L - 1/\omega C)^2}} \end{aligned}$$

This result would have taken pages of algebra without the complex number technique!

Example 34: A second-order stereo crossover filter

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^2R. Putting an inductor in series with the woofer, ag/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}.



ag / Example 34.

A second-order filter, like ag/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}.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

10.6.1 Gauss' law

The flea of subsection 10.3.2 had a long and illustrious scientific career, and we're now going to pick up her story where we left off. This flea, whose name is Gauss[9], has derived the equation $E_{\perp} = 2\pi k\sigma$ for the electric field very close to a charged surface with charge density σ. Next we will describe two improvements she is going to make to that equation.

First, she realizes that the equation is not as useful as it could be, because it only gives the part of the field *due to the surface*. If other charges are nearby, then their fields will add to this field as vectors, and the equation will not be true unless we carefully subtract out the field from the other charges. This is especially problematic for her because the planet on which she lives, known for obscure reasons as planet Flatcat, is itself electrically charged, and so are all the fleas --- the only thing that keeps them from floating off into outer space is that they are negatively charged, while Flatcat carries a positive charge, so they are electrically attracted to it. When Gauss found the original version of her equation, she wanted to demonstrate it to her skeptical colleagues in the laboratory, using electric field meters and charged pieces of metal foil. Even if she set up the measurements by remote control, so that her the charge on her own body would be too far away to have any effect, they would be disrupted by the ambient field of planet Flatcat. Finally, however, she realized that she could improve her equation by rewriting it as follows:

$$E_{\text{outward, on side 1}} + E_{\text{outward, on side 2}} = 4\pi k\sigma.$$

The tricky thing here is that “outward” means a different thing, depending on which side of the foil we're on. On the left side, “outward” means to the left, while on the right side, “outward” is right. A positively charged piece of metal foil has a field that points leftward on the left side, and rightward on its right side, so the two contributions of $2\pi k\sigma$ are both positive, and we get $4\pi k\sigma$. On the other hand, suppose there is a field created by other charges, not by the charged foil, that happens to point to the right. On the right side, this externally created field is in the same direction as the foil's field, but on the left side, the it *reduces* the strength of the leftward field created by the foil. The increase in one term of the equation balances the decrease in the other term. This new version of the equation is thus exactly correct regardless of what externally generated fields are present!

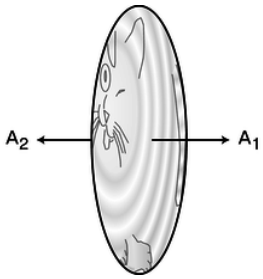
Her next innovation starts by multiplying the equation on both sides by the area, A, of one side of the foil:

$$\left(E_{\text{outward, on side 1}} + E_{\text{outward, on side 2}}\right) A = 4\pi k\sigma A$$

or

$$E_{\text{outward, on side 1}}\, A + E_{\text{outward, on side 2}}\, A = 4\pi kq,$$

where q is the charge of the foil. The reason for this modification is that she can now make the whole thing more attractive by defining a new vector, the area vector $\mathbf{A}$.

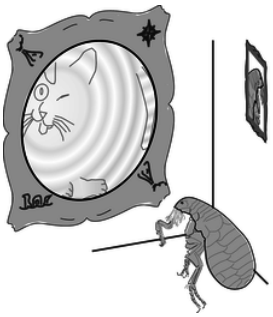


a / The area vector is defined to be perpendicular to the surface, in the outward direction. Its magnitude tells how much the area is.

As shown in figure a, she defines an area vector for side 1 which has magnitude A and points outward from side 1, and an area vector for side 2 which has the same magnitude and points outward from that side, which is in the opposite direction. The dot product of two vectors, $\mathbf{u}\cdot\mathbf{v}$, can be interpreted as $u_{\text{parallel to } v}|\mathbf{v}|$, and she can therefore rewrite her equation as

$$\mathbf{E}_1\cdot\mathbf{A}_1 + \mathbf{E}_2\cdot\mathbf{A}_2 = 4\pi kq.$$

The quantity on the left side of this equation is called the *flux* through the surface, written Φ.



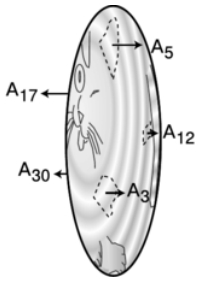
b / Gauss contemplates a map of the known world.

Gauss now writes a grant proposal to her favorite funding agency, the BSGS (Blood-Suckers' Geological Survey), and it is quickly approved. Her audacious plan is to send out exploring teams to chart the electric fields of the whole planet of Flatcat, and thereby determine the total electric charge of the planet. The fleas' world is commonly assumed to be a flat disk, and its size is known to be finite, since the sun passes behind it at sunset and comes back around on the other side at dawn. The most daring part of the plan is that it requires surveying not just the known side of the planet but the uncharted Far Side as well. No flea has ever actually gone around the edge and returned to tell the tale, but Gauss assures them that they won't fall off --- their negatively charged bodies will be attracted to the disk no matter which side they are on.

Of course it is possible that the electric charge of planet Flatcat is not perfectly uniform, but that isn't a problem. As discussed in subsection 10.3.2, as long as one is very close to the surface, the field only depends on the *local* charge density. In fact, a side-benefit of Gauss's program of exploration is that any such local irregularities will be mapped out. But what the newspapers find exciting is the idea that once all the teams get back from their voyages and tabulate their data, the *total* charge of the planet will have been determined for the first time. Each surveying team is assigned to visit a certain list of republics, duchies, city-states, and so on. They are to record each territory's electric field vector, as well as its area. Because the electric field may be nonuniform, the final equation for determining the planet's electric charge will have many terms, not just one for each side of the planet:

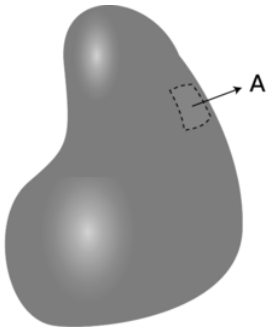
$$\Phi = \sum \mathbf{E}_j \cdot \mathbf{A}_j = 4\pi k q_{total}$$

Gauss herself leads one of the expeditions, which heads due east, toward the distant Tail Kingdom, known only from fables and the occasional account from a caravan of traders. A strange thing happens, however. Gauss embarks from her college town in the wetlands of the Tongue Republic, travels straight east, passes right through the Tail Kingdom, and one day finds herself right back at home, all without ever seeing the edge of the world! What can have happened? All at once she realizes that the world isn't flat.



c / Each part of the surface has its own area vector. Note the differences in lengths of the vectors, corresponding to the unequal areas.

Now what? The surveying teams all return, the data are tabulated, and the result for the total charge of Flatcat is $(1/4\pi k)\sum \mathbf{E}_j \cdot \mathbf{A}_j = 37$ nC (units of nanocoulombs). But the equation was derived under the assumption that Flatcat was a disk. If Flatcat is really round, then the result may be completely wrong. Gauss and two of her grad students go to their favorite bar, and decide to keep on ordering Bloody Marys until they either solve their problems or forget them. One student suggests that perhaps Flatcat really is a disk, but the edges are rounded. Maybe the surveying teams really did flip over the edge at some point, but just didn't realize it. Under this assumption, the original equation will be approximately valid, and 37 nC really is the total charge of Flatcat.



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

A second student, named Newton, suggests that they take seriously the possibility that Flatcat is a sphere. In this scenario, their planet's surface is really curved, but the surveying teams just didn't notice the curvature, since they were close to the surface, and the surface was so big compared to them. They divided up the surface into a patchwork, and each patch was fairly small compared to the whole planet, so each patch was nearly flat. Since the patch is nearly flat, it makes sense to define an area vector that is perpendicular to it. In general, this is how we define the direction of an area vector, as shown in figure d. This only works if the areas are small. For instance, there would be no way to define an area vector for an entire sphere, since "outward" is in more than one direction.

If Flatcat is a sphere, then the inside of the sphere must be vast, and there is no way of knowing exactly how the charge is arranged below the surface. However, the survey teams all found that the electric field was approximately perpendicular to the surface everywhere, and that its strength didn't change very much from one location to another. The simplest explanation is that the charge is all concentrated in one small lump at the center of the sphere. They have no way of knowing if this is really the case, but it's a hypothesis that allows them to see how much their 37 nC result would change if they assumed a different geometry. Making this assumption, Newton performs the following simple computation on a napkin. The field at the surface is related to the charge at the center by

$$|\mathbf{E}| = \frac{kq_{total}}{r^2}, \tag{11.6.1}$$

where r is the radius of Flatcat. The flux is then

$$\Phi = \sum \mathbf{E}_j \cdot \mathbf{A}_j, \tag{11.6.2}$$

and since the $\mathbf{E}_j$ and $\mathbf{A}_j$ vectors are parallel, the dot product equals $|\mathbf{E}_j||\mathbf{A}_j|$, so

$$\Phi = \sum \frac{kq_{total}}{r^2}|\mathbf{A}_j|. \tag{11.6.3}$$

But the field strength is always the same, so we can take it outside the sum, giving

$$\Phi = \frac{kq_{total}}{r^2}\sum|\mathbf{A}_j| \qquad = \frac{kq_{total}}{r^2}A_{total} \ = \frac{kq_{total}}{r^2}4\pi r^2 \qquad = 4\pi k q_{total}.$$

Not only have all the factors of r canceled out, but the result is the same as for a disk!

Everyone is pleasantly surprised by this apparent mathematical coincidence, but is it anything more than that? For instance, what if the charge wasn't concentrated at the center, but instead was evenly distributed throughout Flatcat's interior volume? Newton, however, is familiar with a result called the shell theorem (page 102), which states that the field of a uniformly charged sphere is the same as if all the charge had been concentrated at its center.[10] We now have three different assumptions about the shape of Flatcat and the arrangement of the charges inside it, and all three lead to exactly the *same* mathematical result, $\Phi = 4\pi k q_{total}$. This is starting to look like more than a coincidence. In fact, there is a general mathematical theorem, called Gauss' theorem, which states the following:

For any region of space, the flux through the surface equals $4\pi k q_{in}$, where q_{in} is the total charge in that region.

Don't memorize the factor of 4π in front --- you can rederive it any time you need to, by considering a spherical surface centered on a point charge.

Note that although region and its surface had a definite physical existence in our story --- they are the planet Flatcat and the surface of planet Flatcat --- Gauss' law is true for any region and surface we choose, and in general, the Gaussian surface has no direct physical significance. It's simply a computational tool.

Rather than proving Gauss' theorem and then presenting some examples and applications, it turns out to be easier to show some examples that demonstrate its salient properties. Having understood these properties, the proof becomes quite simple.

self-check:

Suppose we have a negative point charge, whose field points inward, and we pick a Gaussian surface which is a sphere centered on that charge. How does Gauss' theorem apply here?

(answer in the back of the PDF version of the book)

Contributors and Attributions

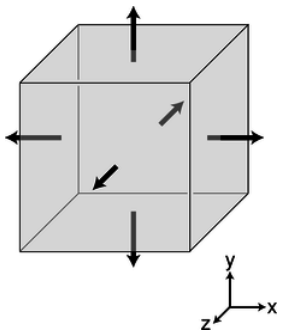
- Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

11.7: Gauss' Law In Differential Form

Gauss' law is a bit spooky. It relates the field on the Gaussian surface to the charges inside the surface. What if the charges have been moving around, and the field at the surface right now is the one that was created by the charges in their previous locations? Gauss' law --- unlike Coulomb's law --- still works in cases like these, but it's far from obvious how the flux and the charges can still stay in agreement if the charges have been moving around.

For this reason, it would be more physically attractive to restate Gauss' law in a different form, so that it related the behavior of the field at one point to the charges that were actually present at that point. This is essentially what we were doing in the fable of the flea named Gauss: the fleas' plan for surveying their planet was essentially one of dividing up the surface of their planet (which they believed was flat) into a patchwork, and then constructing *small* a Gaussian pillbox around each *small* patch. The equation $E_{\perp} = 2\pi k\sigma$ then related a particular property of the *local* electric field to the *local* charge density.

In general, charge distributions need not be confined to a flat surface --- life is three-dimensional --- but the general approach of defining very small Gaussian surfaces is still a good one. Our strategy is to divide up space into tiny cubes, like the one on page 621. Each such cube constitutes a Gaussian surface, which may contain some charge. Again we approximate the field using its six values at the center of each of the six sides. Let the cube extend from x to $x+dx$, from y to $y+dy$, and from y to $y+dy$.



a / A tiny cubical Gaussian surface.

The sides at x and $x+dx$ have area vectors $-dydz\hat{\mathbf{x}}$ and $dydz\hat{\mathbf{x}}$, respectively. The flux through the side at x is $-E_x(x)dydz$, and the flux through the opposite side, at $x+dx$ is $E_x(x+dx)dydz$. The sum of these is $(E_x(x+dx)-E_x(x))dydz$, and if the field was uniform, the flux through these two opposite sides would be zero. It will only be zero if the field's x component changes as a function of x. The difference $E_x(x+dx)-E_x(x)$ can be rewritten as $dE_x = (dE_x)/(dx)dx$, so the contribution to the flux from these two sides of the cube ends up being

$$\frac{dE_x}{dx}dxdydz.$$

Doing the same for the other sides, we end up with a total flux

$$\begin{aligned} d\Phi &= \left(\frac{dE_x}{dx}+\frac{dE_y}{dy}+\frac{dE_z}{dz}\right)dxdydz \\ &= \left(\frac{dE_x}{dx}+\frac{dE_y}{dy}+\frac{dE_z}{dz}\right)dv, \end{aligned}$$

where dv is the volume of the cube. In evaluating each of these three derivatives, we are going to treat the other two variables as constants, to emphasize this we use the partial derivative notation ∂ introduced in chapter 3,

$$d\Phi = \left(\frac{\partial E_x}{\partial x}+\frac{\partial E_y}{\partial y}+\frac{\partial E_z}{\partial z}\right)dv.$$

Using Gauss' law,

$$4\pi kq_{in} = \left(\frac{\partial E_x}{\partial x}+\frac{\partial E_y}{\partial y}+\frac{\partial E_z}{\partial z}\right)dv,$$

and we introduce the notation ρ (Greek letter rho) for the charge per unit volume, giving

$$4\pi k\rho = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}.$$

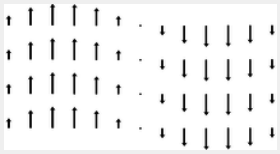
The quantity on the right is called the *divergence* of the electric field, written $\text{div}\mathbf{E}$. Using this notation, we have $\text{div}\mathbf{E} = 4\pi k\rho$.

This equation has all the same physical implications as Gauss' law. After all, we proved Gauss' law by breaking down space into little cubes like this. We therefore refer to it as the differential form of Gauss' law, as opposed to $\Phi = 4\pi k q_{in}$, which is called the integral form.

b / A meter for measuring $\text{div}\mathbf{E}$.

Figure b shows an intuitive way of visualizing the meaning of the divergence. The meter consists of some electrically charged balls connected by springs. If the divergence is positive, then the whole cluster will expand, and it will contract its volume if it is placed at a point where the field has $\text{div}\mathbf{E} < 0$. What if the field is constant? We know based on the definition of the divergence that we should have $\text{div}\mathbf{E} = 0$ in this case, and the meter does give the right result: all the balls will feel a force in the same direction, but they will neither expand nor contract.

Example 36: Divergence of a sine wave



c / Example 36.

▷ Figure c shows an electric field that varies as a sine wave. This is in fact what you'd see in a light wave: light is a wave pattern made of electric and magnetic fields. (The magnetic field would look similar, but would be in a plane perpendicular to the page.) What is the divergence of such a field, and what is the physical significance of the result?

▷ Intuitively, we can see that no matter where we put the div-meter in this field, it will neither expand nor contract. For instance, if we put it at the center of the figure, it will start spinning, but that's it.

Mathematically, let the x axis be to the right and let y be up. The field is of the form

$$\mathbf{E} = (\sin Kx)\,\hat{\mathbf{y}},$$

where the constant K is not to be confused with Coulomb's constant. Since the field has only a y component, the only term in the divergence we need to evaluate is

$$\mathbf{E} = \frac{\partial E_y}{\partial y},$$

but this vanishes, because E_y depends only on x, not y : we treat y as a constant when evaluating the partial derivative $\partial E_y/\partial y$, and the derivative of an expression containing only constants must be zero.

Physically this is a very important result: it tells us that a light wave can exist without any charges along the way to “keep it going.” In other words, light can travel through a vacuum, a region with no particles in it. If this wasn't true, we'd be dead, because the sun's light wouldn't be able to get to us through millions of kilometers of empty space!

Example 37: Electric field of a point charge

The case of a point charge is tricky, because the field behaves badly right on top of the charge, blowing up and becoming discontinuous. At this point, we cannot use the component form of the divergence, since none of the derivatives are well defined. However, a little visualization using the original definition of the divergence will quickly convince us that div E is infinite here, and that makes sense, because the density of charge has to be infinite at a point where there is a zero-size point of charge (finite charge in zero volume).
At all other points, we have

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

where $\hat{\mathbf{r}} = \mathbf{r}/r = (x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}})/r$ is the unit vector pointing radially away from the charge. The field can therefore be written as

$$\begin{aligned}\mathbf{E} &= \frac{kq}{r^3}\hat{\mathbf{r}} \\ &= \frac{kq(x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}})}{\left(x^2+y^2+z^2\right)^{3/2}}.\end{aligned}$$

The three terms in the divergence are all similar, e.g.,

$$\begin{aligned}\frac{\partial E_x}{\partial x} &= kq\frac{\partial}{\partial x}\left[\frac{x}{\left(x^2+y^2+z^2\right)^{3/2}}\right] \\ &= kq\left[\frac{1}{\left(x^2+y^2+z^2\right)^{3/2}} - \frac{3}{2}\,\frac{2x^2}{\left(x^2+y^2+z^2\right)^{5/2}}\right] \\ &= kq\left(r^{-3} - 3x^2r^{-5}\right).\end{aligned}$$

Straightforward algebra shows that adding in the other two terms results in zero, which makes sense, because there is no charge except at the origin.

Gauss' law in differential form lends itself most easily to finding the charge density when we are give the field. What if we want to find the field given the charge density? As demonstrated in the following example, one technique that often works is to guess the general form of the field based on experience or physical intuition, and then try to use Gauss' law to find what specific version of that general form will be a solution.

Example 38: The field inside a uniform sphere of charge

▷ Find the field inside a uniform sphere of charge whose charge density is ρ. (This is very much like finding the gravitational field at some depth below the surface of the earth.)
▷ By symmetry we know that the field must be purely radial (in and out). We guess that the solution might be of the form

$$\mathbf{E} = br^p\hat{\mathbf{r}},$$

where r is the distance from the center, and b and p are constants. A negative value of p would indicate a field that was strongest at the center, while a positive p would give zero field at the center and stronger fields farther out. Physically, we know by symmetry that the field is zero at the center, so we expect p to be positive.
As in the example 37, we rewrite $\hat{\mathbf{r}}$ as $\mathbf{r}/r$, and to simplify the writing we define $n = p - 1$, so

$$\mathbf{E} = br^n\mathbf{r}.$$

Gauss' law in differential form is

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

so we want a field whose divergence is constant. For a field of the form we guessed, the divergence has terms in it like

$$\begin{aligned}\frac{\partial E_x}{\partial x} &= \frac{\partial}{\partial x}(br^n x)\\ &= b\left(nr^{n-1}\frac{\partial r}{\partial x}x + r^n\right)\end{aligned}$$

The partial derivative $\partial r/\partial x$ is easily calculated to be x/r, so

$$\frac{\partial E_x}{\partial x} = b\left(nr^{n-2}x^2 + r^n\right)$$

Adding in similar expressions for the other two terms in the divergence, and making use of $x^2 + y^2 + z^2 = r^2$, we have

$$\mathrm{div}\mathbf{E} = \mathrm{b(n+3)r^n}.$$

This can indeed be constant, but only if n is 0 or -3, i.e., p is 1 or -2. The second solution gives a divergence which is constant and *zero* : this is the solution for the *outside* of the sphere! The first solution, which has the field directly proportional to r, must be the one that applies to the inside of the sphere, which is what we care about right now. Equating the coefficient in front to the one in Gauss' law, the field is

$$\mathbf{E} = \frac{4\pi k\rho}{3}r\,\hat{\mathbf{r}}.$$

The field is zero at the center, and gets stronger and stronger as we approach the surface.

Contributors and Attributions

11.8: Footnotes

1. rhymes with “mouse”
2. Current is a scalar, since the definition $I = dq/dt$ is the derivative of a scalar. However, there is a closely related quantity called the current *density*, **J**, which is a vector, and **J** is in fact the more fundamentally important quantity.
3. As in chapter 2, we use the word “frequency” to mean either f or $\omega = 2\pi f$ when the context makes it clear which is being referred to.
4. I 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.
5. In general, the use of complex number techniques to do an integral could result in a complex number, but that complex number would be a constant, which could be subsumed within the usual constant of integration.
6. A resistor always turns electrical energy into heat. It never turns heat into electrical energy!
7. To many people, the word “radiation” implies nuclear contamination. Actually, the word simply means something that “radiates” outward. Natural sunlight is “radiation.” So is the light from a lightbulb, or the infrared light being emitted by your skin right now.
8. Note 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.
9. no relation to the human mathematician of the same name
10. Newton's human namesake actually proved this for gravity, not electricity, but they're both $1/r^2$ forces, so the proof works equally well in both cases.
11. The math gets messy for the off-axis case. This part of the proof can be completed more easily and transparently using the techniques of section 10.7, and that is exactly we'll do in example 37 on page 631.

11.E: Fields (Exercises)

1. The gap between the electrodes in an automobile engine's spark plug is 0.060 cm. To produce an electric spark in a gasoline-air mixture, an electric field of 3.0×10^6 V/m must be achieved. On starting a car, what minimum voltage must be supplied by the ignition circuit? Assume the field is uniform.(answer check available at lightandmatter.com)

(b) The small size of the gap between the electrodes is inconvenient because it can get blocked easily, and special tools are needed to measure it. Why don't they design spark plugs with a wider gap?

2. (a) As suggested in example 9 on page 573, use approximations to show that the expression given for the electric field approaches kQ/d^2 for large d.
(b) Do the same for the result of example 12 on page 577.

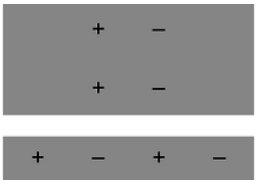
3. Astronomers believe that the mass distribution (mass per unit volume) of some galaxies may be approximated, in spherical coordinates, by $\rho = ae^{-br}$, for $0 \le r \le \infty$, where ρ is the density. Find the total mass.

4. (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.(answer check available at lightandmatter.com)
(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.)

5. Show that the alternative definition of the magnitude of the electric field, \(|**E**|=\tau/D_t\sin\theta\), has units that make sense.

6. Redo the calculation of example 5 on page 566 using a different origin for the coordinate system, and show that you get the same result.

7. The definition of the dipole moment, $\mathbf{D} = \sum q_i \mathbf{r}_i$, involves the vector $\mathbf{r}_i$ stretching from the origin of our coordinate system out to the charge q_i. There are clearly cases where this causes the dipole moment to be dependent on the choice of coordinate system. For instance, if there is only one charge, then we could make the dipole moment equal zero if we chose the origin to be right on top of the charge, or nonzero if we put the origin somewhere else.
(a) Make up a numerical example with two charges of equal magnitude and opposite sign. Compute the dipole moment using two different coordinate systems that are oriented the same way, but differ in the choice of origin. Comment on the result.
(b) Generalize the result of part a to any pair of charges with equal magnitude and opposite sign. This is supposed to be a proof for *any* arrangement of the two charges, so don't assume any numbers.
(c) Generalize further, to n charges.

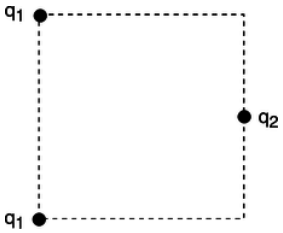


a / Problem 8.

8. Compare the two dipole moments.

9. Find an arrangement of charges that has zero total charge and zero dipole moment, but that will make nonvanishing electric fields.

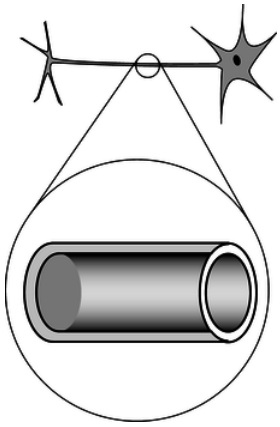
10. As suggested in example 11 on page 575, show that you can get the same result for the on-axis field by differentiating the voltage



b / Problem 11.

11. 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?(answer check available at lightandmatter.com)

12. This is a one-dimensional problem, with everything confined to the x axis. Dipole A consists of a -1.000 C charge at $x = 0.000$ m and a 1.000 C charge at $x = 1.000$ m. Dipole B has a -2.000 C charge at $x = 0.000$ m and a 2.000 C charge at $x = 0.500$ m.
(a) Compare the two dipole moments.
(b) Calculate the field created by dipole A at $x = 10.000$ m, and compare with the field dipole B would make. Comment on the result.(answer check available at lightandmatter.com)



c / Problem 13.

13. In our by-now-familiar neuron, the voltage difference between the inner and outer surfaces of the cell membrane is about $V_{out} - V_{in} = -70$ mV in the resting state, and the thickness of the membrane is about 6.0 nm (i.e., only about a hundred atoms thick). What is the electric field inside the membrane?(answer check available at lightandmatter.com)

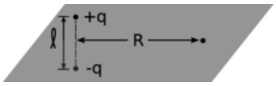
14. A proton is in a region in which the electric field is given by $E = a + bx^3$. If the proton starts at rest at $x_1 = 0$, find its speed, v, when it reaches position x_2. Give your answer in terms of a, b, x_2, and e and m, the charge and mass of the proton.(answer check available at lightandmatter.com)

15. (a) Given that the on-axis field of a dipole at large distances is proportional to D/r^3, show that its voltage varies as D/r^2. (Ignore positive and negative signs and numerical constants of proportionality.)
(b) Write down an exact expression for the voltage of a two-charge dipole at an on-axis point, without assuming that the distance is large compared to the size of the dipole. Your expression will have to contain the actual charges and size of the dipole, not just its dipole moment. Now use approximations to show that, at large distances, this is consistent with your answer to part a.\hwhint{hwhint:dipolev}

16. 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 voltage in and around the hydrogen atom can be approximated using an expression of the form $V = r^{-1}e^{-r}$. (The units come out wrong, because I've left out some constants.) Find the electric field corresponding to this voltage, and comment on its behavior at very large and very small r. (solution in the pdf version of the book)

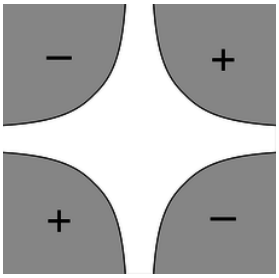
17. A carbon dioxide molecule is structured like O-C-O, with all three atoms along a line. The oxygen atoms grab a little bit of extra negative charge, leaving the carbon positive. The molecule's symmetry, however, means that it has no overall dipole moment, unlike a V-shaped water molecule, for instance. Whereas the voltage of a dipole of magnitude D is proportional to D/r^2 (see problem 15), it turns out that the voltage 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 (cf. problem 9). 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?(answer check available at lightandmatter.com)

18. A hydrogen atom in a particular state has the charge density (charge per unit volume) of the electron cloud given by $\rho = ae^{-br}z^2$, where r is the distance from the proton, and z is the coordinate measured along the z axis. Given that the total charge of the electron cloud must be $-e$, find a in terms of the other variables.



d / Problem 19.

19. A dipole has a midplane, i.e., the plane that cuts through the dipole's center, and is perpendicular to the dipole's axis. Consider a two-charge dipole made of point charges $\pm q$ located at $z = \pm\ell/2$. Use approximations to find the field at a distant point in the midplane, and show that its magnitude comes out to be kD/R^3 (half what it would be at a point on the axis lying an equal distance from the dipole).



e / Problem 20.

20. The figure shows a vacuum chamber surrounded by four metal electrodes shaped like hyperbolas. (Yes, physicists do sometimes ask their university machine shops for things machined in mathematical shapes like this. They have to be made on computer-controlled mills.) We assume that the electrodes extend far into and out of the page along the unseen z axis, so that by symmetry, the electric fields are the same for all z. The problem is therefore effectively two-dimensional. Two of the electrodes are at voltage $+V_\text{o}$, and the other two at $-V_\text{o}$, as shown. The equations of the hyperbolic surfaces are $|xy| = b^2$, where b is a constant. (We can interpret b as giving the locations $x = \pm b$, $y = \pm b$ of the four points on the surfaces that are closest to the central axis.) There is no obvious, pedestrian way to determine the field or voltage in the central vacuum region, but there's a trick that works: with a little mathematical insight, we see that the voltage $V = V_\text{o}b^{-2}xy$ is consistent with all the given information. (Mathematicians could prove that this solution was unique, but a physicist knows it on physical grounds: if there were two different solutions, there would be no physical way for the system to decide which one to do!) (a) Use the techniques of subsection 10.2.2 to find the field in the vacuum region, and (b) sketch the field as a "sea of arrows."(answer check available at lightandmatter.com)

21. (a) A certain region of three-dimensional space has a voltage that varies as $V = br^2$, where r is the distance from the origin. Use the techniques of subsection 10.2.2 to find the field.(answer check available at lightandmatter.com)
(b) Write down another voltage that gives exactly the same field.

22. (a) Example 10 on page 574 gives the field of a charged rod in its midplane. Starting from this result, take the limit as the length of the rod approaches infinity. Note that λ is not changing, so as L gets bigger, the total charge Q increases. \hwans{hwans:estrips}
(b) In the text, I have shown (by several different methods) that the field of an infinite, uniformly charged plane is $2\pi k\sigma$. Now you're going to rederive the same result by a different method. Suppose that it is the $x-y$ plane that is charged, and we want to find the field at the point $(0, 0, z)$. (Since the plane is infinite, there is no loss of generality in assuming $x = 0$ and $y = 0$.) Imagine that we slice the plane into an infinite number of straight strips parallel to the y axis. Each strip has infinitesimal width dx, and extends from x to $x + dx$. The contribution any one of these strips to the field at our point has a magnitude which can be found from part a. By vector addition, prove the desired result for the field of the plane of charge.

f / Problem 23.

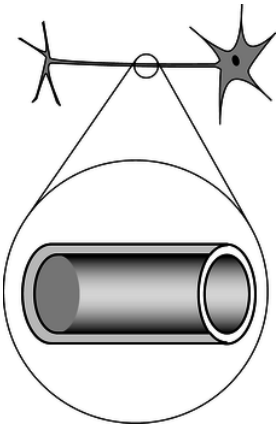
23. Consider the electric field created by a uniformly charged cylindrical surface that extends to infinity in one direction.
(a) Show that the field at the center of the cylinder's mouth is $2\pi k\sigma$, which happens to be the same as the field of an infinite *flat* sheet of charge!
(b) This expression 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?

24. In an electrical storm, the cloud and the ground act like a parallel-plate capacitor, which typically charges up due to frictional electricity in collisions of ice particles in the cold upper atmosphere. Lightning occurs when the magnitude of the electric field builds up to 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.(answer check available at lightandmatter.com)
(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 V/m.

25. (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? \hwhint{hwhint:epointinfty}
[4] (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. (Example 38 shows this using a different technique.)

The calculation in part a seems to show that infinite energy would be required in order to create a charged, pointlike particle. However, there are processes that, for example, create electron-positron pairs, and these processes don't require infinite energy. According to Einstein's famous equation $E = mc^2$, the energy required to create such a pair should only be $2mc^2$, which is finite. One way out of this difficulty is to assume that no particle is really pointlike, and this is in fact the main motivation behind a speculative physical theory called string theory, which posits that charged particles are actually tiny loops, not points.

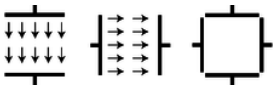
(answer check available at lightandmatter.com)



g / Problem 26.

26. 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.) \hwhint{hwhint:neuronenergy}(answer check available at lightandmatter.com)

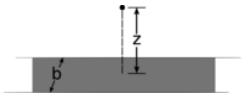
(b) An organism's evolutionary fitness should be better if it needs less energy to operate its nervous system. Based on your answer to part a, what would you expect evolution to do to the dimensions t and r? What other constraints would keep these evolutionary trends from going too far?



h / Problem 27.

27. The figure shows cross-sectional views of two cubical capacitors, and a cross-sectional view of the same two capacitors put together so that their interiors coincide. A capacitor with the plates close together has a nearly uniform electric field between the plates, and almost zero field outside; these capacitors don't have their plates very close together compared to the dimensions of the plates, but for the purposes of this problem, assume that they still have approximately the kind of idealized field pattern shown in the figure. Each capacitor has an interior volume of 1.00 m^3, and is charged up to the point where its internal field is 1.00 V/m.
(a) Calculate the energy stored in the electric field of each capacitor when they are separate. (answer check available at lightandmatter.com)
(b) Calculate the magnitude of the interior field when the two capacitors are put together in the manner shown. Ignore effects arising from the redistribution of each capacitor's charge under the influence of the other capacitor.(answer check available at lightandmatter.com)
(c) Calculate the energy of the put-together configuration. Does assembling them like this release energy, consume energy, or neither?(answer check available at lightandmatter.com)

28. 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.)(answer check available at lightandmatter.com)



i / Problem 29.

29. (a) Show that the field found in example 10 on page 574 reduces to $E = 2k\lambda/R$ in the limit of $L \to \infty$.
(b) 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. (answer check available at lightandmatter.com)
(c) Show that this expression has the correct behavior in the limit where z approaches zero, and also in the limit of $z \gg b$. For the latter, you'll need the result of problem 22a, which is given on page 930.

30. 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.

31. Find the voltage at the edge of a uniformly charged disk. (Define $V = 0$ to be infinitely far from the disk.) (answer check available at lightandmatter.com)\hwhint{hwhint:vedgedisk}

32. Find the energy stored in a capacitor in terms of its capacitance and the voltage difference across it.(answer check available at lightandmatter.com)

33. (a) Find the capacitance of two identical capacitors in series.
(b) Based on this, how would you expect the capacitance of a parallel-plate capacitor to depend on the distance between the plates?

34. (a) Use complex number techniques to rewrite the function $f(t) = 4\sin\omega t + 3\cos\omega t$ in the form $A\sin(\omega t + \delta)$.(answer check available at lightandmatter.com)
(b) Verify the result using the trigonometric identity $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \sin\beta\cos\alpha$.

35. (a) Show that the equation $V_L = LdI/dt$ has the right units.
(b) Verify that RC has units of time.
(c) Verify that L/R has units of time.

36. Find the inductance of two identical inductors in parallel.

37. Calculate the quantity i^i (i.e., find its real and imaginary parts).(answer check available at lightandmatter.com)

38. The wires themselves in a circuit can have resistance, inductance, and capacitance. Would "stray" inductance and capacitance be most important for low-frequency or for high-frequency circuits? For simplicity, assume that the wires act like they're in *series* with an inductor or capacitor.

39. Starting from the relation $V = LdI/dt$ for the voltage difference across an inductor, show that an inductor has an impedance equal to $L\omega$.

40. 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.

41. 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 acclerates 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.

42. If an FM radio tuner consisting of an LRC circuit contains a 1.0 μH inductor, what range of capacitances should the variable capacitor be able to provide?(answer check available at lightandmatter.com)

43. (a) Find the parallel impedance of a 37 kΩ resistor and a 1.0 nF capacitor at $f = 1.0 \times 10^4$ Hz.(answer check available at lightandmatter.com)
(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? (answer check available at lightandmatter.com)

44. A series LRC circuit consists of a 1.000 Ω resistor, a 1.000 F capacitor, and a 1.000 H inductor. (These are not particularly easy values to find on the shelf at Radio Shack!)
(a) Plot its impedance as a point in the complex plane for each of the following frequencies: ω=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?(answer check available at lightandmatter.com)
(c) What is the resonant frequency f_{res} corresponding to your answer in part b?(answer check available at lightandmatter.com)

45. At a frequency ω, a certain series LR circuit has an impedance of $1\ \Omega + (2\ \Omega)i$. Suppose that instead we want to achieve the same impedance using two circuit elements in parallel. What must the elements be?

46. (a) Use Gauss' law to find the fields inside and outside an infinite cylindrical surface with radius b and uniform surface charge density σ.(answer check available at lightandmatter.com)
(b) Show that there is a discontinuity in the electric field equal to $4\pi k\sigma$ between one side of the surface and the other, as there should be (see page 628).
(c) Reexpress your result in terms of the charge per unit length, and compare with the field of a line of charge.
(d) A coaxial cable has two conductors: a central conductor of radius a, and an outer conductor of radius b. These two conductors are separated by an insulator. Although such a cable is normally used for time-varying signals, assume throughout this problem that there is simply a DC voltage between the two conductors. The outer conductor is thin, as in part c. The inner conductor is solid, but, as is always the case with a conductor in electrostatics, the charge is concentrated on the surface. Thus, you can find all the fields in part b by superposing the fields due to each conductor, as found in part c. (Note that on a given length of the cable, the total charge of the inner and outer conductors is zero, so $\lambda_1 = -\lambda_2$, but $\sigma_1 \neq \sigma_2$, since the areas are unequal.) Find the capacitance per unit length of such a cable.(answer check available at lightandmatter.com)

47. In a certain region of space, the electric field is constant (i.e., the vector always has the same magnitude and direction). For simplicity, assume that the field points in the positive x direction. (a) Use Gauss's law to prove that there is no charge in this region of space. This is most easily done by considering a Gaussian surface consisting of a rectangular box, whose edges are parallel to the x, y, and z axes.
(b) If there are no charges in this region of space, what could be making this electric field?

48. (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. (answer check available at lightandmatter.com)

49. Use Gauss' law to find the field inside an infinite cylinder with radius b and uniform charge density ρ. (The external field has the same form as the one in problem 46.)(answer check available at lightandmatter.com)

50. (a) In a certain region of space, the electric field is given by $\mathbf{E} = bx\hat{\mathbf{x}}$, where b is a constant. Find the amount of charge contained within a cubical volume extending from $x = 0$ to $x = a$, from $y = 0$ to $y = a$, and from $z = 0$ to $z = a$.
(b) Repeat for $\mathbf{E} = bx\hat{\mathbf{z}}$.
(c) Repeat for $\mathbf{E} = 13bz\hat{\mathbf{z}} - 7cz\hat{\mathbf{y}}$.
(d) Repeat for $\mathbf{E} = bxz\hat{\mathbf{z}}$.

51. Light is a wave made of electric and magnetic fields, and the fields are perpendicular to the direction of the wave's motion, i.e., they're transverse. An example would be the electric field given by $\mathbf{E} = b\hat{\mathbf{x}}\sin cz$, where b and c are constants. (There would also be an associated magnetic field.) We observe that light can travel through a vacuum, so we expect that this wave pattern is consistent with the nonexistence of any charge in the space it's currently occupying. Use Gauss's law to prove that this is true.

52. This is an alternative approach to problem 49, using a different technique. Suppose that a long cylinder contains a uniform charge density ρ throughout its interior volume.
(a) Use the methods of section 10.7 to find the electric field inside the cylinder. (answer check available at lightandmatter.com)
(b) Extend your solution to the outside region, using the same technique. Once you find the general form of the solution, adjust it so that the inside and outside fields match up at the surface.(answer check available at lightandmatter.com)

53. The purpose of this homework problem is to prove that the divergence is invariant with respect to translations. That is, it doesn't matter where you choose to put the origin of your coordinate system. Suppose we have a field of the form $\mathbf{E} = ax\hat{\mathbf{x}} + by\hat{\mathbf{y}} + cz\hat{\mathbf{z}}$. This is the most general field we need to consider in any small region as far as the divergence is concerned. (The dependence on x, y, and z is linear, but any smooth function looks linear close up. We also don't need to put in terms like $x\hat{\mathbf{y}}$, because they don't contribute to the divergence.) Define a new set of coordinates (u, v, w) related to (x, y, z) by

$$\begin{aligned} x &= u + p \\ y &= v + q \\ z &= w + r, \end{aligned}$$

where p, q, and r are constants. Show that the field's divergence is the same in these new coordinates. Note that $\hat{\mathbf{x}}$ and $\hat{\mathbf{u}}$ are identical, and similarly for the other coordinates.

54. Using a techniques similar to that of problem 53, show that the divergence is rotationally invariant, in the special case of rotations about the z axis. In such a rotation, we rotate to a new (u, v, z) coordinate system, whose axes are rotated by an angle θ with respect to those of the (x, y, z) system. The coordinates are related by

$$\begin{aligned} x &= u\cos\theta + v\sin\theta \\ y &= -u\sin\theta + v\cos\theta \end{aligned}$$

Find how the u and v components the field $\mathbf{E}$ depend on u and v, and show that its divergence is the same in this new coordinate system.

55. An electric field is given in cylindrical coordinates (R, ϕ, z) by $E_R = ce^{-u|z|}R^{-1}\cos^2\phi$, where the notation E_R indicates the component of the field pointing directly away from the axis, and the components in the other directions are zero. (This isn't a completely impossible expression for the field near a radio transmitting antenna.) (a) Find the total charge enclosed within the infinitely long cylinder extending from the axis out to $R = b$. (b) Interpret the R-dependence of your answer to part a.

56. 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 in the pdf version of the book)

57. Find every complex number z such that $z^3 = 1$. (solution in the pdf version of the book)

58. Factor the expression $x^3 - y^3$ into factors of the lowest possible order, using complex coefficients. (Hint: use the result of problem 57.) Then do the same using real coefficients. (solution in the pdf version of the book)

Contributors and Attributions

Benjamin Crowell (Fullerton College).

CHAPTER OVERVIEW

12: Electromagnetism

Contributors and Attributions

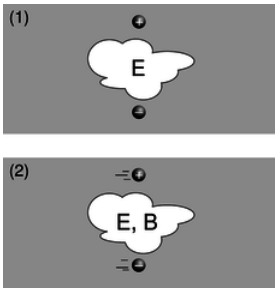
Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

12.1: More About the Magnetic Field

11.1.1 Magnetic forces

In this chapter, I assume you know a few basic ideas about Einstein's theory of relativity, as described in sections 7.1 and 7.2. Unless your typical workday involves rocket ships or particle accelerators, all this relativity stuff might sound like a description of some bizarre futuristic world that is completely hypothetical. There is, however, a relativistic effect that occurs in everyday life, and it is obvious and dramatic: magnetism. Magnetism, as we discussed previously, is an interaction between a moving charge and another moving charge, as opposed to electric forces, which act between any pair of charges, regardless of their motion. Relativistic effects are weak for speeds that are small compared to the speed of light, and the average speed at which electrons drift through a wire is quite low (centimeters per second, typically), so how can relativity be behind an impressive effect like a car being lifted by an electromagnet hanging from a crane? The key is that matter is almost perfectly electrically neutral, and electric forces therefore cancel out almost perfectly. Magnetic forces really aren't very strong, but electric forces are even weaker.

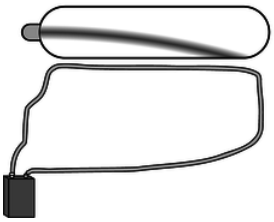
What about the word “relativity” in the name of the theory? It would seem problematic if moving charges interact differently than stationary charges, since motion is a matter of opinion, depending on your frame of reference. Magnetism, however, comes not to destroy relativity but to fulfill it. Magnetic interactions *must* exist according to the theory of relativity. To understand how this can be, consider how time and space behave in relativity. Observers in different frames of reference disagree about the lengths of measuring sticks and the speeds of clocks, but the laws of physics are valid and self-consistent in either frame of reference. Similarly, observers in different frames of reference disagree about what electric and magnetic fields and forces there are, but they agree about concrete physical events.



a / The pair of charged particles, as seen in two different frames of reference.

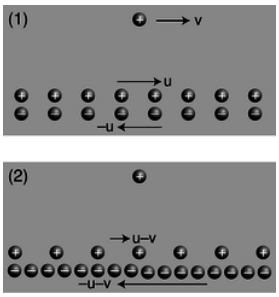
For instance, figure a/1 shows two particles, with opposite charges, which are not moving at a particular moment in time. An observer in this frame of reference says there are electric fields around the particles, and predicts that as time goes on, the particles will begin to accelerate towards one another, eventually colliding. A different observer, a/2, says the particles are moving. This observer also predicts that the particles will collide, but explains their motion in terms of both an electric field, $\mathbf{E}$, and a magnetic field, $\mathbf{B}$. As we'll see shortly, the magnetic field is *required* in order to maintain consistency between the predictions made in the two frames of reference.

To see how this really works out, we need to find a nice simple example that is easy to calculate. An example like figure a is *not* easy to handle, because in the second frame of reference, the moving charges create fields that change over time at any given location. Examples like figure b are easier, because there is a steady flow of charges, and all the fields stay the same over time.[1] What is remarkable about this demonstration is that there can be no electric fields acting on the electron beam at all, since the total charge density throughout the wire is zero. Unlike figure a/2, figure b is purely magnetic.



b / A large current is created by shorting across the leads of the battery. The moving charges in the wire attract the moving charges in the electron beam, causing the electrons to curve.

To see why this must occur based on relativity, we make the mathematically idealized model shown in figure c.



c / A charged particle and a current, seen in two different frames of reference. The second frame is moving at velocity v with respect to the first frame, so all the velocities have v subtracted from them. (As discussed in the main text, this is only approximately correct.)

The charge by itself is like one of the electrons in the vacuum tube beam of figure b, and a pair of moving, infinitely long line charges has been substituted for the wire. The electrons in a real wire are in rapid thermal motion, and the current is created only by a slow drift superimposed on this chaos. A second deviation from reality is that in the real experiment, the protons are at rest with respect to the tabletop, and it is the electrons that are in motion, but in c/1 we have the positive charges moving in one direction and the negative ones moving the other way. If we wanted to, we could construct a third frame of reference in which the positive charges were at rest, which would be more like the frame of reference fixed to the tabletop in the real demonstration. However, as we'll see shortly, frames c/1 and c/2 are designed so that they are particularly easy to analyze. It's important to note that even though the two line charges are moving in opposite directions, their currents don't cancel. A negative charge moving to the left makes a current that goes to the right, so in frame c/1, the total current is twice that contributed by either line charge.

Frame 1 is easy to analyze because the charge densities of the two line charges cancel out, and the electric field experienced by the lone charge is therefore zero:

$$\mathbf{E}_1 = 0$$

In frame 1, any force experienced by the lone charge must therefore be attributed solely to magnetism.

Frame 2 shows what we'd see if we were observing all this from a frame of reference moving along with the lone charge. Why don't the charge densities also cancel in this frame? Here's where the relativity comes in. Relativity tells us that moving objects appear contracted to an observer who is not moving along with them. Both line charges are in motion in both frames of reference, but in frame 1, the line charges were moving at equal speeds, so their contractions were equal, and their charge densities canceled out. In frame 2, however, their speeds are unequal. The positive charges are moving more slowly than in frame 1, so in frame 2 they are less contracted. The negative charges are moving more quickly, so their contraction is greater now. Since the charge densities don't cancel, there is an electric field in frame 2, which points into the wire, attracting the lone charge. Furthermore, the attraction felt by the lone charge must be purely electrical, since the lone charge is at rest in this frame of reference, and magnetic effects occur only between moving charges and other moving charges.[2]

To summarize, frame 1 displays a purely magnetic attraction, while in frame 2 it is purely electrical.

Now we can calculate the force in frame 2, and equating it to the force in frame 1, we can find out how much magnetic force occurs. To keep the math simple, and to keep from assuming too much about your knowledge of relativity, we're going to carry out this whole calculation in the approximation where all the speeds are fairly small compared to the speed of light.[3] For instance, if we find an expression such as $(v/c)^2 + (v/c)^4$, we will assume that the fourth-order term is negligible by comparison. This is known as a calculation "to leading order in v/c." In fact, I've already used the leading-order approximation twice without saying so! The first time I used it implicitly was in figure c, where I assumed that the velocities of the two line charges were $u-v$ and $-u-v$. Relativistic velocities don't just combine by simple addition and subtraction like this, but this is an effect we can ignore in the present approximation. The second sleight of hand occurred when I stated that we could equate the forces in the two frames of reference. Force, like time and distance, is distorted relativistically when we change from one frame of reference to another. Again, however, this is an effect that we can ignore to the desired level of approximation.

Let $\pm\lambda$ be the charge per unit length of each line charge without relativistic contraction, i.e., in the frame moving with that line charge. Using the approximation $\gamma = (1 - v^2/c^2)^{-1/2} \approx 1 + v^2/2c^2$ for $v \ll c$, the total charge per unit length in frame 2 is

$$\begin{aligned}\lambda_{total,\,2} &\approx \lambda\left[1+\frac{(u-v)^2}{2c^2}\right]-\lambda\left[1+\frac{(-u-v)^2}{2c^2}\right]\\ &= \frac{-2\lambda uv}{c^2}.\end{aligned}$$

Let R be the distance from the line charge to the lone charge. Applying Gauss' law to a cylinder of radius R centered on the line charge, we find that the magnitude of the electric field experienced by the lone charge in frame 2 is

$$E=\frac{4k\lambda uv}{c^2R},$$

and the force acting on the lone charge q is

$$F=\frac{4k\lambda quv}{c^2R}.$$

In frame 1, the current is $I=2\lambda_1 u$ (see homework problem 5), which we can approximate as $I=2\lambda u$, since the current, unlike $\lambda_{total,\,2}$, doesn't vanish completely without the relativistic effect. The magnetic force on the lone charge q due to the current I is

$$F=\frac{2kIqv}{c^2R}.$$

Discussion Question

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

11.1.2 The magnetic field

Definition in terms of the force on a moving particle

With electricity, it turned out to be useful to define an electric field rather than always working in terms of electric forces. Likewise, we want to define a magnetic field, $\mathbf{B}$. Let's look at the result of the preceding subsection for insight. The equation

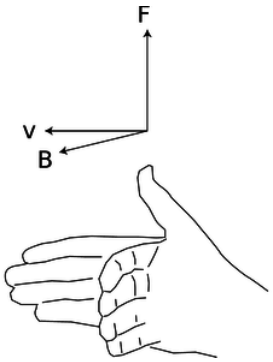
$$F=\frac{2kIqv}{c^2R}$$

shows that when we put a moving charge near other moving charges, there is an extra magnetic force on it, in addition to any electric forces that may exist. Equations for electric forces always have a factor of k in front --- the Coulomb constant k is called the coupling constant for electric forces. Since magnetic effects are relativistic in origin, they end up having a factor of k/c^2 instead of just k. In a world where the speed of light was infinite, relativistic effects, including magnetism, would be absent, and the coupling constant for magnetism would be zero. A cute feature of the metric system is that we have $k/c^2=10^{-7}\ \mathrm{N\cdot s^2/C^2}$ exactly, as a matter of definition.

Naively, we could try to work by analogy with the electric field, and define the magnetic field as the magnetic force per unit charge. However, if we think of the lone charge in our example as the test charge, we'll find that this approach fails, because the force depends not just on the test particle's charge, but on its velocity, v, as well. Although we only carried out calculations for the case where the particle was moving parallel to the wire, in general this velocity is a vector, **v**, in three dimensions. We can also anticipate that the magnetic field will be a vector. The electric and gravitational fields are vectors, and we expect intuitively based on our experience with magnetic compasses that a magnetic field has a particular direction in space. Furthermore, reversing the current I in our example would have reversed the force, which would only make sense if the magnetic field had a direction in space that could be reversed. Summarizing, we think there must be a magnetic field vector **B**, and the force on a test particle moving through a magnetic field is proportional both to the **B** vector and to the particle's own **v** vector. In other words, the magnetic force vector **F** is found by some sort of vector multiplication of the vectors **v** and **B**. As proved on page 912, however, there is only one physically useful way of defining such a multiplication, which is the cross product.

We therefore define the magnetic field vector, **B**, as the vector that determines the force on a charged particle according to the following rule:

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B} [\text{definition of the magnetic field}]$$



d / The right-hand relationship between the velocity of a positively charged particle, the magnetic field through which it is moving, and the magnetic force on it.

From this definition, we see that the magnetic field's units are $\mathrm{N \cdot s/C \cdot m}$, which are usually abbreviated as teslas, $1\ \mathrm{T} = 1\ \mathrm{N \cdot s/C \cdot m}$. The definition implies a right-hand-rule relationship among the vectors, figure d, if the charge q is positive, and the opposite handedness if it is negative.

e / The unit of magnetic field, the tesla, is named after Serbian-American inventor Nikola Tesla.

This is not just a definition but a bold prediction! Is it really true that for any point in space, we can always find a vector **B** that successfully predicts the force on any passing particle, regardless of its charge and velocity vector? Yes --- it's not obvious that it can be done, but experiments verify that it can. How? Well for example, the cross product of parallel vectors is zero, so we can try particles moving in various directions, and hunt for the direction that produces zero force; the **B** vector lies along that line, in either the same direction the particle was moving, or the opposite one. We can then go back to our data from one of the other cases, where the force was nonzero, and use it to choose between these two directions and find the magnitude of the **B** vector. We could then verify that this vector gave correct force predictions in a variety of other cases.

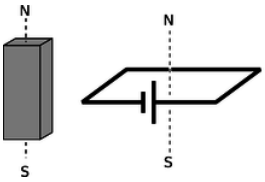
Even with this empirical reassurance, the meaning of this equation is not intuitively transparent, nor is it practical in most cases to measure a magnetic field this way. For these reasons, let's look at an alternative method of defining the magnetic field which, although not as fundamental or mathematically simple, may be more appealing.

Definition in terms of the torque on a dipole

A compass needle in a magnetic field experiences a torque which tends to align it with the field. This is just like the behavior of an electric dipole in an electric field, so we consider the compass needle to be a *magnetic dipole*. In subsection 10.1.3 on page 567, we gave an alternative definition of the electric field in terms of the torque on an electric dipole.

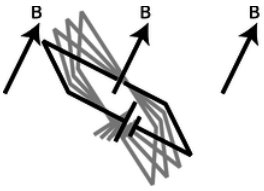
To define the strength of a magnetic field, however, we need some way of defining the strength of a test dipole, i.e., we need a definition of the magnetic dipole moment. We could use an iron permanent magnet constructed according to certain specifications, but such an object is really an extremely complex system consisting of many iron atoms, only some of which are aligned with each

other. A more fundamental standard dipole is a square current loop. This could be little resistive circuit consisting of a square of wire shorting across a battery, f.



f / A standard dipole made from a square loop of wire shorting across a battery. It acts very much like a bar magnet, but its strength is more easily quantified.

Applying $\mathbf{F} = \mathbf{v} \times \mathbf{B}$, we find that such a loop, when placed in a magnetic field, g, experiences a torque that tends to align plane so that its interior "face" points in a certain direction.



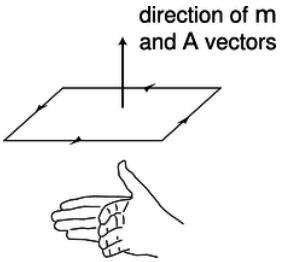
g / A dipole tends to align itself to the surrounding magnetic field.

Since the loop is symmetric, it doesn't care if we rotate it like a wheel without changing the plane in which it lies. It is this preferred facing direction that we will end up using as our alternative definition of the magnetic field.

If the loop is out of alignment with the field, the torque on it is proportional to the amount of current, and also to the interior area of the loop. The proportionality to current makes sense, since magnetic forces are interactions between moving charges, and current is a measure of the motion of charge. The proportionality to the loop's area is also not hard to understand, because increasing the length of the sides of the square increases both the amount of charge contained in this circular "river" and the amount of leverage supplied for making torque. Two separate physical reasons for a proportionality to length result in an overall proportionality to length squared, which is the same as the area of the loop. For these reasons, we define the magnetic dipole moment of a square current loop as

$$\mathbf{m} = I\mathbf{A},$$

where the direction of the vectors is defined as shown in figure h.



h / The **m** and **A** vectors.

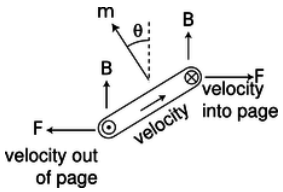
We can now give an alternative definition of the magnetic field:

The magnetic field vector, **B**, at any location in space is defined by observing the torque exerted on a magnetic test dipole $\mathbf{m}_t$ consisting of a square current loop. The field's magnitude is

$$|\mathbf{B}| = \frac{\tau}{|\mathbf{m}_t| \sin\theta},$$

where θ is the angle between the dipole vector and the field. This is equivalent to the vector cross product $\boldsymbol{\tau} = \mathbf{m}_t \times \mathbf{B}$.

Let's show that this is consistent with the previous definition, using the geometry shown in figure i.



i / The torque on a current loop in a magnetic field. The current comes out of the page, goes across, goes back into the page, and then back across the other way in the hidden side of the loop.

The velocity vector that point in and out of the page are shown using the convention defined in figure j.

⊙ out of the page

⊗ into the page

j / A vector coming out of the page is shown with the tip of an arrowhead. A vector going into the page is represented using the tailfeathers of the arrow.

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 $I = \lambda v$, and $m = h^2 \lambda v$. 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

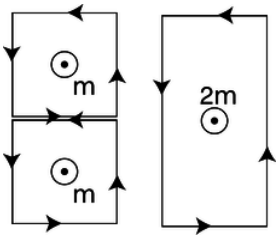
$$\begin{aligned} |\boldsymbol{\tau}| &= 2\frac{h}{2}|\mathbf{F}|\sin\theta \\ &= h\ qvB\ \sin\theta, \end{aligned}$$

and substituting $q = \lambda h$ and $v = m/h^2\lambda$, we have

$$\begin{aligned} |\boldsymbol{\tau}| &= h\ \lambda h\ \frac{m}{h^2\lambda} B\sin\theta \\ &= mB\sin\theta, \end{aligned}$$

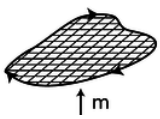
which is consistent with the second definition of the field.

It undoubtedly seems artificial to you that we have discussed dipoles only in the form of a square loop of current. A permanent magnet, for example, is made out of atomic dipoles, and atoms aren't square! However, it turns out that the shape doesn't matter. To see why this is so, consider the additive property of areas and dipole moments, shown in figure k.



k / Dipole vectors can be added.

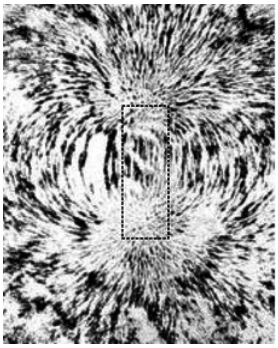
Each of the square dipoles has a dipole moment that points out of the page. When they are placed side by side, the currents in the adjoining sides cancel out, so they are equivalent to a single rectangular loop with twice the area. We can break down any irregular shape into little squares, as shown in figure l, so the dipole moment of any planar current loop can be calculated based on its area, regardless of its shape.



l / An irregular loop can be broken up into little squares.

Example 1: The magnetic dipole moment of an atom

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 $(10^{-10}\ \mathrm{m})^2 = 10^{-20}\ \mathrm{m}^2$. 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 $m \sim 10^{-23}\ \mathrm{A{\cdot}m^2}$.
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 $10^{3}\ \mathrm{A{\cdot}m^2}$, 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.



m / The magnetic field pattern around a bar magnet is created by the superposition of the dipole fields of the individual iron atoms. Roughly speaking, it looks like the field of one big dipole, especially farther away from the magnet. Closer in, however, you can see a hint of the magnet's rectangular shape. The picture was made by placing iron filings on a piece of paper, and then bringing a magnet up underneath.

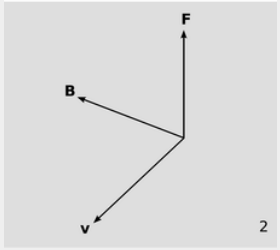
Discussion Questions

◊ The physical situation shown in figure c on page 648 was analyzed entirely in terms of forces. Now let's go back and think about it in terms of fields. The charge by itself up above the wire is like a test charge, being used to determine the magnetic and electric fields created by the wire. In figures c/1 and c/2, are there fields that are purely electric or purely magnetic? Are there fields that are a mixture of $\mathbf{E}$ and $\mathbf{B}$? How does this compare with the forces?

◊ Continuing the analysis begun in discussion question A, can we come up with a scenario involving some charged particles such that the fields are purely magnetic in one frame of reference but a mixture of $\mathbf{E}$ and $\mathbf{B}$ in another frame? How about an example where the fields are purely electric in one frame, but mixed in another? Or an example where the fields are purely electric in one frame, but purely magnetic in another?

11.1.3 Some applications

Example 2: Magnetic levitation



n / Example 2.

In figure n, 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 through the wire. The electrons moving through the wire feel a force from the magnetic field made by the permanent magnet, and this force levitates the wire.

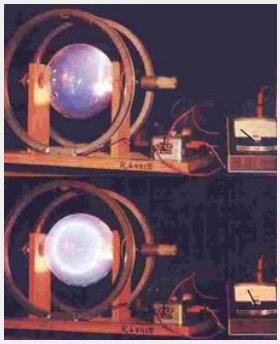
From the photo, it's possible to find the direction of the magnetic field made by the permanent magnet. The electrons in the copper wire are negatively charged, so they flow from the negative (flat) terminal of the battery to the positive terminal (the one with the bump, in front). As the electrons pass 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. By the right-hand rule (figure d on page 651), the field must be toward the battery.

Example 3: Nervous-system effects during an MRI scan<</th>

During an MRI scan of the head, the patient's nervous system is exposed to intense magnetic fields, and there are ions moving around in the nerves. The resulting forces on the ions can cause symptoms such as vertigo.

Example 4: A circular orbit

The magnetic force is always perpendicular to the motion of the particle, so it can never do any work, and a charged particle moving through a magnetic field does not experience any change in its kinetic energy: its velocity vector can change its direction, but not its magnitude. If the velocity vector is initially perpendicular to the field, then the curve of its motion will remain in the plane perpendicular to the field, so the magnitude of the magnetic force on it will stay the same. When an object experiences a force with constant magnitude, which is always perpendicular to the direction of its motion, the result is that it travels in a circle.



o / Magnetic forces cause a beam of electrons to move in a circle.

Figure o shows a beam of electrons in a spherical vacuum tube. In the top photo, the beam is emitted near the right side of the tube, and travels straight up. In the bottom photo, a magnetic field has been imposed by an electromagnet surrounding the vacuum tube; the ammeter on the right shows that the current through the electromagnet is now nonzero. We observe that the beam is bent into a circle.

self-check:

Infer the direction of the magnetic field. Don't forget that the beam is made of electrons, which are negatively charged!

(answer in the back of the PDF version of the book)

Homework problem 12 is a quantitative analysis of circular orbits.

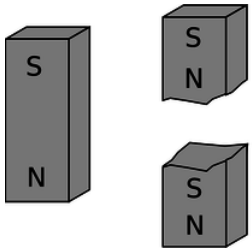
Example 5: A velocity filter

Suppose you see the electron beam in figure o, and you want to determine how fast the electrons are going. You certainly can't do it with a stopwatch! Physicists may also encounter situations where they have a beam of unknown charged particles, and they don't even know their charges. This happened, for instance, when alpha and beta radiation were discovered. One solution to this problem relies on the fact that the force experienced by a charged particle in an electric field, $\mathbf{F}_E = q\mathbf{E}$, is independent of its velocity, but the force due to a magnetic field, $\mathbf{F}_B = q\mathbf{v} \times \mathbf{B}$, isn't. One can send a beam of charged particles through a space containing both an electric and a magnetic field, setting up the fields so that the two forces will cancel out perfectly for a certain velocity. Note that since both forces are proportional to the charge of the particles, the cancellation is independent of charge. Such a *velocity filter* can be used either to determine the velocity of an unknown beam or particles, or to select from a beam of particles only those having velocities within a certain desired range. Homework problem 7 is an analysis of this application.

11.1.4 No magnetic monopoles

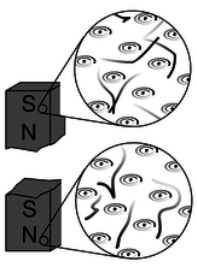
If you could play with a handful of electric dipoles and a handful of bar magnets, they would appear very similar. For instance, a pair of bar magnets wants to align themselves head-to-tail, and a pair of electric dipoles does the same thing. (It is unfortunately not that easy to make a permanent electric dipole that can be handled like this, since the charge tends to leak.)

You would eventually notice an important difference between the two types of objects, however. The electric dipoles can be broken apart to form isolated positive charges and negative charges. The two-ended device can be broken into parts that are not two-ended. But if you break a bar magnet in half, p, you will find that you have simply made two smaller two-ended objects.



p / You can't isolate the poles of a magnet by breaking it in half.

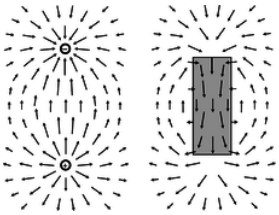
The reason for this behavior is not hard to divine from our microscopic picture of permanent iron magnets. An electric dipole has extra positive “stuff” concentrated in one end and extra negative in the other. The bar magnet, on the other hand, gets its magnetic properties not from an imbalance of magnetic “stuff” at the two ends but from the orientation of the rotation of its electrons. One end is the one from which we could look down the axis and see the electrons rotating clockwise, and the other is the one from which they would appear to go counterclockwise. There is no difference between the “stuff” in one end of the magnet and the other, q.



q / A magnetic dipole is made out of other dipoles, not out of monopoles.

Nobody has ever succeeded in isolating a single magnetic pole. In technical language, we say that magnetic *monopoles* not seem to exist. Electric monopoles *do* exist --- that's what charges are.

Electric and magnetic forces seem similar in many ways. Both act at a distance, both can be either attractive or repulsive, and both are intimately related to the property of matter called charge. (Recall that magnetism is an interaction between moving charges.) Physicists's aesthetic senses have been offended for a long time because this seeming symmetry is broken by the existence of electric monopoles and the absence of magnetic ones. Perhaps some exotic form of matter exists, composed of particles that are magnetic monopoles. If such particles could be found in cosmic rays or moon rocks, it would be evidence that the apparent asymmetry was only an asymmetry in the composition of the universe, not in the laws of physics. For these admittedly subjective reasons, there have been several searches for magnetic monopoles. Experiments have been performed, with negative results, to look for magnetic monopoles embedded in ordinary matter. Soviet physicists in the 1960's made exciting claims that they had created and detected magnetic monopoles in particle accelerators, but there was no success in attempts to reproduce the results there or at other accelerators. The most recent search for magnetic monopoles, done by reanalyzing data from the search for the top quark at Fermilab, turned up no candidates, which shows that either monopoles don't exist in nature or they are extremely massive and thus hard to create in accelerators.



r / Magnetic fields have no sources or sinks.

The nonexistence of magnetic monopoles means that unlike an electric field, a magnetic one, can never have sources or sinks. The magnetic field vectors lead in paths that loop back on themselves, without ever converging or diverging at a point, as in the fields shown in figure r. Gauss' law for magnetism is therefore much simpler than Gauss' law for electric fields:

$$\Phi_B = \sum \mathbf{B}_j \cdot \mathbf{A}_j = 0$$

The magnetic flux through any closed surface is zero.

self-check:

Draw a Gaussian surface on the electric dipole field of figure r that has nonzero electric flux through it, and then draw a similar surface on the magnetic field pattern. What happens?

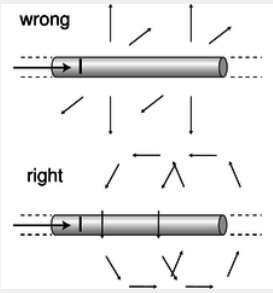
(answer in the back of the PDF version of the book)

Example 6: The field of a wire

▷ On page 650, we showed that a long, straight wire carrying current I exerts a magnetic force

$$F = \frac{2kIqv}{c^2R}$$

on a particle with charge q moving parallel to the wire with velocity v. What, then, is the magnetic field of the wire?



s / Example 6.

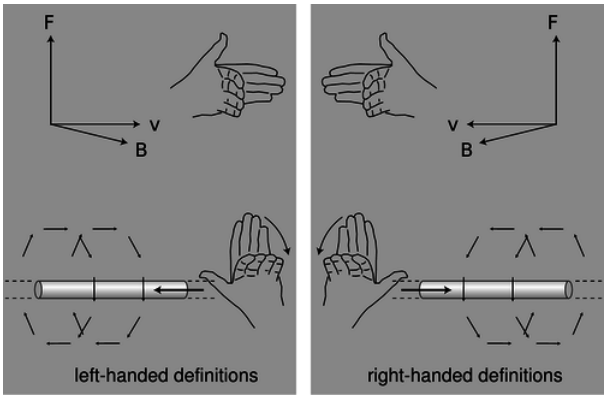
▷ Comparing the equation above to the first definition of the magnetic field, $\mathbf{F} = \mathbf{v} \times \mathbf{B}$, it appears that the magnetic field is one that falls off like $1/R$, where R is the distance from the wire. However, it's not so easy to determine the direction of the field vector. There are two other axes along which the particle could have been moving, and the brute-force method would be to carry out relativistic calculations for these cases as well. Although this would probably be enough information to determine the field, we don't want to do that much work.

Instead, let's consider what the possibilities are. The field can't be parallel to the wire, because a cross product vanishes when the two vectors are parallel, and yet we know from the case we analyzed that the force doesn't vanish when the particle is moving parallel to the wire. The other two possibilities that are consistent with the symmetry of the problem are shown in figure s. One is like a bottle brush, and the other is like a spool of thread. The bottle brush pattern, however, violates Gauss' law for magnetism. If we made a cylindrical Gaussian surface with its axis coinciding with the wire, the flux through it would *not* be zero. We therefore conclude that the spool-of-thread pattern is the correct one.[4] Since the particle in our example was moving perpendicular to the field, we have $|F| = |q||v||B|$, so

$$\begin{aligned} |B| &= \frac{|F|}{|q||v|} \\ &= \frac{2kI}{c^2R} \end{aligned}$$

11.1.5 Symmetry and handedness

Imagine that you establish radio contact with an alien on another planet. Neither of you even knows where the other one's planet is, and you aren't able to establish any landmarks that you both recognize. You manage to learn quite a bit of each other's languages, but you're stumped when you try to establish the definitions of left and right (or, equivalently, clockwise and counterclockwise). Is there any way to do it?



t / Left-handed and right-handed definitions.

If there was any way to do it without reference to external landmarks, then it would imply that the laws of physics themselves were asymmetric, which would be strange. Why should they distinguish left from right? The gravitational field pattern surrounding a star

or planet looks the same in a mirror, and the same goes for electric fields. However, the magnetic field patterns shown in figure s seems to violate this principle. Could you use these patterns to explain left and right to the alien? No. If you look back at the definition of the magnetic field, it also contains a reference to handedness: the direction of the vector cross product. The aliens might have reversed their definition of the magnetic field, in which case their drawings of field patterns would look like mirror images of ours, as in the left panel of figure t.

Until the middle of the twentieth century, physicists assumed that any reasonable set of physical laws would have to have this kind of symmetry between left and right.

u / In this scene from Swan Lake, the the choreography has a symmetry with respect to left and right.

An asymmetry would be grotesque. Whatever their aesthetic feelings, they had to change their opinions about reality when experiments by C.S. Wu et al. showed that the weak nuclear force violates right-left symmetry!

v / C.S. Wu

It is still a mystery why right-left symmetry is observed so scrupulously in general, but is violated by one particular type of physical process.

Contributors

Benjamin Crowell (Fullerton College).

12.2: Magnetic Fields by Superposition

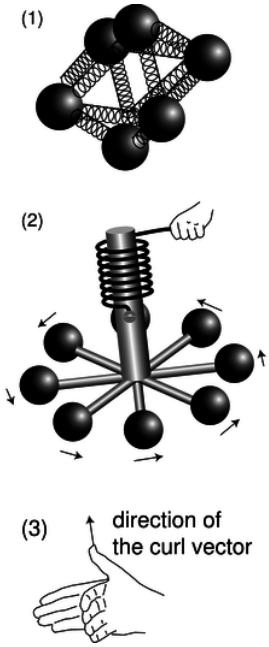
Contributors and Attributions

12.3: Magnetic Fields by Ampère's Law

Contributors and Attributions

11.4.1 The curl operator

The differential form of Gauss' law is more physically satisfying than the integral form, because it relates the charges that are present at some point to the properties of the electric field *at the same point*. Likewise, it would be more attractive to have a differential version of Ampère's law that would relate the currents to the magnetic field at a single point.



a / The div-meter, 1, and the curl-meter, 2 and 3.

Intuitively, the divergence was based on the idea of the div-meter, a/1. The corresponding device for measuring the curliness of a field is the curl-meter, a/2. If the field is curly, then the torques on the charges will not cancel out, and the wheel will twist against the resistance of the spring. If your intuition tells you that the curlmeter will never do anything at all, then your intuition is doing a beautiful job on static fields; for nonstatic fields, however, it is perfectly possible to get a curly electric field.

Gauss' law in differential form relates $\text{div}\,\mathbf{E}$, a scalar, to the charge density, another scalar. Ampère's law, however, deals with directions in space: if we reverse the directions of the currents, it makes a difference. We therefore expect that the differential form of Ampère's law will have vectors on both sides of the equal sign, and we should be thinking of the curl-meter's result as a vector. First we find the orientation of the curl-meter that gives the strongest torque, and then we define the direction of the curl vector using the right-hand rule shown in figure a/3.

To convert the div-meter concept to a mathematical definition, we found the infinitesimal flux, $d\Phi$ through a tiny cubical Gaussian surface containing a volume dv. By analogy, we imagine a tiny square Ampèrian surface with infinitesimal area $d\mathbf{A}$. We assume this surface has been oriented in order to get the maximum circulation. The area vector $d\mathbf{A}$ will then be in the same direction as the one defined in figure a/3. Ampère's law is

$$d\Gamma = \frac{4\pi k}{c^2}\, dI_{through}\,.$$

We define a current density per unit area, $\mathbf{j}$, which is a vector pointing in the direction of the current and having magnitude $\mathbf{j} = dI/|d\mathbf{A}|$. In terms of this quantity, we have

$$d\Gamma = \frac{4\pi k}{c^2}\, j|\mathbf{j}|\,|d\mathbf{A}|$$

$$\frac{d\Gamma}{|d\mathbf{A}|} = \frac{4\pi k}{c^2}\,|\mathbf{j}|$$

With this motivation, we define the magnitude of the curl as

$$|\text{curl } \mathbf{B}| = \frac{d\Gamma}{|d\mathbf{A}|}.$$

Note that the curl, just like a derivative, has a differential divided by another differential. In terms of this definition, we find Amp To convert the div-meter concept to a mathematical definition, we found the infinitesimal flux, $d\Phi$ through a tiny cubical Gaussian surface containing a volume dv. By analogy, we imagine a tiny square Ampèrian surface with infinitesimal area $d\mathbf{A}$. We assume this surface has been oriented in order to get the maximum circulation. The area vector $d\mathbf{A}$ will then be in the same direction as the one defined in figure a/3. Ampère's law is {e}re's law in differential form:

$$\text{curl } \mathbf{B} = \frac{4\pi k}{c^2}\mathbf{j}$$

The complete set of Maxwell's equations in differential form is collected on page 914.

11.4.2 Properties of the curl operator

Contributors and Attributions

12.5: Induced Electric Fields

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

12.6: Maxwell's Equations

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

12.7: 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 592, 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.

Contributors and Attributions

Benjamin Crowell (Fullerton College).

12.8: Footnotes

1. For a more practical demonstration of this effect, you can put an ordinary magnet near a computer monitor. The picture will be distorted. Make sure that the monitor has a demagnetizing ("degaussing") button, however! Otherwise you may permanently damage it. Don't use a television tube, because TV tubes don't have demagnetizing buttons.
2. One could object that this is circular reasoning, since the whole purpose of this argument is to prove from first principles that magnetic effects follow from the theory of relativity. Could there be some extra interaction which occurs between a moving charge and *any* other charge, regardless of whether the other charge is moving or not? We can argue, however, that such a theory would lack self-consistency, since we have to define the electric field somehow, and the only way to define it is in terms of F/q, where F is the force on a test charge q which is at rest. In other words, we'd have to say that there was some extra contribution to the *electric* field if the charge making it was in motion. This would, however, violate Gauss' law, and Gauss' law is amply supported by experiment, even when the sources of the electric field are moving. It would also violate the time-reversal symmetry of the laws of physics.
3. The reader who wants to see the full relativistic treatment is referred to E.M. Purcell, *Electricity and Magnetism*, McGraw Hill, 1985, p. 174.
4. Strictly speaking, there is a hole in this logic, since I've only ruled out a field that is purely along one of these three perpendicular directions. What if it has components along more than one of them? A little more work is required to eliminate these mixed possibilities. For example, we can rule out a field with a nonzero component parallel to the wire based on the following symmetry argument. Suppose a charged particle is moving in the plane of the page directly toward the wire. If the field had a component parallel to the wire, then the particle would feel a force into or out of the page, but such a force is impossible based on symmetry, since the whole arrangement is symmetric with respect to mirror-reflection across the plane of the page.
5. If you've taken a course in differential equations, this won't seem like a very surprising assertion. The differential form of Gauss' law is a differential equation, and by giving the value of the field in the midplane, we've specified a boundary condition for the differential equation. Normally if you specify the boundary conditions, then there is a unique solution to the differential equation. In this particular case, it turns out that to ensure uniqueness, we also need to demand that the solution satisfy the differential form of Ampère's law, which is discussed in section 11.4.
6. If you didn't read this optional subsection, don't worry, because the point is that we need to try a whole new approach anyway.
7. Note that the magnetic field never does work on a charged particle, because its force is perpendicular to the motion; the electric power is actually coming from the mechanical work that had to be done to spin the coil. Spinning the coil is more difficult due to the presence of the magnet.
8. If 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.
9. The 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.
10. They can't be blamed too much for this. As a consequence of Faraday's work, it soon became apparent that light was an electromagnetic wave, and to reconcile this with the relative nature of motion requires Einstein's version of relativity, with all its subversive ideas how space and time are not absolute.
11. One way to prove this rigorously is that in a frame of reference where the particle is at rest, it has an electric field that surrounds it on all sides. If the particle has been moving with constant velocity for a long time, then this is just an ordinary Coulomb's-law field, extending off to very large distances, since disturbances in the field ripple outward at the speed of light. In a frame where the particle is moving, this pure electric field is experienced instead as a combination of an electric field and a magnetic field, so the magnetic field must exist throughout the same vast region of space.
12. Even if the fields can't be parallel to the direction of propagation, one might wonder whether they could form some angle other than 90 degrees with it. No. One proof is given on page 703. A alternative argument, which is simpler but more esoteric, is that if there was such a pattern, then there would be some other frame of reference in which it would look like figure i.
13. A young Einstein worried about what would happen if you rode a motorcycle alongside a light wave, traveling at the speed of light. Would the light wave have a zero velocity in this frame of reference? The only solution lies in the theory of relativity, one

of whose consequences is that a material object like a student or a motorcycle cannot move at the speed of light.

14. Actually, this is only exactly true of the rectangular strip is made infinitesimally thin.
15. You may know already that different colors of light have different speeds when they pass through a material substance, such as the glass or water. This is not in contradiction with what I'm saying here, since this whole analysis is for light in a vacuum.
16. What makes them appear to be unrelated phenomena is that we experience them through their interaction with atoms, and atoms are complicated, so they respond to various kinds of electromagnetic waves in complicated ways.
17. This current will soon come to a grinding halt, because we don't have a complete circuit, but let's say we're talking about the first picosecond during which the radio wave encounters the wire. This is why real radio antennas are *not* very short compared to a wavelength!

12.E: Electromagnetism (Exercises)

1. 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 $(1.0\ \mathrm{m/s})\hat{\mathbf{x}}$. The electric field at point P is $(1.0\ \mathrm{V/m})\hat{\mathbf{y}}$, and the magnetic field is $(2.0\ \mathrm{T})\hat{\mathbf{y}}$. Find the force experienced by the particle.(answer check available at lightandmatter.com)

2. For a positively charged particle moving through a magnetic field, the directions of the $\mathbf{v}$, $\mathbf{B}$, and $\mathbf{F}$ vectors are related by a right-hand rule:

$$\begin{aligned}&\mathbf{v}\text{along the fingers, with the hand flat}\\&\mathbf{B}\text{along the fingers, with the knuckles bent}\\&\mathbf{F}\text{along the thumb}\end{aligned}$$

Make a three-dimensional model of the three vectors using pencils or rolled-up pieces of paper to represent the vectors assembled with their tails together. Make all three vectors perpendicular to each other. Now write down every possible way in which the rule could be rewritten by scrambling up the three symbols $\mathbf{v}$, $\mathbf{B}$, and $\mathbf{F}$. Referring to your model, which are correct and which are incorrect?

3. 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.

4. A charged particle is in a region of space in which there is a uniform magnetic field $\mathbf{B} = B\hat{\mathbf{z}}$. 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. $\mathbf{v}_o = (1\ \mathrm{m/s})\hat{\mathbf{z}}$
c. $\mathbf{v}_o = (1\ \mathrm{m/s})\hat{\mathbf{y}}$

5. (a) A line charge, with charge per unit length λ, moves at velocity v along its own length. How much charge passes a given point in time dt? What is the resulting current? \hwans{hwans:linechargecurrent}
(b) Show that the units of your answer in part a work out correctly.

This constitutes a physical model of an electric current, and it would be a physically realistic model of a beam of particles moving in a vacuum, such as the electron beam in a television tube. It is not a physically realistic model of the motion of the electrons in a current-carrying wire, or of the ions in your nervous system; the motion of the charge carriers in these systems is much more complicated and chaotic, and there are charges of both signs, so that the total charge is zero. But even when the model is physically unrealistic, it still gives the right answers when you use it to compute magnetic effects. This is a remarkable fact, which we will not prove. The interested reader is referred to E.M. Purcell, *Electricity and Magnetism*, McGraw Hill, 1963.

6. Two parallel wires of length L carry currents I_1 and I_2. They are separated by a distance R, and we assume R is much less than L, so that our results for long, straight wires are accurate. The goal of this problem is to compute the magnetic forces acting between the wires.
(a) Neither wire can make a force on *itself*. Therefore, our first step in computing wire 1's force on wire 2 is to find the magnetic field made only by wire 1, in the space *occupied* by wire 2. Express this field in terms of the given quantities.(answer check available at lightandmatter.com)
(b) Let's model the current in wire 2 by pretending that there is a line charge inside it, possessing density per unit length λ_2 and moving at velocity v_2. Relate λ_2 and v_2 to the current I_2, using the result of problem 5a. Now find the magnetic force wire 1 makes on wire 2, in terms of I_1, I_2, L, and R. \hwans{hwans:forcebetweentwowires}
(c) Show that the units of the answer to part b work out to be newtons.

7. 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. 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.)

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:

q_1 = 1 μC	$\mathbf{v}_1 = (1m/s)\hat{\mathbf{x}}$	$\mathbf{F}_1 = (-1mN)\hat{\mathbf{y}}$
q_2 = - 2 μC	$\mathbf{v}_2 = (-1m/s)\hat{\mathbf{x}}$	$\mathbf{F}_2 = (-2mN)\hat{\mathbf{y}}$

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:

q_1 = 1 nC	$\mathbf{v}_1 = (1m/s)\hat{\mathbf{z}}$	$\mathbf{F}_1 = (5mN)\hat{\mathbf{x}} + (2mN)\hat{\mathbf{y}}$
q_2 = 1 nC	$\mathbf{v}_2 = (3m/s)\hat{\mathbf{z}}$	$\mathbf{F}_2 = (10mN)\hat{\mathbf{x}} + (4mN)\hat{\mathbf{y}}$

Is there a nonzero electric field at this point? A nonzero magnetic field?

10. This problem is a continuation of problem 6. Note that the answer to problem 6b is given on page 930.

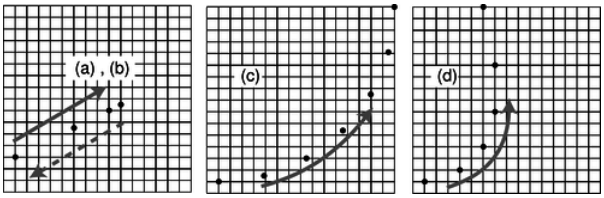
a. Interchanging the 1's and 2's in the answer to problem 6b, what is the magnitude of the magnetic force from wire 2 acting on wire 1? Is this consistent with Newton's third law?
b. Suppose the currents are in the same direction. Make a sketch, and use the right-hand rule to determine whether wire 1 pulls wire 2 towards it, or pushes it away.
c. Apply the right-hand rule again to find the direction of wire 2's force on wire 1. Does this agree with Newton's third law?
d. What would happen if wire 1's current was in the opposite direction compared to wire 2's?

11. (a) In the photo of the vacuum tube apparatus in figure o on page 656, infer the direction of the magnetic field from the motion of the electron beam. (The answer is given in the answer to the self-check on that page.)
(b) Based on your answer to part a, find the direction of the currents in the coils.
(c) What direction are the electrons in the coils going?
(d) Are the currents in the coils repelling the currents consisting of the beam inside the tube, or attracting them? Check your answer by comparing with the result of problem 10.

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.(answer check available at lightandmatter.com)
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. (answer check available at lightandmatter.com)

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.



a / Problem 13.

13. Each figure represents the motion of a positively charged particle. The dots give the particles' positions at equal time intervals. In each case, determine whether the motion was caused by an electric force, a magnetic force, or a frictional force, and explain your reasoning. If possible, determine the direction of the magnetic or electric field. All fields are uniform. In (a), the particle stops for an instant at the upper right, but then comes back down and to the left, retracing the same dots. In (b), it stops on the upper right and stays there.

14. One model of the hydrogen atom has the electron circling around the proton at a speed of 2.2×10^6 m/s, in an orbit with a radius of 0.05 nm. (Although the electron and proton really orbit around their common center of mass, the center of mass is very close to the proton, since it is 2000 times more massive. For this problem, assume the proton is stationary.)

a. Treat the circling electron as a current loop, and calculate the current.
b. Estimate the magnetic field created at the center of the atom by the electron.(answer check available at lightandmatter.com)
c. Does the proton experience a nonzero force from the electron's magnetic field? Explain.
d. Does the electron experience a magnetic field from the proton? Explain.
e. Does the electron experience a magnetic field created by its own current? Explain.
f. Is there an electric force acting between the proton and electron? If so, calculate it.(answer check available at lightandmatter.com)
g. Is there a gravitational force acting between the proton and electron? If so, calculate it.
h. An inward force is required to keep the electron in its orbit -- otherwise it would obey Newton's first law and go straight, leaving the atom. Based on your answers to the previous parts, which force or forces (electric, magnetic and gravitational) contributes significantly to this inward force? (Based on a problem by Arnold Arons.)

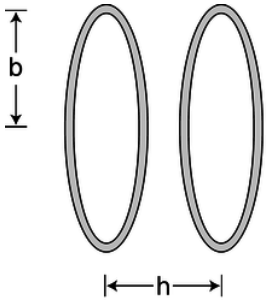
15. The equation $B_z = \beta k I A / c^2 r^3$ was found on page 666 for the distant field of a dipole. Show, as asserted there, that the constant β must be unitless.

16. 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:

q_1 = 1 C	$\mathbf{v}_1 = 0$	$\mathbf{F}_1 = (1N)\hat{\mathbf{y}}$
q_2 =1 C	$\mathbf{v}_2 = (1m/s)\hat{\mathbf{x}}$	$\mathbf{F}_2 = (1N)\hat{\mathbf{y}}$
q_3 =1 C	$\mathbf{v}_3 = (1m/s)\hat{\mathbf{z}}$	$\mathbf{F}_3 = 0$

Determine the electric and magnetic fields.(answer check available at lightandmatter.com)

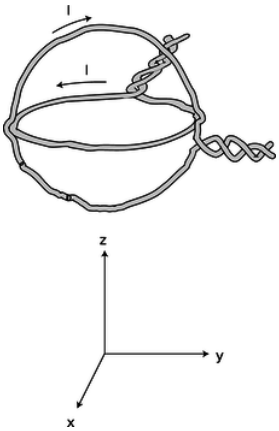
17. If you put four times more current through a solenoid, how many times more energy is stored in its magnetic field?(answer check available at lightandmatter.com)



b / Problem 18.

18. 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 may have more than one turn 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. A photograph of a Helmholtz coil was shown in figure o on page 656.
(a) Find the percentage drop in the field at the center of one coil, compared to the full strength at the center of the whole apparatus. (answer check available at lightandmatter.com)

(b) What value of h (not equal to b) would make this percentage difference equal to zero?(answer check available at lightandmatter.com)



c / Problem 19.

19. 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. (answer check available at lightandmatter.com)
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.(answer check available at lightandmatter.com)

d / Problem 20.

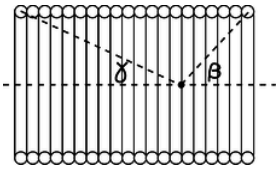
20. Four long wires are arranged, as shown, so that their cross-section forms a square, with connections at the ends so that current flows through all four before exiting. Note that the current is to the right in the two back wires, but to the left in the front wires. If the dimensions of the cross-sectional square (height and front-to-back) are b, find the magnetic field (magnitude and direction) along the long central axis.(answer check available at lightandmatter.com)

21. In problem 16, 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?

22. Use the Biot-Savart law to derive the magnetic field of a long, straight wire, and show that this reproduces the result of example 6 on page 658.

23. (a) Modify the calculation on page 663 to determine the component of the magnetic field of a sheet of charge that is perpendicular to the sheet.(answer check available at lightandmatter.com)
(b) Show that your answer has the right units.
(c) Show that your answer approaches zero as z approaches infinity.
(d) What happens to your answer in the case of $a=b$? Explain why this makes sense.

24. Consider two solenoids, one of which is smaller so that it can be put inside the other. Assume they are long enough so that each one only contributes significantly to the field inside itself, and the interior fields are nearly uniform. Consider the configuration where the small one is inside the big one with their currents circulating in the same direction, and a second configuration in which the currents circulate in opposite directions. Compare the energies of these configurations with the energy when the solenoids are far apart. Based on this reasoning, which configuration is stable, and in which configuration will the little solenoid tend to get twisted around or spit out? \hwhint{hwhint:nestedsolenoids}



e / Problem 25.

25. (a) A solenoid can be imagined as a series of circular current loops that are spaced along their common axis. Integrate the result of example 12 on page 671 to show that the field on the axis of a solenoid can be written as $B = (2\pi k\eta/c^2)(\cos\beta + \cos\gamma)$, where the angles β and γ are defined in the figure.
(b) Show that in the limit where the solenoid is very long, this exact result agrees with the approximate one derived in example 13 on page 674 using Ampère's law.
(c) Note that, unlike the calculation using Ampère's law, this one is valid at points that are near the mouths of the solenoid, or even outside it entirely. If the solenoid is long, at what point on the axis is the field equal to one half of its value at the center of the solenoid?
(d) What happens to your result when you apply it to points that are very far away from the solenoid? Does this make sense?

26. The first step in the proof of Ampère's law on page 675 is to show that Ampère's law holds in the case shown in figure f/1, where a circular Ampèrian loop is centered on a long, straight wire that is perpendicular to the plane of the loop. Carry out this calculation, using the result for the field of a wire that was established without using Ampère's law.

27. A certain region of space has a magnetic field given by $\mathbf{B} = bx\hat{\mathbf{y}}$. Find the electric current flowing through the square defined by $z = 0$, $0 \le x \le a$, and $0 \le y \le a$.(answer check available at lightandmatter.com)

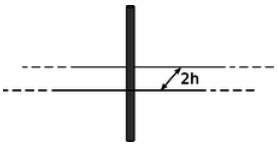
f / A nautilus shell is approximately a logarithmic spiral, of the type in problem 28.

28. Perform a calculation similar to the one in problem 54, but for a logarithmic spiral, defined by $r = we^{u\theta}$, and show that the field is $B = (kI/c^2u)(1/a - 1/b)$. Note that the solution to problem 54 is given in the back of the book.

29. (a) For the geometry described in example 8 on page 661, find the field at a point the lies in the plane of the wires, but not between the wires, at a distance b from the center line. Use the same technique as in that example.
(b) Now redo the calculation using the technique demonstrated on page 666. The integrals are nearly the same, but now the reasoning is reversed: you already know $\beta = 1$, and you want to find an unknown field. The only difference in the integrals is that you are tiling a different region of the plane in order to mock up the currents in the two wires. Note that you can't tile a region that contains a point of interest, since the technique uses the field of a distant dipole.(answer check available at lightandmatter.com)

30. (a) A long, skinny solenoid consists of N turns of wire wrapped uniformly around a hollow cylinder of length ℓ and cross-sectional area A. Find its inductance.(answer check available at lightandmatter.com)
(b) Show that your answer has the right units to be an inductance.

31. Consider two solenoids, one of which is smaller so that it can be put inside the other. Assume they are long enough to act like ideal solenoids, so that each one only contributes significantly to the field inside itself, and the interior fields are nearly uniform. Consider the configuration where the small one is partly inside and partly hanging out of the big one, with their currents circulating in the same direction. Their axes are constrained to coincide.
(a) Find the difference in the magnetic energy between the configuration where the solenoids are separate and the configuration where the small one is inserted into the big one. Your equation will include the length x of the part of the small solenoid that is inside the big one, as well as other relevant variables describing the two solenoids.(answer check available at lightandmatter.com)
(b) Based on your answer to part a, find the force acting between the solenoids.(answer check available at lightandmatter.com)

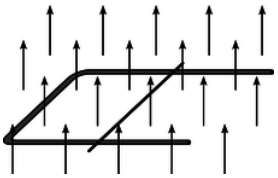


g / Problem 32.

32. 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 Ampèrian surface. The length of the rectangle is infinite, so it's not necessary to compute the contributions of the ends.

33. The purpose of this problem is to find how the gain of a transformer depends on its construction.
(a) The number of loops of wire, N, in a solenoid is changed, while keeping the length constant. How does the impedance depend on N? State your answer as a proportionality, e.g., $Z \propto N^3$ or $Z \propto N^{-5}$.
(b) For a given AC voltage applied across the inductor, how does the magnetic field depend on N? You need to take into account both the dependence of a solenoid's field on N for a given current and your answer to part a, which affects the current.
(c) Now consider a transformer consisting of two solenoids. The input side has N_1 loops, and the output N_2. We wish to find how the output voltage V_2 depends on N_1, N_2, and the input voltage V_1. The text has already established $V_2 \propto V_1 N_2$, so it only remains to find the dependence on N_1. Use your result from part b to accomplish this. The ratio V_2/V_1 is called the voltage gain.

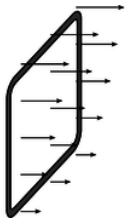
34. Problem 33 dealt with the dependence of a transformer's gain on the number of loops of wire in the input solenoid. Carry out a similar analysis of how the gain depends on the frequency at which the circuit is operated.



h / Problem 35.

35. A U-shaped wire makes electrical contact with a second, straight wire, which rolls along it to the right, as shown in the figure. The whole thing is immersed in a uniform magnetic field, which is perpendicular to the plane of the circuit. The resistance of the rolling wire is much greater than that of the U.
(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 magnitude of the force acting on the wire. There is more than one way to do this, but please do it using Faraday's law (which works even though it's the Ampèrian surface itself that is changing, rather than the field).(answer check available at lightandmatter.com)
(d) 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 c.

36. A charged particle is in motion at speed v, in a region of vacuum through which an electromagnetic wave is passing. In what direction should the particle be moving in order to minimize the total force acting on it? Consider both possibilities for the sign of the charge. (Based on a problem by David J. Raymond.)



i / Problem 37.

37. 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.(answer check available at lightandmatter.com)

38. A capacitor has parallel plates of area A, separated by a distance h. If there is a vacuum between the plates, then Gauss's law gives $E = 4\pi k\sigma = 4\pi kq/A$ for the field between the plates, and combining this with $E = V/h$, we find $C = q/V = (1/4\pi k)A/h$. (a) Generalize this derivation to the case where there is a dielectric between the plates. (b) Suppose we have a list of possible materials we could choose as dielectrics, and we wish to construct a capacitor that will have the highest possible energy density, U_e/v, where v is the volume. For each dielectric, we know its permittivity ϵ, and also the maximum electric field E it can sustain without breaking down and allowing sparks to cross between the plates. Write the maximum energy density in terms of these two variables, and determine a figure of merit that could be used to decide which material would be the best choice.

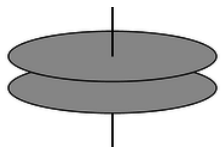
39. (a) For each term appearing on the right side of Maxwell's equations, give an example of an everyday situation it describes.
(b) Most people doing calculations in the SI system of units don't use k and k/c^2. Instead, they express everything in terms of the constants

$$\epsilon_o = \frac{1}{4\pi k} \text{and}$$
$$\mu_o = \frac{4\pi k}{c^2}.$$

Rewrite Maxwell's equations in terms of these constants, eliminating k and c everywhere.

40. (a) Prove that in an electromagnetic plane wave, half the energy is in the electric field and half in the magnetic field.
(b) Based on your result from part a, find the proportionality constant in the relation $d\mathbf{p} \propto \mathbf{E} \times \mathbf{B} dv$, where $d\mathbf{p}$ is the momentum of the part of a plane light wave contained in the volume dv. The vector $\mathbf{E} \times \mathbf{B}$ is known as the Poynting vector. (To do this problem, you need to know the relativistic relationship between the energy and momentum of a beam of light.)(answer check available at lightandmatter.com)

41. (a) A beam of light has cross-sectional area A and power P, i.e., P is the number of joules per second that enter a window through which the beam passes. Find the energy density U/v in terms of P, A, and universal constants.
(b) Find $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{B}}$, the amplitudes of the electric and magnetic fields, in terms of P, A, and universal constants (i.e., your answer should *not* include U or v). You will need the result of problem 40a. A real beam of light usually consists of many short wavetrains, not one big sine wave, but don't worry about that.(answer check available at lightandmatter.com)\hwhint{hwhint:solarconstant}
(c) A beam of sunlight has an intensity of $P/A = 1.35 \times 10^3 \text{ W}/\text{m}^2$, assuming no clouds or atmospheric absorption. This is known as the solar constant. Compute $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{B}}$, and compare with the strengths of static fields you experience in everyday life: $E \sim 10^6 \text{ V}/\text{m}$ in a thunderstorm, and $B \sim 10^{-3}$ T for the Earth's magnetic field.(answer check available at lightandmatter.com)



j / Problem 42.

42. 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. \hwhint{hwhint:circularcap}(answer check available at lightandmatter.com)

43. A positively charged particle is released from rest at the origin at $t = 0$, in a region of vacuum through which an electromagnetic wave is passing. The particle accelerates in response to the wave. In this region of space, the wave varies as $\mathbf{E} = \hat{\mathbf{x}}\tilde{E}\sin\omega t$, $\mathbf{B} = \hat{\mathbf{y}}\tilde{B}\sin\omega t$, and we assume that the particle has a relatively large value of m/q, so that its response to the wave is sluggish, and it never ends up moving at any speed comparable to the speed of light. Therefore we don't have to worry about the spatial variation of the wave; we can just imagine that these are uniform fields imposed by some external mechanism on this region of space.

(a) Find the particle's coordinates as functions of time.(answer check available at lightandmatter.com)
(b) Show that the motion is confined to $-z_{max} \le z \le z_{max}$, where $z_{max} = 1.101 \left(q^2 \tilde{E} \tilde{B} / m^2 \omega^3 \right)$.

44. Electromagnetic waves are supposed to have their electric and magnetic fields perpendicular to each other. (Throughout this problem, assume we're talking about waves traveling through a vacuum, and that there is only a single sine wave traveling in a single direction, not a superposition of sine waves passing through each other.) Suppose someone claims they can make an electromagnetic wave in which the electric and magnetic fields lie in the same plane. Prove that this is impossible based on Maxwell's equations.

45. Repeat the self-check on page 710, but with one change in the procedure: after we charge the capacitor, we open the circuit, and then continue with the observations.

46. On page 713, I proved that $\mathbf{H}_{\parallel,1} = \mathbf{H}_{\parallel,2}$ at the boundary between two substances if there is no free current and the fields are static. In fact, each of Maxwell's four equations implies a constraint with a similar structure. Some are constraints on the field components parallel to the boundary, while others are constraints on the perpendicular parts. Since some of the fields referred to in Maxwell's equations are the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$, while others are the auxiliary fields $\mathbf{D}$ and $\mathbf{H}$, some of the constraints deal with $\mathbf{E}$ and $\mathbf{B}$, others with $\mathbf{D}$ and $\mathbf{H}$. Find the other three constraints.

47. (a) Figure j on page 714 shows a hollow sphere with $\mu/\mu_o = x$, inner radius a, and outer radius b, which has been subjected to an external field $\mathbf{B}_o$. Finding the fields on the exterior, in the shell, and on the interior requires finding a set of fields that satisfies five boundary conditions: (1) far from the sphere, the field must approach the constant $\mathbf{B}_o$; (2) at the outer surface of the sphere, the field must have $\mathbf{H}_{\parallel,1} = \mathbf{H}_{\parallel,2}$, as discussed on page 713; (3) the same constraint applies at the inner surface of the sphere; (4) and (5) there is an additional constraint on the fields at the inner and outer surfaces, as found in problem 46. The goal of this problem is to find the solution for the fields, and from it, to prove that the interior field is uniform, and given by

$$\mathbf{B} = \left[\frac{9x}{(2x+1)(x+2) - 2\frac{a^3}{b^3}(x-1)^2} \right] \mathbf{B}_o .$$

This is a very difficult problem to solve from first principles, because it's not obvious what form the fields should have, and if you hadn't been told, you probably wouldn't have guessed that the interior field would be uniform. We could, however, guess that once the sphere becomes polarized by the external field, it would become a dipole, and at $r \gg b$, the field would be a uniform field superimposed on the field of a dipole. It turns out that even close to the sphere, the solution has exactly this form. In order to complete the solution, we need to find the field in the shell ($a < r < b$), but the only way this field could match up with the detailed angular variation of the interior and exterior fields would be if it was also a superposition of a uniform field with a dipole field. The final result is that we have four unknowns: the strength of the dipole component of the external field, the strength of the uniform and dipole components of the field within the shell, and the strength of the uniform interior field. These four unknowns are to be determined by imposing constraints (2) through (5) above.
(b) Show that the expression from part a has physically reasonable behavior in its dependence on x and a/b.

48. Two long, parallel strips of thin metal foil form a configuration like a long, narrow sandwich. The air gap between them has height h, the width of each strip is w, and their length is ℓ. Each strip carries current I, and we assume for concreteness that the currents are in opposite directions, so that the magnetic force, F, between the strips is repulsive.
(a) Find the force in the limit of $w \gg h$.(answer check available at lightandmatter.com)
(b) Find the force in the limit of $w \ll h$, which is like two ordinary wires.
(c) Discuss the relationship between the two results.

49. Suppose we are given a permanent magnet with a complicated, asymmetric shape. Describe how a series of measurements with a magnetic compass could be used to determine the strength and direction of its magnetic field at some point of interest. Assume that you are only able to see the direction to which the compass needle settles; you cannot measure the torque acting on it.

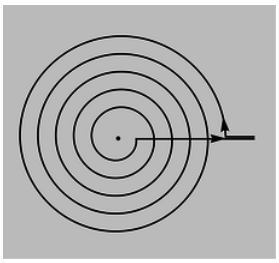
50. On page 680, the curl of $x\hat{\mathbf{y}}$ was computed. Now consider the fields $x\hat{\mathbf{x}}$ and $y\hat{\mathbf{y}}$.
(a) Sketch these fields.
(b) Using the same technique of explicitly constructing a small square, prove that their curls are both zero. Do not use the component form of the curl; this was one step in *deriving* the component form of the curl.

51. If you watch a movie played backwards, some vectors reverse their direction. For instance, people walk backwards, with their velocity vectors flipped around. Other vectors, such as forces, keep the same direction, e.g., gravity still pulls down. An electric

field is another example of a vector that doesn't turn around: positive charges are still positive in the time-reversed universe, so they still make diverging electric fields, and likewise for the converging fields around negative charges.
(a) How does the momentum of a material object behave under time-reversal?(solution in the pdf version of the book)
(b) The laws of physics are still valid in the time-reversed universe. For example, show that if two material objects are interacting, and momentum is conserved, then momentum is still conserved in the time-reversed universe.(solution in the pdf version of the book)
(c) Discuss how currents and magnetic fields would behave under time reversal. \hwhint{hwhint:timereversalem}
(d) Similarly, show that the equation $d\mathbf{p} \propto \mathbf{E} \times \mathbf{B}$ is still valid under time reversal.

52. This problem is a more advanced exploration of the time-reversal ideas introduced in problem 51.
(a) In that problem, we assumed that charge did not flip its sign under time reversal. Suppose we make the opposite assumption, that charge *does* change its sign. This is an idea introduced by Richard Feynman: that antimatter is really matter traveling backward in time! Determine the time-reversal properties of $\mathbf{E}$ and $\mathbf{B}$ under this new assumption, and show that $d\mathbf{p} \propto \mathbf{E} \times \mathbf{B}$ is still valid under time-reversal.
(b) Show that Maxwell's equations are time-reversal symmetric, i.e., that if the fields $\mathbf{E}(x, y, z, t)$ and $\mathbf{B}(x, y, z, t)$ satisfy Maxwell's equations, then so do $\mathbf{E}(x, y, z, -t)$ and $\mathbf{B}(x, y, z, -t)$. Demonstrate this under both possible assumptions about charge, $q \to q$ and $q \to -q$.

53. The purpose of this problem is to prove that the constant of proportionality a in the equation $dU_m = aB^2 dv$, for the energy density of the magnetic field, is given by $a = c^2/8\pi k$ as asserted on page 665. The geometry we'll use consists of two sheets of current, like a sandwich with nothing in between but some vacuum in which there is a magnetic field. The currents are in opposite directions, and we can imagine them as being joined together at the ends to form a complete circuit, like a tube made of paper that has been squashed almost flat. The sheets have lengths L in the direction parallel to the current, and widths w. They are separated by a distance d, which, for convenience, we assume is small compared to L and w. Thus each sheet's contribution to the field is uniform, and can be approximated by the expression $2\pi k\eta/c^2$.
(a) Make a drawing similar to the one in figure 11.2.1 on page 664, and show that in this opposite-current configuration, the magnetic fields of the two sheets reinforce in the region between them, producing double the field, but cancel on the outside.
(b) By analogy with the case of a single strand of wire, one sheet's force on the other is ILB_1, were $I = \eta w$ is the total current in one sheet, and $B_1 = B/2$ is the field contributed by only one of the sheets, since the sheet can't make any net force on itself. Based on your drawing and the right-hand rule, show that this force is repulsive.
For the rest of the problem, consider a process in which the sheets start out touching, and are then separated to a distance d. Since the force between the sheets is repulsive, they do mechanical work on the outside world as they are separated, in much the same way that the piston in an engine does work as the gases inside the cylinder expand. At the same time, however, there is an induced emf which would tend to extinguish the current, so in order to maintain a constant current, energy will have to be drained from a battery. There are three types of energy involved: the increase in the magnetic field energy, the increase in the energy of the outside world, and the decrease in energy as the battery is drained. (We assume the sheets have very little resistance, so there is no ohmic heating involved.)(answer check available at lightandmatter.com)
(c) Find the mechanical work done by the sheets, which equals the increase in the energy of the outside world. Show that your result can be stated in terms of η, the final volume $v = wLd$, and nothing else but numerical and physical constants.(answer check available at lightandmatter.com)
(d) The power supplied by the battery is $P = I\Gamma_E$ (like $P = I\Delta V$, but with an emf instead of a voltage difference), and the circulation is given by $\Gamma = -d\Phi_B/dt$. The negative sign indicates that the battery is being drained. Calculate the energy supplied by the battery, and, as in part c, show that the result can be stated in terms of η, v, and universal constants.(answer check available at lightandmatter.com)
(e) Find the increase in the magnetic-field energy, in terms of η, v, and the unknown constant a.(answer check available at lightandmatter.com)
(f) Use conservation of energy to relate your answers from parts c, d, and e, and solve for a.(answer check available at lightandmatter.com)



k / Problem 54.

54. 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 in the pdf version of the book)

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

CHAPTER OVERVIEW

13: Optics

Contributors and Attributions

13.1: The Ray Model of Light

Ads for one Macintosh computer bragged that it could do an arithmetic calculation in less time than it took for the light to get from the screen to your eye. We find this impressive because of the contrast between the speed of light and the speeds at which we interact with physical objects in our environment. Perhaps it shouldn't surprise us, then, that Newton succeeded so well in explaining the motion of objects, but was far less successful with the study of light.

The climax of our study of electricity and magnetism was discovery that light is an electromagnetic wave. Knowing this, however, is not the same as knowing everything about eyes and telescopes. In fact, the full description of light as a wave can be rather cumbersome. We will instead spend most of our treatment of optics making use of a simpler model of light, the ray model, which does a fine job in most practical situations. Not only that, but we will even backtrack a little and start with a discussion of basic ideas about light and vision that predated the discovery of electromagnetic waves.

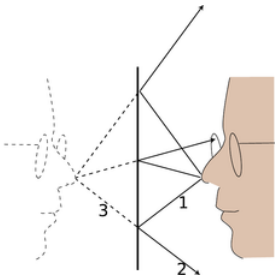
Contributors and Attributions

13.2: Images by Reflection

Infants are always fascinated by the antics of the Baby in the Mirror. Now if you want to know something about mirror images that most people don't understand, try this. First bring this page closer and closer to your eyes, until you can no longer focus on it without straining. Then go in the bathroom and see how close you can get your face to the surface of the mirror before you can no longer easily focus on the image of your own eyes. You will find that the shortest comfortable eye-mirror distance is much less than the shortest comfortable eye-paper distance. This demonstrates that the image of your face in the mirror acts as if it had depth and existed in the space *behind* the mirror. If the image was like a flat picture in a book, then you wouldn't be able to focus on it from such a short distance.

In this chapter we will study the images formed by flat and curved mirrors on a qualitative, conceptual basis. Although this type of image is not as commonly encountered in everyday life as images formed by lenses, images formed by reflection are simpler to understand, so we discuss them first. In section 12.3 we will turn to a more mathematical treatment of images made by reflection. Surprisingly, the same equations can also be applied to lenses, which are the topic of section 12.4.

12.2.1 A virtual image



a / An image formed by a mirror.

We can understand a mirror image using a ray diagram. Figure a shows several light rays, 1, that originated by diffuse reflection at the person's nose. They bounce off the mirror, producing new rays, 2. To anyone whose eye is in the right position to get one of these rays, they appear to have come from a behind the mirror, 3, where they would have originated from a single point. This point is where the tip of the image-person's nose appears to be. A similar analysis applies to every other point on the person's face, so it looks as though there was an entire face behind the mirror. The customary way of describing the situation requires some explanation:

- **Customary description in physics:** There is an image of the face behind the mirror.
- **Translation:** The pattern of rays coming from the mirror is exactly the same as it would be if there were a face behind the mirror. Nothing is really behind the mirror.

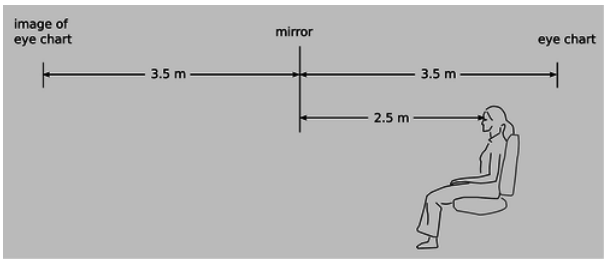
This is referred to as a *virtual* image, because the rays do not actually cross at the point behind the mirror. They only appear to have originated there.

self-check:

Imagine that the person in figure a moves his face down quite a bit --- a couple of feet in real life, or a few inches on this scale drawing. The mirror stays where it is. Draw a new ray diagram. Will there still be an image? If so, where is it visible from?

(answer in the back of the PDF version of the book)

The geometry of specular reflection tells us that rays 1 and 2 are at equal angles to the normal (the imaginary perpendicular line piercing the mirror at the point of reflection). This means that ray 2's imaginary continuation, 3, forms the same angle with the mirror as ray 1. Since each ray of type 3 forms the same angles with the mirror as its partner of type 1, we see that the distance of the image from the mirror is the same as that of the actual face from the mirror, and it lies directly across from it. The image therefore appears to be the same size as the actual face.



b / Example 2.

Example 2: An eye exam

Figure b shows a typical setup in an optometrist's examination room. The patient's vision is supposed to be tested at a distance of 6 meters (20 feet in the U.S.), but this distance is larger than the amount of space available in the room. Therefore a mirror is used to create an image of the eye chart behind the wall.

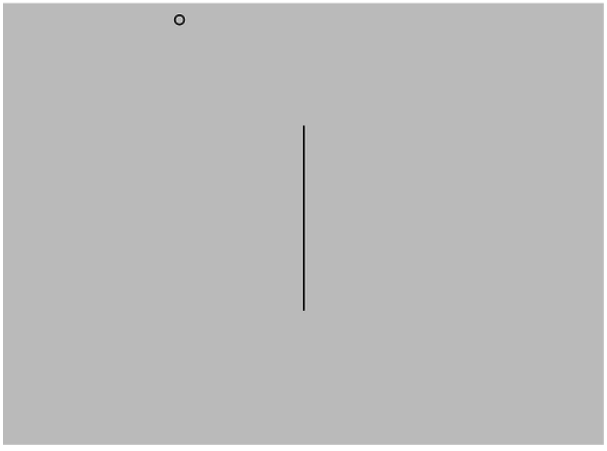
Example 3: The Praxinoscope

c / The praxinoscope.

Figure c shows an old-fashioned device called a praxinoscope, which displays an animated picture when spun. The removable strip of paper with the pictures printed on it has twice the radius of the inner circle made of flat mirrors, so each picture's virtual image is at the center. As the wheel spins, each picture's image is replaced by the next.

Discussion Question

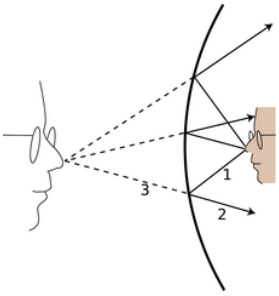
◊ The figure shows an object that is off to one side of a mirror. Draw a ray diagram. Is an image formed? If so, where is it, and from which directions would it be visible?



12.2.2 Curved mirrors

An image in a flat mirror is a pretechnological example: even animals can look at their reflections in a calm pond. We now pass to our first nontrivial example of the manipulation of an image by technology: an image in a curved mirror. Before we dive in, let's consider why this is an important example. If it was just a question of memorizing a bunch of facts about curved mirrors, then you

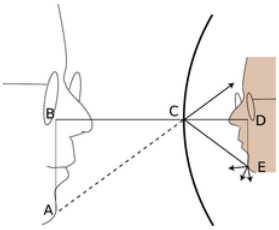
would rightly rebel against an effort to spoil the beauty of your liberally educated brain by force-feeding you technological trivia. The reason this is an important example is not that curved mirrors are so important in and of themselves, but that the results we derive for curved bowl-shaped mirrors turn out to be true for a large class of other optical devices, including mirrors that bulge outward rather than inward, and lenses as well. A microscope or a telescope is simply a combination of lenses or mirrors or both. What you're really learning about here is the basic building block of all optical devices from movie projectors to octopus eyes.



d / An image formed by a curved mirror.

Because the mirror in figure d is curved, it bends the rays back closer together than a flat mirror would: we describe it as *converging*. Note that the term refers to what it does to the light rays, not to the physical shape of the mirror's surface . (The surface itself would be described as *concave*. The term is not all that hard to remember, because the hollowed-out interior of the mirror is like a cave.) It is surprising but true that all the rays like 3 really do converge on a point, forming a good image. We will not prove this fact, but it is true for any mirror whose curvature is gentle enough and that is symmetric with respect to rotation about the perpendicular line passing through its center (not asymmetric like a potato chip). The old-fashioned method of making mirrors and lenses is by grinding them in grit by hand, and this automatically tends to produce an almost perfect spherical surface.

Bending a ray like 2 inward implies bending its imaginary continuation 3 outward, in the same way that raising one end of a seesaw causes the other end to go down. The image therefore forms deeper behind the mirror. This doesn't just show that there is extra distance between the image-nose and the mirror; it also implies that the image itself is bigger from front to back. It has been *magnified* in the front-to-back direction.



e / The image is magnified by the same factor in depth and in its other dimensions.

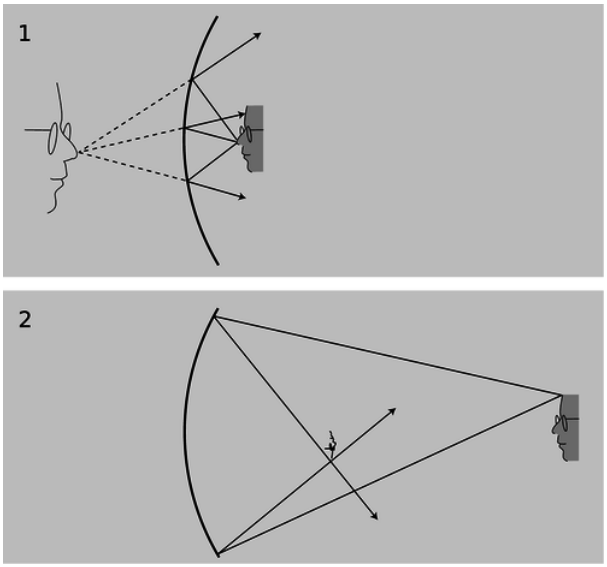
It is easy to prove that the same magnification also applies to the image's other dimensions. Consider a point like E in figure e. The trick is that out of all the rays diffusely reflected by E, we pick the one that happens to head for the mirror's center, C. The equal-angle property of specular reflection plus a little straightforward geometry easily leads us to the conclusion that triangles ABC and CDE are the same shape, with ABC being simply a scaled-up version of CDE. The magnification of depth equals the ratio BC/CD, and the up-down magnification is AB/DE. A repetition of the same proof shows that the magnification in the third dimension (out of the page) is also the same. This means that the image-head is simply a larger version of the real one, without any distortion. The scaling factor is called the magnification, M. The image in the figure is magnified by a factor $M = 1.9$.

f / Increased magnification always comes at the expense of decreased field of view.

Note that we did not explicitly specify whether the mirror was a sphere, a paraboloid, or some other shape. However, we assumed that a focused image would be formed, which would not necessarily be true, for instance, for a mirror that was asymmetric or very deeply curved.

12.2.3 A real image

If we start by placing an object very close to the mirror, g/1, and then move it farther and farther away, the image at first behaves as we would expect from our everyday experience with flat mirrors, receding deeper and deeper behind the mirror. At a certain point, however, a dramatic change occurs. When the object is more than a certain distance from the mirror, g/2, the image appears upside-down and in *front* of the mirror.



g / 1. A virtual image. 2. A real image. As you'll verify in homework problem 12, the image is upside-down

Here's what's happened. The mirror bends light rays inward, but when the object is very close to it, as in g/1, the rays coming from a given point on the object are too strongly diverging (spreading) for the mirror to bring them back together. On reflection, the rays are still diverging, just not as strongly diverging. But when the object is sufficiently far away, g/2, the mirror is only intercepting the rays that came out in a narrow cone, and it is able to bend these enough so that they will reconverge.

Note that the rays shown in the figure, which both originated at the same point on the object, reunite when they cross. The point where they cross is the image of the point on the original object. This type of image is called a *real image*, in contradistinction to the virtual images we've studied before.

Definition: A real image is one where rays actually cross. A virtual image is a point from which rays only appear to have come.

The use of the word “real” is perhaps unfortunate. It sounds as though we are saying the image was an actual material object, which of course it is not.

The distinction between a real image and a virtual image is an important one, because a real image can be projected onto a screen or photographic film. If a piece of paper is inserted in figure g/2 at the location of the image, the image will be visible on the paper (provided the object is bright and the room is dark). Your eye uses a lens to make a real image on the retina.

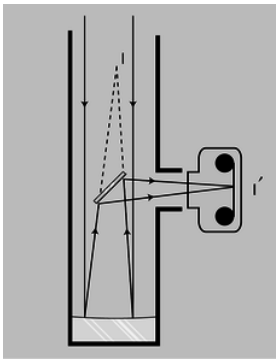
self-check:

Sketch another copy of the face in figure g/1, even farther from the mirror, and draw a ray diagram. What has happened to the location of the image?

(answer in the back of the PDF version of the book)

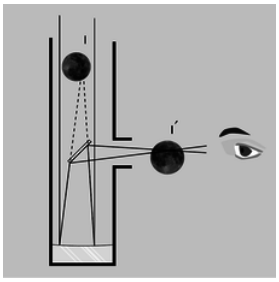
12.2.4 Images of images

If you are wearing glasses right now, then the light rays from the page are being manipulated first by your glasses and then by the lens of your eye. You might think that it would be extremely difficult to analyze this, but in fact it is quite easy. In any series of optical elements (mirrors or lenses or both), each element works on the rays furnished by the previous element in exactly the same manner as if the image formed by the previous element was an actual object.



h / A Newtonian telescope being used with a camera.

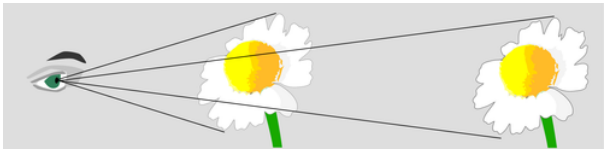
Figure h shows an example involving only mirrors. The Newtonian telescope, invented by Isaac Newton, consists of a large curved mirror, plus a second, flat mirror that brings the light out of the tube. (In very large telescopes, there may be enough room to put a camera or even a person inside the tube, in which case the second mirror is not needed.) The tube of the telescope is not vital; it is mainly a structural element, although it can also be helpful for blocking out stray light. The lens has been removed from the front of the camera body, and is not needed for this setup. Note that the two sample rays have been drawn parallel, because an astronomical telescope is used for viewing objects that are extremely far away. These two "parallel" lines actually meet at a certain point, say a crater on the moon, so they can't actually be perfectly parallel, but they are parallel for all practical purposes since we would have to follow them upward for a quarter of a million miles to get to the point where they intersect.



i / A Newtonian telescope being used for visual rather than photographic observing. In real life, an eyepiece lens is normally used for additional magnification, but this simpler setup will also work.

The large curved mirror by itself would form an image I, but the small flat mirror creates an image of the image, I'. The relationship between I and I' is exactly the same as it would be if I was an actual object rather than an image: I and I' are at equal distances from the plane of the mirror, and the line between them is perpendicular to the plane of the mirror.

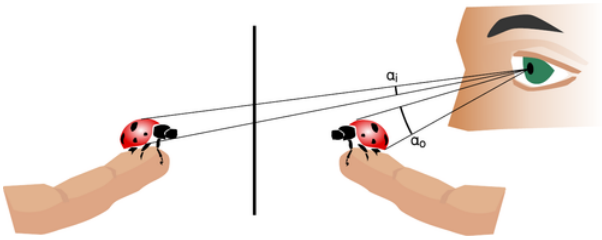
One surprising wrinkle is that whereas a flat mirror used by itself forms a virtual image of an object that is real, here the mirror is forming a real image of virtual image I. This shows how pointless it would be to try to memorize lists of facts about what kinds of images are formed by various optical elements under various circumstances. You are better off simply drawing a ray diagram.



j / The angular size of the flower depends on its distance from the eye.

Although the main point here was to give an example of an image of an image, figure i also shows an interesting case where we need to make the distinction between *magnification* and *angular magnification*. If you are looking at the moon through this

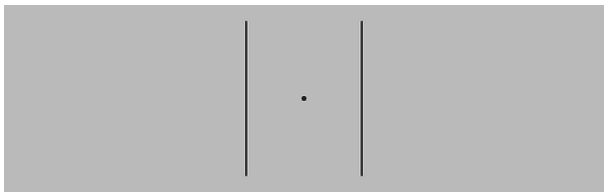
telescope, then the images I and I′ are much *smaller* than the actual moon. Otherwise, for example, image I would not fit inside the telescope! However, these images are very close to your eye compared to the actual moon. The small size of the image has been more than compensated for by the shorter distance. The important thing here is the amount of *angle* within your field of view that the image covers, and it is this angle that has been increased. The factor by which it is increased is called the *angular magnification*, M_a.



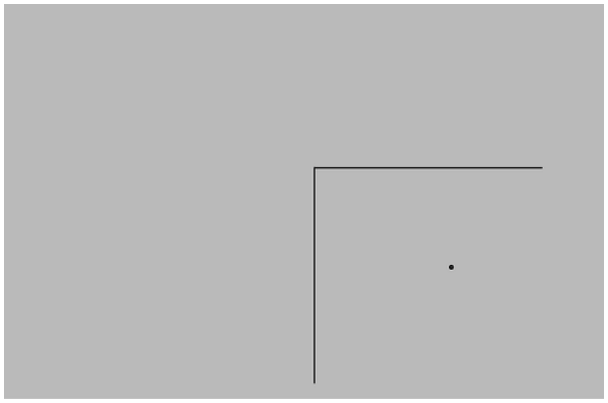
k / The person uses a mirror to get a view of both sides of the ladybug. Although the flat mirror has $M = 1$, it doesn't give an angular magnification of 1. The image is farther from the eye than the object, so the angular magnification $M_a = \alpha_i/\alpha_o$ is less than one.

Discussion Questions

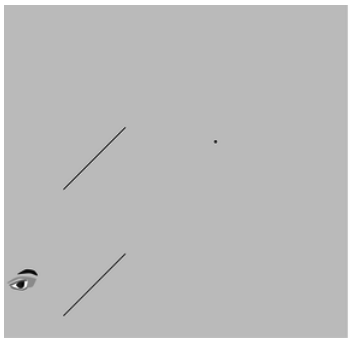
◊ Locate the images of you that will be formed if you stand between two parallel mirrors.



◊ Locate the images formed by two perpendicular mirrors, as in the figure. What happens if the mirrors are not perfectly perpendicular?



◊ Locate the images formed by the periscope.



Contributor

Benjamin Crowell (Fullerton College).

13.3: Images, Quantitatively

It sounds a bit odd when a scientist refers to a theory as “beautiful,” but to those in the know it makes perfect sense. One mark of a beautiful theory is that it surprises us by being simple. The mathematical theory of lenses and curved mirrors gives us just such a surprise. We expect the subject to be complex because there are so many cases: a converging mirror forming a real image, a diverging lens that makes a virtual image, and so on for a total of six possibilities. If we want to predict the location of the images in all these situations, we might expect to need six different equations, and six more for predicting magnifications. Instead, it turns out that we can use just one equation for the location of the image and one equation for its magnification, and these two equations work in all the different cases with no changes except for plus and minus signs. This is the kind of thing the physicist Eugene Wigner referred to as “the unreasonable effectiveness of mathematics.” Sometimes we can find a deeper reason for this kind of unexpected simplicity, but sometimes it almost seems as if God went out of Her way to make the secrets of universe susceptible to attack by the human thought-tool called math.

Contributors and Attributions

13.4: Refraction

Economists normally consider free markets to be the natural way of judging the monetary value of something, but social scientists also use questionnaires to gauge the relative value of privileges, disadvantages, or possessions that cannot be bought or sold. They ask people to *imagine* that they could trade one thing for another and ask which they would choose. One interesting result is that the average light-skinned person in the U.S. would rather lose an arm than suffer the racist treatment routinely endured by African-Americans. Even more impressive is the value of sight. Many prospective parents can imagine without too much fear having a deaf child, but would have a far more difficult time coping with raising a blind one.

So great is the value attached to sight that some have imbued it with mystical aspects. Joan of Arc saw visions, and my college has a "vision statement." Christian fundamentalists who perceive a conflict between evolution and their religion have claimed that the eye is such a perfect device that it could never have arisen through a process as helter-skelter as evolution, or that it could not have evolved because half of an eye would be useless. In fact, the structure of an eye is fundamentally dictated by physics, and it has arisen separately by evolution somewhere between eight and 40 times, depending on which biologist you ask. We humans have a version of the eye that can be traced back to the evolution of a light-sensitive "eye spot" on the head of an ancient invertebrate. A sunken pit then developed so that the eye would only receive light from one direction, allowing the organism to tell where the light was coming from. (Modern flatworms have this type of eye.) The top of the pit then became partially covered, leaving a hole, for even greater directionality (as in the nautilus). At some point the cavity became filled with jelly, and this jelly finally became a lens, resulting in the general type of eye that we share with the bony fishes and other vertebrates. Far from being a perfect device, the vertebrate eye is marred by a serious design flaw due to the lack of planning or intelligent design in evolution: the nerve cells of the retina and the blood vessels that serve them are all in front of the light-sensitive cells, blocking part of the light. Squids and other molluscs, whose eyes evolved on a separate branch of the evolutionary tree, have a more sensible arrangement, with the light-sensitive cells out in front.

12.4.1 Refraction

The fundamental physical phenomenon at work in the eye is that when light crosses a boundary between two media (such as air and the eye's jelly), part of its energy is reflected, but part passes into the new medium. In the ray model of light, we describe the original ray as splitting into a reflected ray and a transmitted one (the one that gets through the boundary). Of course the reflected ray goes in a direction that is different from that of the original one, according to the rules of reflection we have already studied. More surprisingly --- and this is the crucial point for making your eye focus light --- the transmitted ray is bent somewhat as well. This bending phenomenon is called *refraction*. The origin of the word is the same as that of the word "fracture," i.e., the ray is bent or "broken." (Keep in mind, however, that light rays are not physical objects that can really be "broken.") Refraction occurs with all waves, not just light waves.

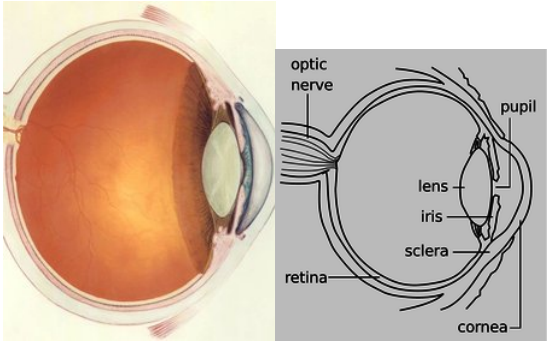


Figure* 13.4.1** *: A human eye.* ***Figure* 13.4.2*: *The anatomy of the eye.*

The actual anatomy of the eye, Figure 13.4.2, is quite complex, but in essence it is very much like every other optical device based on refraction. The rays are bent when they pass through the front surface of the eye, Figure 13.4.3. Rays that enter farther from the central axis are bent more, with the result that an image is formed on the retina. There is only one slightly novel aspect of the situation. In most human-built optical devices, such as a movie projector, the light is bent as it passes into a lens, bent again as it reemerges, and then reaches a focus beyond the lens. In the eye, however, the "screen" is inside the eye, so the rays are only refracted once, on entering the jelly, and never emerge again.

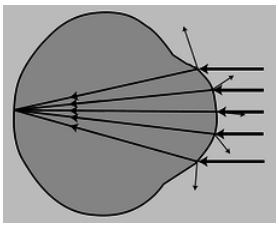


***Figure* 13.4.3:** *A simplified optical diagram of the eye. Light rays are bent when they cross from the air into the eye. (A little of the incident rays' energy goes into the reflected rays rather than the ones transmitted into the eye.)*

A common misconception is that the “lens” of the eye is what does the focusing. All the transparent parts of the eye are made of fairly similar stuff, so the dramatic change in medium is when a ray crosses from the air into the eye (at the outside surface of the cornea). This is where nearly all the refraction takes place. The lens medium differs only slightly in its optical properties from the rest of the eye, so very little refraction occurs as light enters and exits the lens. The lens, whose shape is adjusted by muscles attached to it, is only meant for fine-tuning the focus to form images of near or far objects.

Refractive Properties of Media

What are the rules governing refraction? The first thing to observe is that just as with reflection, the new, bent part of the ray lies in the same plane as the normal (perpendicular) and the incident ray, Figure 13.4.4

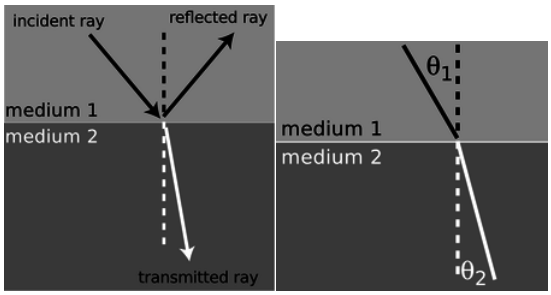


***Figure* 13.4.4:** *The incident, reflected, and transmitted (refracted) rays all lie in a plane that includes the normal (dashed line).*
***Figure* 13.4.5:** *The angles θ_1 and θ_2 are related to each other, and also depend on the properties of the two media. Because refraction is time-reversal symmetric, there is no need to label the rays with arrowheads.*

If you try shooting a beam of light at the boundary between two substances, say water and air, you'll find that regardless of the angle at which you send in the beam, the part of the beam in the water is always closer to the normal line, Figure 13.4.5

It doesn't matter if the ray is entering the water or leaving, so refraction is symmetric with respect to time-reversal, Figure 13.4.6

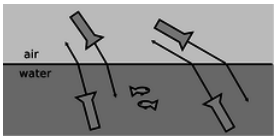


***Figure* 13.4.6:** Refraction has time-reversal symmetry. Regardless of whether the light is going into or out of the water, the relationship between the two angles is the same, and the ray is closer to the normal while in the water.

If, instead of water and air, you try another combination of substances, say plastic and gasoline, again you'll find that the ray's angle with respect to the normal is consistently smaller in one and larger in the other. Also, we find that if substance A has rays closer to normal than in B, and B has rays closer to normal than in C, then A has rays closer to normal than C. This means that we can rank-order all materials according to their refractive properties. Isaac Newton did so, including in his list many amusing substances, such as “Danzig vitriol” and “a pseudo-topazius, being a natural, pellucid, brittle, hairy stone, of a yellow color.” Several general rules can be inferred from such a list:

- Vacuum lies at one end of the list. In refraction across the interface between vacuum and any other medium, the other medium has rays closer to the normal.
- Among gases, the ray gets closer to the normal if you increase the density of the gas by pressurizing it more.

- The refractive properties of liquid mixtures and solutions vary in a smooth and systematic manner as the proportions of the mixture are changed.
- Denser substances usually, but not always, have rays closer to the normal.

The second and third rules provide us with a method for measuring the density of an unknown sample of gas, or the concentration of a solution. The latter technique is very commonly used, and the CRC Handbook of Physics and Chemistry, for instance, contains extensive tables of the refractive properties of sugar solutions, cat urine, and so on.

Snell's Law

The numerical rule governing refraction was discovered by Snell, who must have collected experimental data something like what is shown on this graph and then attempted by trial and error to find the right equation. The equation he came up with was

$$\frac{\sin\theta_1}{\sin\theta_2} = \text{constant}. \tag{13.4.1}$$

The value of the constant would depend on the combination of media used. For instance, any one of the data points in the graph would have sufficed to show that the constant was 1.3 for an air-water interface (taking air to be substance 1 and water to be substance 2).

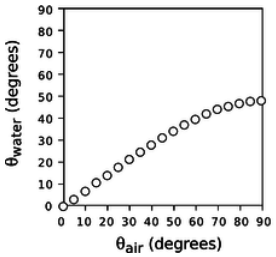


Figure* 13.4.7*: The relationship between the angles in refraction.

Snell further found that if media A and B gave a constant K_{AB} and media B and C gave a constant K_{BC}, then refraction at an interface between A and C would be described by a constant equal to the product, $K_{AC} = K_{AB}K_{BC}$. This is exactly what one would expect if the constant depended on the ratio of some number characterizing one medium to the number characteristic of the second medium. This number is called the *index of refraction* of the medium, written as n in equations. Since measuring the angles would only allow him to determine the *ratio* of the indices of refraction of two media, Snell had to pick some medium and define it as having $n = 1$. He chose to define vacuum as having $n = 1$. (The index of refraction of air at normal atmospheric pressure is 1.0003, so for most purposes it is a good approximation to assume that air has $n = 1$.) He also had to decide which way to define the ratio, and he chose to define it so that media with their rays closer to the normal would have larger indices of refraction. This had the advantage that denser media would typically have higher indices of refraction, and for this reason the index of refraction is also referred to as the optical density. Written in terms of indices of refraction, Snell's equation becomes

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_2}{n_1}, \tag{13.4.2}$$

but rewriting it in the form

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{13.4.3}$$

[relationship between angles of rays at the interface between media with indices of refraction n_1 and n_2; angles are defined with respect to the normal] makes us less likely to get the 1's and 2's mixed up, so this the way most people remember Snell's law. A few indices of refraction are given in the back of the book.

Exercise 13.4.1

1. What would the graph look like for two substances with the same index of refraction?
2. Based on the graph, when does refraction at an air-water interface change the direction of a ray most strongly?

(answer in the back of the PDF version of the book)

Example 13.4.1: Finding an angle using Snell's law

A submarine shines its searchlight up toward the surface of the water. What is the angle α shown in Figure 13.4.8?

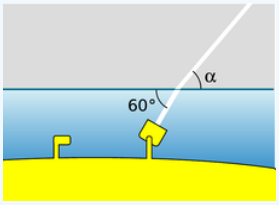


Figure* 13.4.8*: *Example 10.*

Solution

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

$$\theta_1 = \sin^{-1}\left(\frac{n_2}{n_1}\sin\theta_2\right). \tag{13.4.4}$$

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

What neither Snell nor Newton knew was that there is a very simple interpretation of the index of refraction. This may come as a relief to the reader who is taken aback by the complex reasoning involving proportionalities that led to its definition. Later experiments showed that the index of refraction of a medium was inversely proportional to the speed of light in that medium. Since c is defined as the speed of light in vacuum, and $n = 1$ is defined as the index of refraction of vacuum, we have

$$n = \frac{c}{v}. \tag{13.4.5}$$

[$n =$ medium's index of refraction, $v =$ speed of light in that medium, $c =$ speed of light in a vacuum]

Many textbooks start with this as the definition of the index of refraction, although that approach makes the quantity's name somewhat of a mystery, and leaves students wondering why c/v was used rather than v/c. It should also be noted that measuring angles of refraction is a far more practical method for determining n than direct measurement of the speed of light in the substance of interest.

A mechanical model of Snell's law

Why should refraction be related to the speed of light? The mechanical model shown in the figure may help to make this more plausible. Suppose medium 2 is thick, sticky mud, which slows down the car. The car's right wheel hits the mud first, causing the right side of the car to slow down. This will cause the car to turn to the right until is moves far enough forward for the left wheel to cross into the mud. After that, the two sides of the car will once again be moving at the same speed, and the car will go straight.

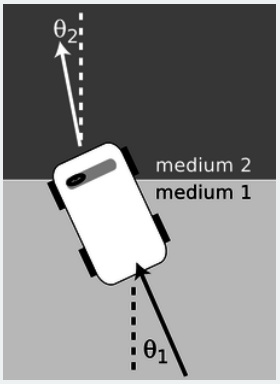


Figure **13.4.8:** A mechanical model of refraction.

Of course, light isn't a car. Why should a beam of light have anything resembling a “left wheel” and “right wheel?” After all, the mechanical model would predict that a motorcycle would go straight, and a motorcycle seems like a better approximation to a ray of light than a car. The whole thing is just a model, not a description of physical reality.

A Derivation of Snell's law

However intuitively appealing the mechanical model may be, light is a wave, and we should be using wave models to describe refraction. In fact Snell's law can be derived quite simply from wave concepts. Figure 13.4.9 shows the refraction of a water wave. The water in the upper left part of the tank is shallower, so the speed of the waves is slower there, and their wavelengths is shorter. The reflected part of the wave is also very faintly visible.

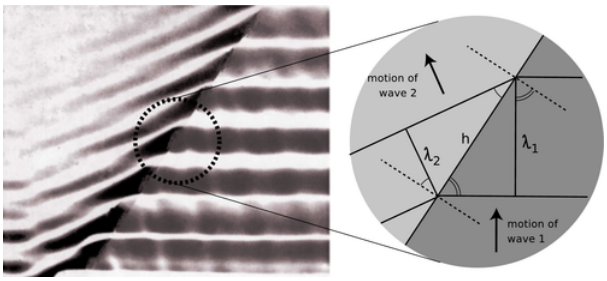


Figure **13.4.9:** A derivation of Snell's law.

In the close-up view on the right, the dashed lines are normals to the interface. The two marked angles on the right side are both equal to θ_1, and the two on the left to θ_2.

Trigonometry gives

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

Eliminating h by dividing the equations, we find

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{\lambda_1}{\lambda_2}. \tag{13.4.6}$$

The frequencies of the two waves must be equal or else they would get out of step, so by $v = f\lambda$ we know that their wavelengths are proportional to their velocities. Combining $\lambda \propto v$ with $v \propto 1/n$ gives $\lambda \propto 1/n$, so we find

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_2}{n_1}, \tag{13.4.7}$$

which is one form of Snell's law.

Example 13.4.1: Ocean waves near and far from shore

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

Color and Refraction

In general, the speed of light in a medium depends both on the medium and on the wavelength of the light. Another way of saying it is that a medium's index of refraction varies with wavelength. This is why a prism can be used to split up a beam of white light into a rainbow. Each wavelength of light is refracted through a different angle.

How much light is reflected, and how much is transmitted?

In section 6.2 we developed an equation for the percentage of the wave energy that is transmitted and the percentage reflected at a boundary between media. This was only done in the case of waves in one dimension, however, and rather than discuss the full three dimensional generalization it will be more useful to go into some qualitative observations about what happens. First, reflection happens only at the interface between two media, and two media with the same index of refraction act as if they were a single medium. Thus, at the interface between media with the same index of refraction, there is no reflection, and the ray keeps going straight. Continuing this line of thought, it is not surprising that we observe very little reflection at an interface between media with similar indices of refraction.

The next thing to note is that it is possible to have situations where no possible angle for the refracted ray can satisfy Snell's law. Solving Snell's law for θ_2, we find

$$\theta_2 = \sin^{-1}\left(\frac{n_1}{n_2}\sin\theta_1\right), \tag{13.4.8}$$

and if n_1 is greater than n_2, then there will be large values of θ_1 for which the quantity $(n_1/n_2)\sin\theta$ is greater than one, meaning that your calculator will flash an error message at you when you try to take the inverse sine. What can happen physically in such a situation? The answer is that all the light is reflected, so there is no refracted ray. This phenomenon is known as *total internal reflection*, and is used in the fiber-optic cables that nowadays carry almost all long-distance telephone calls.

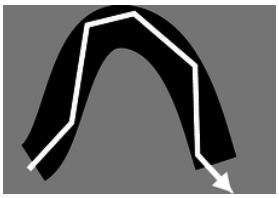


***Figure* 13.4.10:** *Total internal reflection in a fiber-optic cable.*

The electrical signals from your phone travel to a switching center, where they are converted from electricity into light. From there, the light is sent across the country in a thin transparent fiber. The light is aimed straight into the end of the fiber, and as long as the fiber never goes through any turns that are too sharp, the light will always encounter the edge of the fiber at an angle sufficiently oblique to give total internal reflection. If the fiber-optic cable is thick enough, one can see an image at one end of whatever the other end is pointed at.

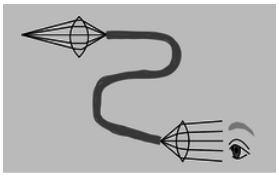


***Figure* 13.4.11:** *A simplified drawing of a surgical endoscope. The first lens forms a real image at one end of a bundle of optical fibers. The light is transmitted through the bundle, and is finally magnified by the eyepiece.*

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

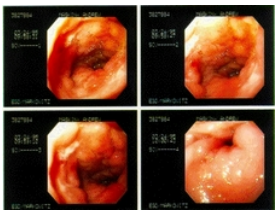


***Figure* 13.4.12:** Endoscopic images of a duodenal ulcer.

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

Discussion Questions

◊ What index of refraction should a fish have in order to be invisible to other fish?

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

◊ A denser sample of a gas has a higher index of refraction than a less dense sample (i.e., a sample under lower pressure), but why would it not make sense for the index of refraction of a gas to be proportional to density?

◊ The earth's atmosphere gets thinner and thinner as you go higher in altitude. If a ray of light comes from a star that is below the zenith, what will happen to it as it comes into the earth's atmosphere?

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

12.4.2 Lenses

Figures n/1 and n/2 show examples of lenses forming images. There is essentially nothing for you to learn about imaging with lenses that is truly new. You already know how to construct and use ray diagrams, and you know about real and virtual images. The concept of the focal length of a lens is the same as for a curved mirror. The equations for locating images and determining magnifications are of the same form. It's really just a question of flexing your mental muscles on a few examples. The following self-checks and discussion questions will get you started.

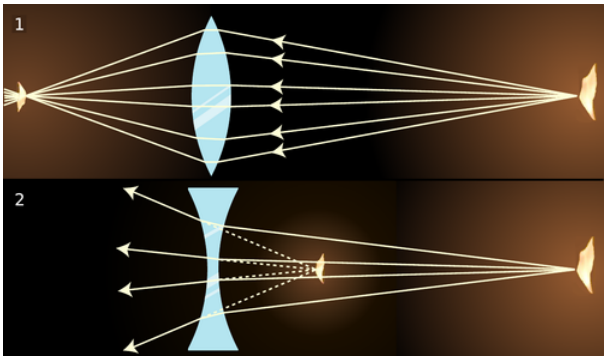


***Figure* 13.4.13:** 1. A converging lens forms an image of a candle flame. 2. A diverging lens.

Exercise 13.4.1

1. In figures Figure $13.4.13; part 1$and Figure $13.4.13; part 2$ classify the images as real or virtual.
2. Glass has an index of refraction that is greater than that of air. Consider the topmost ray in Figure $13.4.13; part 1$ Explain why the ray makes a slight left turn upon entering the lens, and another left turn when it exits.
3. If the flame in Figure $13.4.13; part 2$were moved closer to the lens, what would happen to the location of the image?

Answer

(answer in the back of the PDF version of the book)

Discussion Questions

◊ In figures n/1 and n/2, the front and back surfaces are parallel to each other at the center of the lens. What will happen to a ray that enters near the center, but not necessarily along the axis of the lens? Draw a BIG ray diagram, and show a ray that comes from off axis.

In discussion questions B-F, don't draw ultra-detailed ray diagrams as in A.

◊ Suppose you wanted to change the setup in figure n/1 so that the location of the actual flame in the figure would instead be occupied by an image of a flame. Where would you have to move the candle to achieve this? What about in n/2?

◊ There are three qualitatively different types of image formation that can occur with lenses, of which figures n/1 and n/2 exhaust only two. Figure out what the third possibility is. Which of the three possibilities can result in a magnification greater than one? Cf.

problem 10, p. 797.

◊ Classify the examples shown in figure o according to the types of images delineated in discussion question C.

◊ In figures n/1 and n/2, the only rays drawn were those that happened to enter the lenses. Discuss this in relation to figure o.

◊ In the right-hand side of figure o, the image viewed through the lens is in focus, but the side of the rose that sticks out from behind the lens is not. Why?

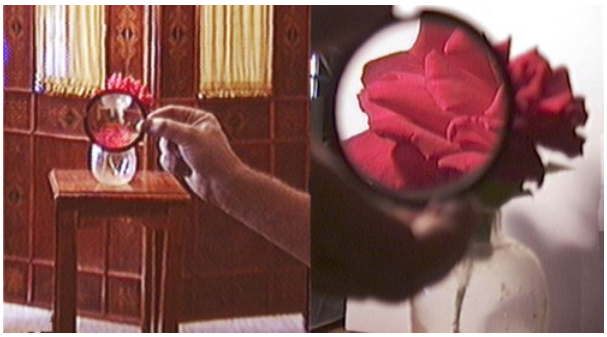


***Figure* 13.4.14:** Two images of a rose created by the same lens and recorded with the same camera.

The Lensmaker's Equation

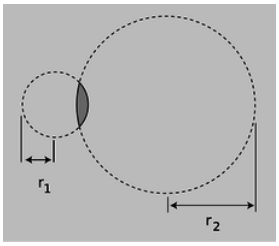


***Figure* 13.4.15:** The radii of curvature appearing in the lensmaker's equation.

The focal length of a spherical mirror is simply $r/2$, but we cannot expect the focal length of a lens to be given by pure geometry, since it also depends on the index of refraction of the lens. Suppose we have a lens whose front and back surfaces are both spherical. (This is no great loss of generality, since any surface with a sufficiently shallow curvature can be approximated with a sphere.) Then if the lens is immersed in a medium with an index of refraction of 1, its focal length is given approximately by

$$f = \frac{1}{(n-1)\left|\frac{1}{r_1} \pm \frac{1}{r_2}\right|} \tag{13.4.9}$$

where n is the index of refraction and r_1 and r_2 are the radii of curvature of the two surfaces of the lens. Equation 13.4.9 is known as the *lensmaker's equation*. In my opinion it is not particularly worthy of memorization. The positive sign is used when both surfaces are curved outward or both are curved inward; otherwise a negative sign applies. The proof of this equation is left as an exercise to those readers who are sufficiently brave and motivated.

12.4.3 Dispersion

For most materials, we observe that the index of refraction depends slightly on wavelength, being highest at the blue end of the visible spectrum and lowest at the red. For example, white light disperses into a rainbow when it passes through a prism, q.

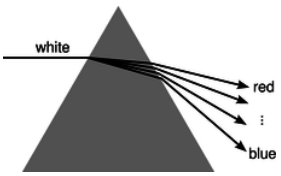


***Figure* 13.4.16:** Dispersion of white light by a prism. White light is a mixture of all the wavelengths of the visible spectrum. Waves of different wavelengths undergo different amounts of refraction.

Even when the waves involved aren't light waves, and even when refraction isn't of interest, the dependence of wave speed on wavelength is referred to as dispersion. Dispersion inside spherical raindrops is responsible for the creation of rainbows in the sky, and in an optical instrument such as the eye or a camera it is responsible for a type of aberration called chromatic aberration (subsection 12.3.3 and problem 28). As we'll see in subsection 13.3.2, dispersion causes a wave that is not a pure sine wave to have its shape distorted as it travels, and also causes the speed at which energy and information are transported by the wave to be different from what one might expect from a naive calculation. The microscopic reasons for dispersion of light in matter are discussed in optional subsection 12.4.6.

The principle of least time for refraction

We have seen previously how the rules governing straight-line motion of light and reflection of light can be derived from the principle of least time. What about refraction? In the figure, it is indeed plausible that the bending of the ray serves to minimize the time required to get from a point A to point B. If the ray followed the unbent path shown with a dashed line, it would have to travel a longer distance in the medium in which its speed is slower. By bending the correct amount, it can reduce the distance it has to cover in the slower medium without going too far out of its way. It is true that Snell's law gives exactly the set of angles that minimizes the time required for light to get from one point to another. The proof of this fact is left as an exercise (problem 38, p. 802).

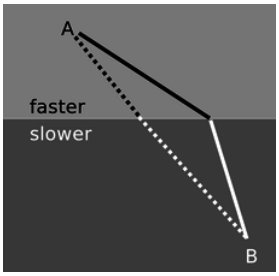


***Figure* 13.4.17:** The principle of least time applied to refraction.

Microscopic description of refraction

Given that the speed of light is different in different media, we've seen two different explanations (on p. 774 and in subsection 12.4.5 above) of why refraction must occur. What we haven't yet explained is why the speed of light does depend on the medium.

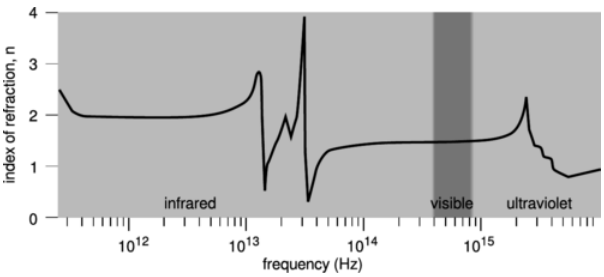


***Figure* 13.4.18:** Index of refraction of silica glass, redrawn from Kitamura, Pilon, and Jonasz, Applied Optics 46 (2007) 8118, reprinted online at http://www.seas.ucla.edu/~pilon/Publ...s/AO2007-1.pdf.

A good clue as to what's going on comes from the figure s. The relatively minor variation of the index of refraction within the visible spectrum was misleading. At certain specific frequencies, n exhibits wild swings in the positive and negative directions. After each such swing, we reach a new, lower plateau on the graph. These frequencies are resonances. For example, the visible part of the spectrum lies on the left-hand tail of a resonance at about 2×10^{15} Hz , corresponding to the ultraviolet part of the spectrum. This resonance arises from the vibration of the electrons, which are bound to the nuclei as if by little springs. Because this resonance is narrow, the effect on visible-light frequencies is relatively small, but it is stronger at the blue end of the spectrum than at the red end. Near each resonance, not only does the index of refraction fluctuate wildly, but the glass becomes nearly opaque; this is because the vibration becomes very strong, causing energy to be dissipated as heat. The “staircase” effect is the same one visible in any resonance, e.g., figure k on p. 180: oscillators have a finite response for $f \ll f_0$, but the response approaches zero for $f \gg f_0$.

So far, we have a qualitative explanation of the frequency-variation of the loosely defined “strength” of the glass's effect on a light wave, but we haven't explained why the effect is observed as a change in speed, or why each resonance is an up-down swing rather than a single positive peak. To understand these effects in more detail, we need to consider the phase response of the oscillator. As shown in the bottom panel of figure j on p. 181, the phase response reverses itself as we pass through a resonance.

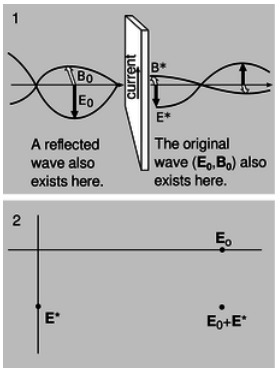


***Figure* 13.4.19:** 1. A wave incident on a sheet of glass excites current in the glass, which produce a secondary wave. 2. The secondary wave superposes with the original wave, as represented in the complex-number representation introduced in subsection 10.5.7.

Suppose that a plane wave is normally incident on the left side of a thin sheet of glass, t/1, at $f \ll f_0$. The light wave observed on the right side consists of a superposition of the incident wave consisting of $\mathbf{E}_0$ and $\mathbf{B}_0$ with a secondary wave $\mathbf{E}^*$ and $\mathbf{B}^*$ generated by the oscillating charges in the glass. Since the frequency is far below resonance, the response $q\mathbf{x}$ of a vibrating charge q is in phase with the driving force $\mathbf{E}_0$. The current is the derivative of this quantity, and therefore 90 degrees ahead of it in phase. The magnetic field generated by a sheet of current has been analyzed in subsection 11.2.1, and the result, shown in figure e on p. 664, is just what we would expect from the right-hand rule. We find, t/1, that the secondary wave is 90 degrees ahead of the incident one in phase. The incident wave still exists on the right side of the sheet, but it is superposed with the secondary one. Their addition is shown in t/2 using the complex number representation introduced in subsection 10.5.7. The superposition of the two fields lags behind the incident wave, which is the effect we would expect if the wave had traveled more slowly through the glass.

In the case $f \gg 0$, the same analysis applies except that the phase of the secondary wave is reversed. The transmitted wave is advanced rather than retarded in phase. This explains the dip observed in figure s after each spike.

All of this is in accord with our understanding of relativity, ch. 7, in which we saw that the universal speed c was to be understood fundamentally as a conversion factor between the units used to measure time and space --- not as the speed of light. Since c isn't defined as the speed of light, it's of no fundamental importance whether light has a different speed in matter than it does in vacuum. In fact, the picture we've built up here is one in which all of our electromagnetic waves travel at c; propagation at some other speed is only what appears to happen because of the superposition of the $(\mathbf{E}_0, \mathbf{B}_0)$ and $(\mathbf{E}^*, \mathbf{B}^*)$ waves, both of which move at c.

But it is worrisome that at the frequencies where $n < 1$, the speed of the wave is greater than c. According to special relativity, information is never supposed to be transmitted at speeds greater than c, since this would produce situations in which a signal could be received before it was transmitted! This difficulty is resolved in subsection 13.3.2, where we show that there are two different velocities that can be defined for a wave in a dispersive medium, the phase velocity and the group velocity. The group velocity is the velocity at which information is transmitted, and it is always less than c.

Contributors and Attributions

13.5: Wave Optics

Electron microscopes can make images of individual atoms, but why will a visible-light microscope never be able to? Stereo speakers create the illusion of music that comes from a band arranged in your living room, but why doesn't the stereo illusion work with bass notes? Why are computer chip manufacturers investing billions of dollars in equipment to etch chips with x-rays instead of visible light?

The answers to all of these questions have to do with the subject of wave optics. So far this book has discussed the interaction of light waves with matter, and its practical applications to optical devices like mirrors, but we have used the ray model of light almost exclusively. Hardly ever have we explicitly made use of the fact that light is an electromagnetic wave. We were able to get away with the simple ray model because the chunks of matter we were discussing, such as lenses and mirrors, were thousands of times larger than a wavelength of light. We now turn to phenomena and devices that can only be understood using the wave model of light.

12.5.1 Diffraction

Figure Figure 13.5.1*a* shows a typical problem in wave optics, enacted with water waves. It may seem surprising that we don't get a simple pattern like Figure 13.5.1*b*, but the pattern would only be that simple if the wavelength was hundreds of times shorter than the distance between the gaps in the barrier and the widths of the gaps.

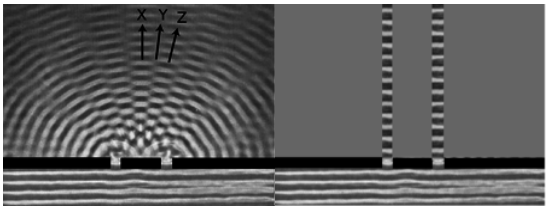


***Figure* 13.5.1** *: a. In this view from overhead, a straight, sinusoidal water wave encounters a barrier with two gaps in it. Strong wave vibration occurs at angles X and Z, but there is none at all at angle Y. (The figure has been retouched from a real photo of water waves. In reality, the waves beyond the barrier would be much weaker than the ones before it, and they would therefore be difficult to see.) b / This doesn't happen.*

Wave optics is a broad subject, but this example will help us to pick out a reasonable set of restrictions to make things more manageable:

1. We restrict ourselves to cases in which a wave travels through a uniform medium, encounters a certain area in which the medium has different properties, and then emerges on the other side into a second uniform region.
2. We assume that the incoming wave is a nice tidy sine-wave pattern with wavefronts that are lines (or, in three dimensions, planes).
3. In Figure 13.5.1*a* we can see that the wave pattern immediately beyond the barrier is rather complex, but farther on it sorts itself out into a set of wedges separated by gaps in which the water is still. We will restrict ourselves to studying the simpler wave patterns that occur farther away, so that the main question of interest is how intense the outgoing wave is at a given angle.

The kind of phenomenon described by restriction (1) is called *diffraction*. Diffraction can be defined as the behavior of a wave when it encounters an obstacle or a nonuniformity in its medium. In general, diffraction causes a wave to bend around obstacles and make patterns of strong and weak waves radiating out beyond the obstacle. Understanding diffraction is the central problem of wave optics. If you understand diffraction, even the subset of diffraction problems that fall within restrictions (2) and (3), the rest of wave optics is icing on the cake.

Diffraction can be used to find the structure of an unknown diffracting object: even if the object is too small to study with ordinary imaging, it may be possible to work backward from the diffraction pattern to learn about the object. The structure of a crystal, for example, can be determined from its x-ray diffraction pattern.

Diffraction can also be a bad thing. In a telescope, for example, light waves are diffracted by all the parts of the instrument. This will cause the image of a star to appear fuzzy even when the focus has been adjusted correctly. By understanding diffraction, one can learn how a telescope must be designed in order to reduce this problem --- essentially, it should have the biggest possible diameter.

There are two ways in which restriction (2) might commonly be violated. First, the light might be a mixture of wavelengths. If we simply want to observe a diffraction pattern or to use diffraction as a technique for studying the object doing the diffracting (e.g., if the object is too small to see with a microscope), then we can pass the light through a colored filter before diffracting it.

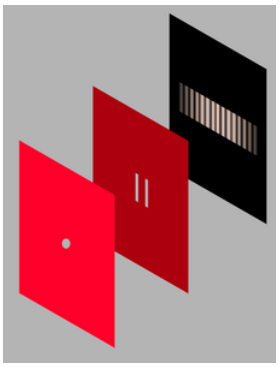
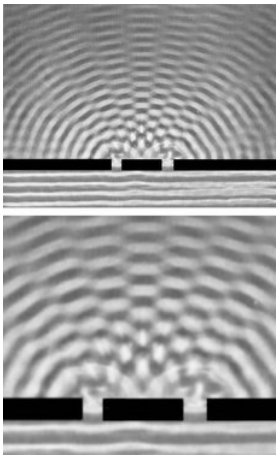


***Figure* 13.5.2:** *A practical, low-tech setup for observing diffraction of light.*

A second issue is that light from sources such as the sun or a lightbulb does not consist of a nice neat plane wave, except over very small regions of space. Different parts of the wave are out of step with each other, and the wave is referred to as *incoherent*. One way of dealing with this is shown in Figure 13.5.2 After filtering to select a certain wavelength of red light, we pass the light through a small pinhole. The region of the light that is intercepted by the pinhole is so small that one part of it is not out of step with another. Beyond the pinhole, light spreads out in a spherical wave; this is analogous to what happens when you speak into one end of a paper towel roll and the sound waves spread out in all directions from the other end. By the time the spherical wave gets to the double slit it has spread out and reduced its curvature, so that we can now think of it as a simple plane wave.



d / The bottom figure is simply a copy of the middle portion of the top one, scaled up by a factor of two. All the angles are the same. Physically, the angular pattern of the diffraction fringes can't be any different if we scale both λ and d by the same factor, leaving λ/d unchanged.

If this seems laborious, you may be relieved to know that modern technology gives us an easier way to produce a single-wavelength, coherent beam of light: the laser.

The parts of the final image on the screen in Figure 13.5.2 are called *diffraction fringes*. The center of each fringe is a point of maximum brightness, and halfway between two fringes is a minimum.

Exercise 13.5.1: Discussion Question

Why would x-rays rather than visible light be used to find the structure of a crystal? Sound waves are used to make images of fetuses in the womb. What would influence the choice of wavelength

12.5.2 Scaling of diffraction

This chapter has “optics” in its title, so it is nominally about light, but we started out with an example involving water waves. Water waves are certainly easier to visualize, but is this a legitimate comparison? In fact the analogy works quite well, despite the fact that a light wave has a wavelength about a million times shorter. This is because diffraction effects scale uniformly. That is, if we enlarge or reduce the whole diffraction situation by the same factor, including both the wavelengths and the sizes of the obstacles the wave encounters, the result is still a valid solution.

This is unusually simple behavior! In subsection 0.2.2 we saw many examples of more complex scaling, such as the impossibility of bacteria the size of dogs, or the need for an elephant to eliminate heat through its ears because of its small surface-to-volume ratio, whereas a tiny shrew's life-style centers around conserving its body heat.

Of course water waves and light waves differ in many ways, not just in scale, but the general facts you will learn about diffraction are applicable to all waves. In some ways it might have been more appropriate to insert this chapter after section 6.2 on bounded waves, but many of the important applications are to light waves, and you would probably have found these much more difficult without any background in optics.

Another way of stating the simple scaling behavior of diffraction is that the diffraction angles we get depend only on the unitless ratio λ/d, where λ is the wavelength of the wave and d is some dimension of the diffracting objects, e.g., the center-to-center spacing between the slits in figure a. If, for instance, we scale up both λ and d by a factor of 37, the ratio λ/d will be unchanged.

12.5.3 The Correspondence Principle

The only reason we don't usually notice diffraction of light in everyday life is that we don't normally deal with objects that are comparable in size to a wavelength of visible light, which is about a millionth of a meter. Does this mean that wave optics contradicts ray optics, or that wave optics sometimes gives wrong results? No. If you hold three fingers out in the sunlight and cast a shadow with them, *either* wave optics or ray optics can be used to predict the straightforward result: a shadow pattern with two bright lines where the light has gone through the gaps between your fingers. Wave optics is a more general theory than ray optics, so in any case where ray optics is valid, the two theories will agree. This is an example of a general idea enunciated by the physicist Niels Bohr, called the *correspondence principle:* when flaws in a physical theory lead to the creation of a new and more general theory, the new theory must still agree with the old theory within its more restricted area of applicability. After all, a theory is only created as a way of describing experimental observations. If the original theory had not worked in any cases at all, it would never have become accepted.

In the case of optics, the correspondence principle tells us that when λ/d is small, both the ray and the wave model of light must give approximately the same result. Suppose you spread your fingers and cast a shadow with them using a coherent light source. The quantity λ/d is about 10^{-4}, so the two models will agree very closely. (To be specific, the shadows of your fingers will be outlined by a series of light and dark fringes, but the angle subtended by a fringe will be on the order of 10^{-4} radians, so they will be invisible and washed out by the natural fuzziness of the edges of sun-shadows, caused by the finite size of the sun.)

Exercise 13.5.1

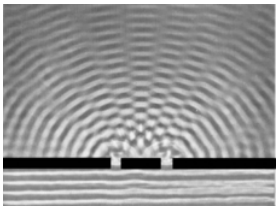
What kind of wavelength would an electromagnetic wave have to have in order to diffract dramatically around your body? Does this contradict the correspondence principle?

Answer

(answer in the back of the PDF version of the book)

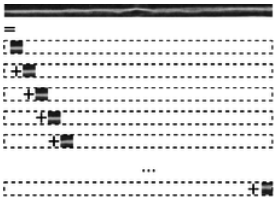
12.5.4 Huygens' Principle

e / Christiaan Huygens (1629-1695).



f / Double-slit diffraction.

Returning to the example of double-slit diffraction, f, note the strong visual impression of two overlapping sets of concentric semicircles. This is an example of *Huygens' principle*, named after a Dutch physicist and astronomer. (The first syllable rhymes with “boy.”) Huygens' principle states that any wavefront can be broken down into many small side-by-side wave peaks, g, which then spread out as circular ripples, h, and by the principle of superposition, the result of adding up these sets of ripples must give the same result as allowing the wave to propagate forward, i.



g / A wavefront can be analyzed by the principle of superposition, breaking it down into many small parts.

h / If it was by itself, each of the parts would spread out as a circular ripple.

i / Adding up the ripples produces a new wavefront.

In the case of sound or light waves, which propagate in three dimensions, the “ripples” are actually spherical rather than circular, but we can often imagine things in two dimensions for simplicity.

In double-slit diffraction the application of Huygens' principle is visually convincing: it is as though all the sets of ripples have been blocked except for two. It is a rather surprising mathematical fact, however, that Huygens' principle gives the right result in the case of an unobstructed linear wave, h and i. A theoretically infinite number of circular wave patterns somehow conspire to add together and produce the simple linear wave motion with which we are familiar.

Since Huygens' principle is equivalent to the principle of superposition, and superposition is a property of waves, what Huygens had created was essentially the first wave theory of light. However, he imagined light as a series of pulses, like hand claps, rather

than as a sinusoidal wave.

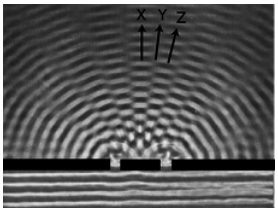
The history is interesting. Isaac Newton loved the atomic theory of matter so much that he searched enthusiastically for evidence that light was also made of tiny particles. The paths of his light particles would correspond to rays in our description; the only significant difference between a ray model and a particle model of light would occur if one could isolate individual particles and show that light had a “graininess” to it. Newton never did this, so although he thought of his model as a particle model, it is more accurate to say he was one of the builders of the ray model.

Almost all that was known about reflection and refraction of light could be interpreted equally well in terms of a particle model or a wave model, but Newton had one reason for strongly opposing Huygens' wave theory. Newton knew that waves exhibited diffraction, but diffraction of light is difficult to observe, so Newton believed that light did not exhibit diffraction, and therefore must not be a wave. Although Newton's criticisms were fair enough, the debate also took on the overtones of a nationalistic dispute between England and continental Europe, fueled by English resentment over Leibniz's supposed plagiarism of Newton's calculus. Newton wrote a book on optics, and his prestige and political prominence tended to discourage questioning of his model.

j / Thomas Young

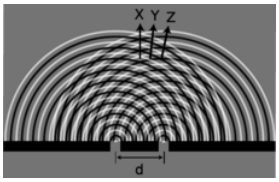
Thomas Young (1773-1829) was the person who finally, a hundred years later, did a careful search for wave interference effects with light and analyzed the results correctly. He observed double-slit diffraction of light as well as a variety of other diffraction effects, all of which showed that light exhibited wave interference effects, and that the wavelengths of visible light waves were extremely short. The crowning achievement was the demonstration by the experimentalist Heinrich Hertz and the theorist James Clerk Maxwell that light was an *electromagnetic* wave. Maxwell is said to have related his discovery to his wife one starry evening and told her that she was the only other person in the world who knew what starlight was.

12.5.5 Double-slit diffraction



k / Double-slit diffraction.

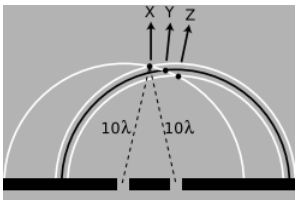
Let's now analyze double-slit diffraction, k, using Huygens' principle. The most interesting question is how to compute the angles such as X and Z where the wave intensity is at a maximum, and the in-between angles like Y where it is minimized. Let's measure all our angles with respect to the vertical center line of the figure, which was the original direction of propagation of the wave.



l / Use of Huygens' principle.

If we assume that the width of the slits is small (on the order of the wavelength of the wave or less), then we can imagine only a single set of Huygens ripples spreading out from each one, l. White lines represent peaks, black ones troughs. The only dimension

of the diffracting slits that has any effect on the geometric pattern of the overlapping ripples is then the center-to-center distance, d, between the slits.



m / Constructive interference along the center-line.

We know from our discussion of the scaling of diffraction that there must be some equation that relates an angle like θ_Z to the ratio λ/d,

$$\frac{\lambda}{d} \leftrightarrow \theta_Z .$$

If the equation for θ_Z depended on some other expression such as $\lambda + d$ or λ^2/d, then it would change when we scaled λ and d by the same factor, which would violate what we know about the scaling of diffraction.

Along the central maximum line, X, we always have positive waves coinciding with positive ones and negative waves coinciding with negative ones. (I have arbitrarily chosen to take a snapshot of the pattern at a moment when the waves emerging from the slit are experiencing a positive peak.) The superposition of the two sets of ripples therefore results in a doubling of the wave amplitude along this line. There is constructive interference. This is easy to explain, because by symmetry, each wave has had to travel an equal number of wavelengths to get from its slit to the center line, m: Because both sets of ripples have ten wavelengths to cover in order to reach the point along direction X, they will be in step when they get there.

At the point along direction Y shown in the same figure, one wave has traveled ten wavelengths, and is therefore at a positive extreme, but the other has traveled only nine and a half wavelengths, so it at a negative extreme. There is perfect cancellation, so points along this line experience no wave motion.

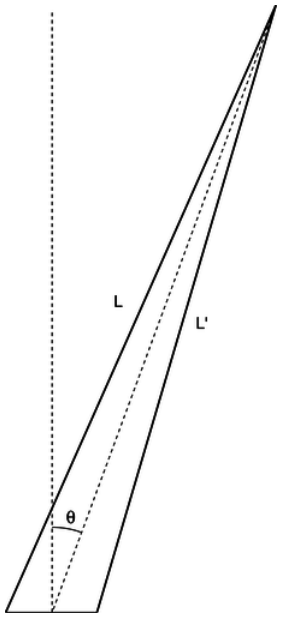
But the distance traveled does not have to be equal in order to get constructive interference. At the point along direction Z, one wave has gone nine wavelengths and the other ten. They are both at a positive extreme.

self-check:

At a point half a wavelength below the point marked along direction X, carry out a similar analysis.

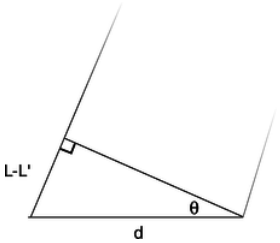
(answer in the back of the PDF version of the book)

To summarize, we will have perfect constructive interference at any point where the distance to one slit differs from the distance to the other slit by an integer number of wavelengths. Perfect destructive interference will occur when the number of wavelengths of path length difference equals an integer plus a half.



n / The waves travel distances L and L' from the two slits to get to the same point in space, at an angle θ from the center line.

Now we are ready to find the equation that predicts the angles of the maxima and minima. The waves travel different distances to get to the same point in space, n. We need to find whether the waves are in phase (in step) or out of phase at this point in order to predict whether there will be constructive interference, destructive interference, or something in between.



o / A close-up view of figure n, showing how the path length difference $L-L'$ is related to d and to the angle θ.

One of our basic assumptions in this chapter is that we will only be dealing with the diffracted wave in regions very far away from the object that diffracts it, so the triangle is long and skinny. Most real-world examples with diffraction of light, in fact, would have triangles with even skinner proportions than this one. The two long sides are therefore very nearly parallel, and we are justified in drawing the right triangle shown in figure o, labeling one leg of the right triangle as the difference in path length , $L-L'$, and labeling the acute angle as θ. (In reality this angle is a tiny bit greater than the one labeled θ in figure n.)

The difference in path length is related to d and θ by the equation

$$\frac{L-L'}{d} = \sin\theta.$$

Constructive interference will result in a maximum at angles for which $L-L'$ is an integer number of wavelengths,

$$L-L' = m\lambda.$$

[condition for a maximum; m is an integer]

Here m equals 0 for the central maximum, -1 for the first maximum to its left, $+2$ for the second maximum on the right, etc. Putting all the ingredients together, we find $m\lambda/d = \sin\theta$, or

$$\frac{\lambda}{d} = \frac{\sin\theta}{m}.$$

[condition for a maximum; m is an integer]

Similarly, the condition for a minimum is

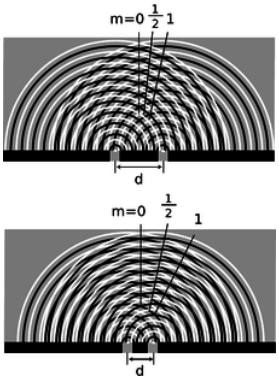
$$\frac{\lambda}{d} = \frac{\sin\theta}{m}.$$

[condition for a minimum; m is an integer plus $1/2$]

That is, the minima are about halfway between the maxima.

As expected based on scaling, this equation relates angles to the unitless ratio λ/d. Alternatively, we could say that we have proven the scaling property in the special case of double-slit diffraction. It was inevitable that the result would have these scaling properties, since the whole proof was geometric, and would have been equally valid when enlarged or reduced on a photocopying machine!

Counterintuitively, this means that a diffracting object with smaller dimensions produces a bigger diffraction pattern, p.



p / Cutting d in half doubles the angles of the diffraction fringes.

Example 12: Double-slit diffraction of blue and red light

Blue light has a shorter wavelength than red. For a given double-slit spacing d, the smaller value of λ/d for leads to smaller values of $\sin\theta$, and therefore to a more closely spaced set of diffraction fringes, q

q / Double-slit diffraction patterns of long-wavelength red light (top) and short-wavelength blue light (bottom).

Example 13: The correspondence principle

Let's also consider how the equations for double-slit diffraction relate to the correspondence principle. When the ratio λ/d is very small, we should recover the case of simple ray optics. Now if λ/d is small, $\sin\theta$ must be small as well, and the spacing between the diffraction fringes will be small as well. Although we have not proven it, the central fringe is always the brightest, and the fringes get dimmer and dimmer as we go farther from it. For small values of λ/d, the part of the diffraction pattern that is bright enough to be detectable covers only a small range of angles. This is exactly what we would expect from ray optics: the rays passing through the two slits would remain parallel, and would continue moving in the $\theta = 0$ direction. (In fact there would be images of the two separate slits on the screen, but our analysis was all in terms of angles, so we should not expect it to address the issue of whether there is structure within a set of rays that are all traveling in the $\theta = 0$ direction.)

Example 14: Spacing of the fringes at small angles

At small angles, we can use the approximation $\sin\theta \approx \theta$, which is valid if θ is measured in radians. The equation for double-slit diffraction becomes simply

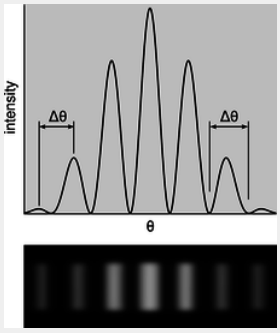
$$\frac{\lambda}{d} = \frac{\theta}{m},$$

which can be solved for θ to give

$$\theta = \frac{m\lambda}{d}.$$

The difference in angle between successive fringes is the change in θ that results from changing m by plus or minus one,

$$\Delta\theta = \frac{\lambda}{d}.$$



r / Interpretation of the angular spacing $\Delta\theta$ in example 14. It can be defined either from maximum to maximum or from minimum to minimum. Either way, the result is the same. It does not make sense to try to interpret $\Delta\theta$ as the width of a fringe; one can see from the graph and from the image below that it is not obvious either that such a thing is well defined or that it would be the same for all fringes.

For example, if we write θ_7 for the angle of the seventh bright fringe on one side of the central maximum and θ_8 for the neighboring one, we have

$$\begin{aligned}\theta_8 - \theta_7 &= \frac{8\lambda}{d} - \frac{7\lambda}{d} \\ &= \frac{\lambda}{d},\end{aligned}$$

and similarly for any other neighboring pair of fringes.

Although the equation $\lambda/d = \sin\theta/m$ is only valid for a double slit, it is can still be a guide to our thinking even if we are observing diffraction of light by a virus or a flea's leg: it is always true that

(1) large values of λ/d lead to a broad diffraction pattern, and

(2) diffraction patterns are repetitive.

In many cases the equation looks just like $\lambda/d = \sin\theta/m$ but with an extra numerical factor thrown in, and with d interpreted as some other dimension of the object, e.g., the diameter of a piece of wire.

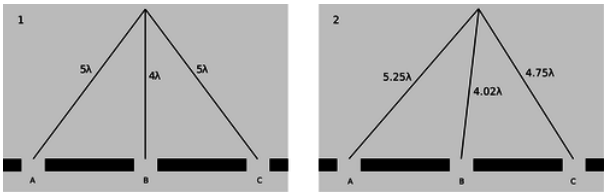
12.5.6 Repetition

s / A triple slit.

Suppose we replace a double slit with a triple slit, s. We can think of this as a third repetition of the structures that were present in the double slit. Will this device be an improvement over the double slit for any practical reasons?

The answer is yes, as can be shown using figure u. For ease of visualization, I have violated our usual rule of only considering points very far from the diffracting object. The scale of the drawing is such that a wavelengths is one cm. In u/1, all three waves travel an integer number of wavelengths to reach the same point, so there is a bright central spot, as we would expect from our experience with the double slit. In figure u/2, we show the path lengths to a new point. This point is farther from slit A by a quarter

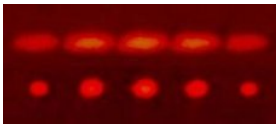
of a wavelength, and correspondingly closer to slit C. The distance from slit B has hardly changed at all. Because the paths lengths traveled from slits A and C differ by half a wavelength, there will be perfect destructive interference between these two waves. There is still some uncanceled wave intensity because of slit B, but the amplitude will be three times less than in figure u/1, resulting in a factor of 9 decrease in brightness. Thus, by moving off to the right a little, we have gone from the bright central maximum to a point that is quite dark.



u / 1. There is a bright central maximum. 2. At this point just off the central maximum, the path lengths traveled by the three waves have changed.

Now let's compare with what would have happened if slit C had been covered, creating a plain old double slit. The waves coming from slits A and B would have been out of phase by 0.23 wavelengths, but this would not have caused very severe interference. The point in figure u/2 would have been quite brightly lit up.

To summarize, we have found that adding a third slit narrows down the central fringe dramatically. The same is true for all the other fringes as well, and since the same amount of energy is concentrated in narrower diffraction fringes, each fringe is brighter and easier to see, t.



t / A double-slit diffraction pattern (top), and a pattern made by five slits (bottom).

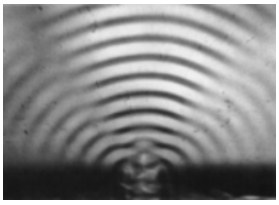
This is an example of a more general fact about diffraction: if some feature of the diffracting object is repeated, the locations of the maxima and minima are unchanged, but they become narrower.

Taking this reasoning to its logical conclusion, a diffracting object with thousands of slits would produce extremely narrow fringes. Such an object is called a diffraction grating.

12.5.7 Single-slit diffraction

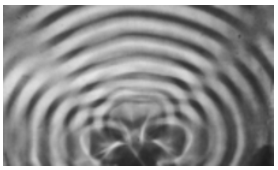
If we use only a single slit, is there diffraction? If the slit is not wide compared to a wavelength of light, then we can approximate its behavior by using only a single set of Huygens ripples. There are no other sets of ripples to add to it, so there are no constructive or destructive interference effects, and no maxima or minima. The result will be a uniform spherical wave of light spreading out in all directions, like what we would expect from a tiny lightbulb. We could call this a diffraction pattern, but it is a completely featureless one, and it could not be used, for instance, to determine the wavelength of the light, as other diffraction patterns could.

All of this, however, assumes that the slit is narrow compared to a wavelength of light. If, on the other hand, the slit is broader, there will indeed be interference among the sets of ripples spreading out from various points along the opening. Figure v shows an example with water waves, and figure w with light.



v / Single-slit diffraction of water waves.

w / Single-slit diffraction of red light. Note the double width of the central maximum.



x / A pretty good simulation of the single-slit pattern of figure v, made by using three motors to produce overlapping ripples from three neighboring points in the water.

self-check:

How does the wavelength of the waves compare with the width of the slit in figure v?

(answer in the back of the PDF version of the book)

We will not go into the details of the analysis of single-slit diffraction, but let us see how its properties can be related to the general things we've learned about diffraction. We know based on scaling arguments that the angular sizes of features in the diffraction pattern must be related to the wavelength and the width, a, of the slit by some relationship of the form

$$\frac{\lambda}{a} \leftrightarrow \theta.$$

This is indeed true, and for instance the angle between the maximum of the central fringe and the maximum of the next fringe on one side equals $1.5\lambda/a$. Scaling arguments will never produce factors such as the 1.5, but they tell us that the answer must involve λ/a, so all the familiar qualitative facts are true. For instance, shorter-wavelength light will produce a more closely spaced diffraction pattern.

y / An image of the Pleiades star cluster. The circular rings around the bright stars are due to single-slit diffraction at the mouth of the telescope's tube.

An important scientific example of single-slit diffraction is in telescopes. Images of individual stars, as in figure y, are a good way to examine diffraction effects, because all stars except the sun are so far away that no telescope, even at the highest magnification, can image their disks or surface features. Thus any features of a star's image must be due purely to optical effects such as diffraction. A prominent cross appears around the brightest star, and dimmer ones surround the dimmer stars. Something like this is seen in most telescope photos, and indicates that inside the tube of the telescope there were two perpendicular struts or supports. Light diffracted around these struts. You might think that diffraction could be eliminated entirely by getting rid of all obstructions in the tube, but the circles around the stars are diffraction effects arising from single-slit diffraction at the mouth of the telescope's tube! (Actually we have not even talked about diffraction through a circular opening, but the idea is the same.) Since the angular sizes of the diffracted images depend on λ/a, the only way to improve the resolution of the images is to increase the diameter, a, of the tube. This is one of the main reasons (in addition to light-gathering power) why the best telescopes must be very large in diameter.

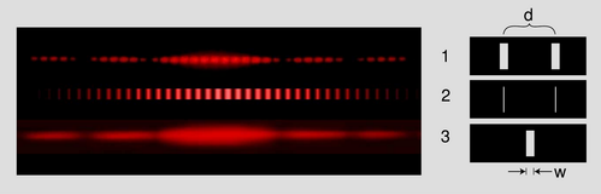
z / A radio telescope.

self-check:

What would this imply about radio telescopes as compared with visible-light telescopes?

(answer in the back of the PDF version of the book)

Double-slit diffraction is easier to understand conceptually than single-slit diffraction, but if you do a double-slit diffraction experiment in real life, you are likely to encounter a complicated pattern like figure aa/1, rather than the simpler one, 2, you were expecting. This is because the slits are fairly big compared to the wavelength of the light being used. We really have two different distances in our pair of slits: d, the distance between the slits, and w, the width of each slit. Remember that smaller distances on the object the light diffracts around correspond to larger features of the diffraction pattern. The pattern 1 thus has two spacings in it: a short spacing corresponding to the large distance d, and a long spacing that relates to the small dimension w.



aa / 1. A diffraction pattern formed by a real double slit. The width of each slit is fairly big compared to the wavelength of the light. This is a real photo. 2. This idealized pattern is not likely to occur in real life. To get it, you would need each slit to be so narrow that its width was comparable to the wavelength of the light, but that's not usually possible. This is not a real photo. 3. A real photo of a single-slit diffraction pattern caused by a slit whose width is the same as the widths of the slits used to make the top pattern.

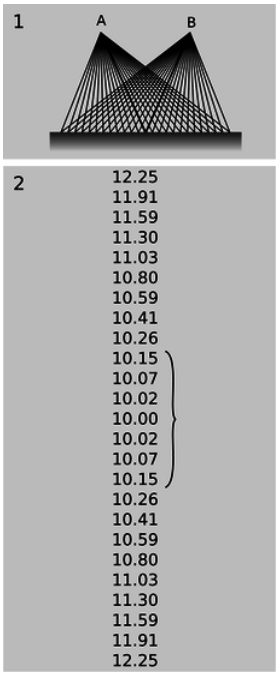
Discussion Question

◊ Why is it optically impossible for bacteria to evolve eyes that use visible light to form images?

The principle of least time

In subsection 12.1.5 and 12.4.5, we saw how in the ray model of light, both refraction and reflection can be described in an elegant and beautiful way by a single principle, the principle of least time. We can now justify the principle of least time based on the wave model of light. Consider an example involving reflection, ab. Starting at point A, Huygens' principle for waves tells us that we can think of the wave as spreading out in all directions. Suppose we imagine all the possible ways that a ray could travel from A to B. We show this by drawing 25 possible paths, of which the central one is the shortest. Since the principle of least time connects the wave model to the ray model, we should expect to get the most accurate results when the wavelength is much shorter than the distances involved --- for the sake of this numerical example, let's say that a wavelength is 1/10 of the shortest reflected path from A to B. The table, 2, shows the distances traveled by the 25 rays.

Note how similar are the distances traveled by the group of 7 rays, indicated with a bracket, that come closest to obeying the principle of least time. If we think of each one as a wave, then all 7 are again nearly in phase at point B. However, the rays that are farther from satisfying the principle of least time show more rapidly changing distances; on reuniting at point B, their phases are a random jumble, and they will very nearly cancel each other out. Thus, almost none of the wave energy delivered to point B goes by these longer paths. Physically we find, for instance, that a wave pulse emitted at A is observed at B after a time interval corresponding very nearly to the shortest possible path, and the pulse is not very “smeared out” when it gets there. The shorter the wavelength compared to the dimensions of the figure, the more accurate these approximate statements become.



ab / Light could take many different paths from A to B.

Instead of drawing a finite number of rays, such 25, what happens if we think of the angle, θ, of emission of the ray as a continuously varying variable? Minimizing the distance L requires

$$\frac{dL}{d\theta} = 0.$$

Because L is changing slowly in the vicinity of the angle that satisfies the principle of least time, all the rays that come out close to this angle have very nearly the same L, and remain very nearly in phase when they reach B. This is the basic reason why the discrete table, ab/2, turned out to have a group of rays that all traveled nearly the same distance.

As discussed in subsection 12.1.5, the principle of least time is really a principle of least *or greatest* time. This makes perfect sense, since $dL/d\theta = 0$ can in general describe either a minimum or a maximum

The principle of least time is very general. It does not apply just to refraction and reflection --- it can even be used to prove that light rays travel in a straight line through empty space, without taking detours! This general approach to wave motion was used by Richard Feynman, one of the pioneers who in the 1950's reconciled quantum mechanics with relativity. A very readable explanation is given in a book Feynman wrote for laypeople, QED: The Strange Theory of Light and Matter.

Contributors and Attributions

- Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

13.6: Footnotes

1. There is a standard piece of terminology which is that the “focal point” is the point lying on the optical axis at a distance from the mirror equal to the focal length. This term isn't particularly helpful, because it names a location where nothing normally happens. In particular, it is *not* normally the place where the rays come to a focus! --- that would be the *image* point. In other words, we don't normally have *[Math Processing Error]*, unless perhaps *[Math Processing Error]*. A recent online discussion among some physics teachers (carnot.physics.buffalo.edu/archives, Feb. 2006) showed that many disliked the terminology, felt it was misleading, or didn't know it and would have misinterpreted it if they had come across it. That is, it appears to be what grammarians call a “skunked term” --- a word that bothers half the population when it's used incorrectly, and the other half when it's used correctly.
2. I would like to thank Fouad Ajami for pointing out the pedagogical advantages of using both equations side by side.

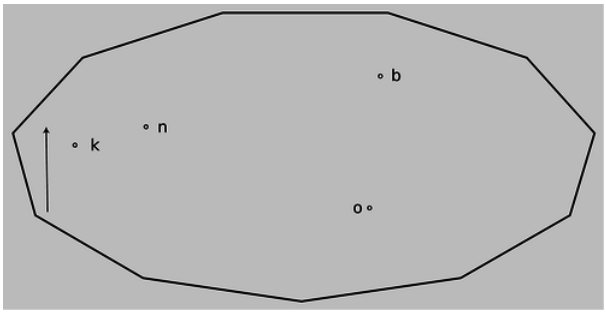
13.E: Optics (Exercises)

1. Draw a ray diagram showing why a small light source (a candle, say) produces sharper shadows than a large one (e.g., a long fluorescent bulb).

2. A Global Positioning System (GPS) receiver is a device that lets you figure out where you are by receiving timed radio signals from satellites. It works by measuring the travel time for the signals, which is related to the distance between you and the satellite. By finding the ranges to several different satellites in this way, it can pin down your location in three dimensions to within a few meters. How accurate does the measurement of the time delay have to be to determine your position to this accuracy?

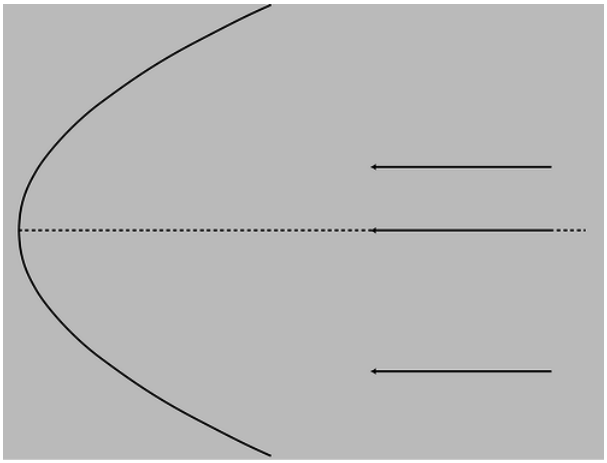
3. Estimate the frequency of an electromagnetic wave whose wavelength is similar in size to an atom (about a nm). Referring back to figure o on p. 703, in what part of the electromagnetic spectrum would such a wave lie (infrared, gamma-rays, ...)?

4. The Stealth bomber is designed with flat, smooth surfaces. Why would this make it difficult to detect via radar?



a / Problem 5.

5. The natives of planet Wumpus play pool using light rays on an eleven-sided table with mirrors for bumpers, shown in the figure on the next page. Trace this shot accurately with a ruler to reveal the hidden message. To get good enough accuracy, you'll need to photocopy the page (or download the book and print the page) and construct each reflection using a protractor.



b / Problem 6.

6. The figure on the next page shows a curved (parabolic) mirror, with three parallel light rays coming toward it. One ray is approaching along the mirror's center line. (a) Trace the drawing accurately, and continue the light rays until they are about to undergo their second reflection. To get good enough accuracy, you'll need to photocopy the page (or download the book and print the page) and draw in the normal at each place where a ray is reflected. What do you notice? (b) Make up an example of a practical use for this device. (c) How could you use this mirror with a small lightbulb to produce a parallel beam of light rays going off to the right?

7. (answer check available at lightandmatter.com) A man is walking at 1.0 m/s directly towards a flat mirror. At what speed is his separation from his image decreasing?

8. If a mirror on a wall is only big enough for you to see yourself from your head down to your waist, can you see your entire body by backing up? Test this experimentally and come up with an explanation for your observations, including a ray diagram.

Note that when you do the experiment, it's easy to confuse yourself if the mirror is even a tiny bit off of vertical. One way to check yourself is to artificially lower the top of the mirror by putting a piece of tape or a post-it note where it blocks your view of the top of your head. You can then check whether you are able to see more of yourself both above *and* below by backing up.

9. In section 12.2 we've only done examples of mirrors with hollowed-out shapes (called concave mirrors). Now draw a ray diagram for a curved mirror that has a bulging outward shape (called a convex mirror). (a) How does the image's distance from the mirror compare with the actual object's distance from the mirror? From this comparison, determine whether the magnification is greater than or less than one. (b) Is the image real, or virtual? Could this mirror ever make the other type of image?

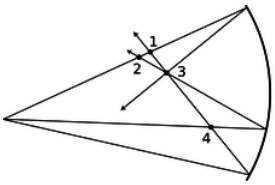
10. As discussed in question 9, there are two types of curved mirrors, concave and convex. Make a list of all the possible combinations of types of images (virtual or real) with types of mirrors (concave and convex). (Not all of the four combinations are physically possible.) Now for each one, use ray diagrams to determine whether increasing the distance of the object from the mirror leads to an increase or a decrease in the distance of the image from the mirror.

Draw BIG ray diagrams! Each diagram should use up about half a page of paper.

Some tips: To draw a ray diagram, you need two rays. For one of these, pick the ray that comes straight along the mirror's axis, since its reflection is easy to draw. After you draw the two rays and locate the image for the original object position, pick a new object position that results in the same type of image, and start a new ray diagram, in a different color of pen, right on top of the first one. For the two new rays, pick the ones that just happen to hit the mirror at the same two places; this makes it much easier to get the result right without depending on extreme accuracy in your ability to draw the reflected rays.

11. If the user of an astronomical telescope moves her head closer to or farther away from the image she is looking at, does the magnification change? Does the angular magnification change? Explain. (For simplicity, assume that no eyepiece is being used.)

12. In figure g/2 in on page 752, only the image of my forehead was located by drawing rays. Either photocopy the figure or download the book and print out the relevant page. On this copy of the figure, make a new set of rays coming from my chin, and locate its image. To make it easier to judge the angles accurately, draw rays from the chin that happen to hit the mirror at the same points where the two rays from the forehead were shown hitting it. By comparing the locations of the chin's image and the forehead's image, verify that the image is actually upside-down, as shown in the original figure.



c / Problem 13.

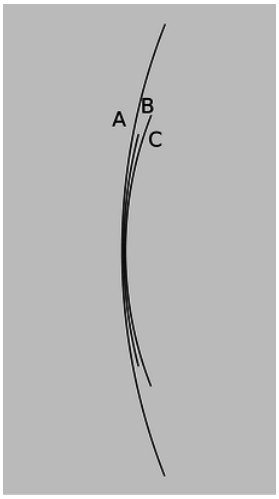
13. The figure shows four points where rays cross. Of these, which are image points? Explain.

14. Here's a game my kids like to play. I sit next to a sunny window, and the sun reflects from the glass on my watch, making a disk of light on the wall or floor, which they pretend to chase as I move it around. Is the spot a disk because that's the shape of the sun, or because it's the shape of my watch? In other words, would a square watch make a square spot, or do we just have a circular image of the circular sun, which will be circular no matter what?

15. Apply the equation $M = d_i/d_o$ to the case of a flat mirror.

16. (solution in the pdf version of the book) Use the method described in the text to derive the equation relating object distance to image distance for the case of a virtual image produced by a converging mirror.

17. Find the focal length of the mirror in problem 6 .(answer check available at lightandmatter.com)



d / Problem 18.

18. Rank the focal lengths of the mirrors in the figure, from shortest to longest. Explain.

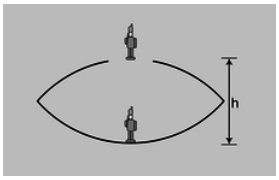
19. (solution in the pdf version of the book) (a) A converging mirror with a focal length of 20 cm is used to create an image, using an object at a distance of 10 cm. Is the image real, or is it virtual? (b) How about $f = 20$ cm and $d_o = 30$ cm? (c) What if it was a *diverging* mirror with $f = 20$ cm and $d_o = 10$ cm? (d) A diverging mirror with $f = 20$ cm and $d_o = 30$ cm?

20. (a) Make up a numerical example of a virtual image formed by a converging mirror with a certain focal length, and determine the magnification. (You will need the result of problem 16.) Make sure to choose values of d_o and f that would actually produce a virtual image, not a real one. Now change the location of the object *a little bit* and redetermine the magnification, showing that it changes. At my local department store, the cosmetics department sells hand mirrors advertised as giving a magnification of 5 times. How would you interpret this?

(b) Suppose a Newtonian telescope is being used for astronomical observing. Assume for simplicity that no eyepiece is used, and assume a value for the focal length of the mirror that would be reasonable for an amateur instrument that is to fit in a closet. Is the angular magnification different for objects at different distances? For example, you could consider two planets, one of which is twice as far as the other.

21. (a) Find a case where the magnification of a curved mirror is infinite. Is the *angular* magnification infinite from any realistic viewing position? (b) Explain why an arbitrarily large magnification can't be achieved by having a sufficiently small value of d_o.

22. A concave surface that reflects sound waves can act just like a converging mirror. Suppose that, standing near such a surface, you are able to find a point where you can place your head so that your own whispers are focused back on your head, so that they sound loud to you. Given your distance to the surface, what is the surface's focal length?



e / Problem 23.

23. The figure shows a device for constructing a realistic optical illusion. Two mirrors of equal focal length are put against each other with their silvered surfaces facing inward. A small object placed in the bottom of the cavity will have its image projected in the air above. The way it works is that the top mirror produces a virtual image, and the bottom mirror then creates a real image of the virtual image. (a) Show that if the image is to be positioned as shown, at the mouth of the cavity, then the focal length of the mirrors is related to the dimension h via the equation

$$\frac{1}{f} = \frac{1}{h} + \frac{1}{h + \left(\frac{1}{h} - \frac{1}{f}\right)^{-1}}.$$

(b) Restate the equation in terms of a single variable $x = h/f$, and show that there are two solutions for x. Which solution is physically consistent with the assumptions of the calculation?

24. (a) A converging mirror is being used to create a virtual image. What is the range of possible magnifications? (b) Do the same for the other types of images that can be formed by curved mirrors (both converging and diverging).

25. A diverging mirror of focal length f is fixed, and faces down. An object is dropped from the surface of the mirror, and falls away from it with acceleration g. The goal of the problem is to find the maximum velocity of the image.
(a) Describe the motion of the image verbally, and explain why we should expect there to be a maximum velocity.
(b) Use arguments based on units to determine the form of the solution, up to an unknown unitless multiplicative constant.
(c) Complete the solution by determining the unitless constant.

26. Diamond has an index of refraction of 2.42, and part of the reason diamonds sparkle is that this encourages a light ray to undergo many total internal reflections before it emerges. (a) Calculate the critical angle at which total internal reflection occurs in diamond. (answer check available at lightandmatter.com) (b) Explain the interpretation of your result: Is it measured from the normal, or from the surface? Is it a minimum, or a maximum? How would the critical angle have been different for a substance such as glass or plastic, with a lower index of refraction?

27. Suppose a converging lens is constructed of a type of plastic whose index of refraction is less than that of water. How will the lens's behavior be different if it is placed underwater?

28. There are two main types of telescopes, refracting (using a lens) and reflecting (using a mirror, as in figure i on p. 754). (Some telescopes use a mixture of the two types of elements: the light first encounters a large curved mirror, and then goes through an eyepiece that is a lens. To keep things simple, assume no eyepiece is used.) What implications would the color-dependence of focal length have for the relative merits of the two types of telescopes? Describe the case where an image is formed of a white star. You may find it helpful to draw a ray diagram.

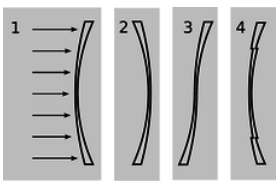
29. Based on Snell's law, explain why rays of light passing through the edges of a converging lens are bent more than rays passing through parts closer to the center. It might seem like it should be the other way around, since the rays at the edge pass through less glass --- shouldn't they be affected less? In your answer:

- Include a ray diagram showing a huge, full-page, close-up view of the relevant part of the lens.
- Make use of the fact that the front and back surfaces aren't always parallel; a lens in which the front and back surfaces *are* always parallel doesn't focus light at all, so if your explanation doesn't make use of this fact, your argument must be incorrect.
- Make sure your argument still works even if the rays don't come in parallel to the axis.

30. When you take pictures with a camera, the distance between the lens and the film has to be adjusted, depending on the distance at which you want to focus. This is done by moving the lens. If you want to change your focus so that you can take a picture of something farther away, which way do you have to move the lens? Explain using ray diagrams. [Based on a problem by Eric Mazur.]

31. When swimming underwater, why is your vision made much clearer by wearing goggles with flat pieces of glass that trap air behind them? [Hint: You can simplify your reasoning by considering the special case where you are looking at an object far away, and along the optic axis of the eye.]

32. (answer check available at lightandmatter.com) An object is more than one focal length from a converging lens. (a) Draw a ray diagram. (b) Using reasoning like that developed in section 12.3, determine the positive and negative signs in the equation $1/f = \pm 1/d_i \pm 1/d_o$. (c) The images of the rose in section 4.2 were made using a lens with a focal length of 23 cm. If the lens is placed 80 cm from the rose, locate the image.

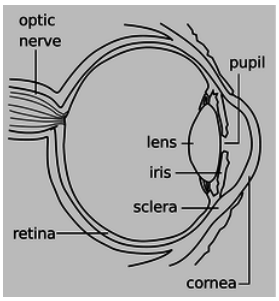


f / Problem 33.

33. The figure shows four lenses. Lens 1 has two spherical surfaces. Lens 2 is the same as lens 1 but turned around. Lens 3 is made by cutting through lens 1 and turning the bottom around. Lens 4 is made by cutting a central circle out of lens 1 and recessing it.

(a) A parallel beam of light enters lens 1 from the left, parallel to its axis. Reasoning based on Snell's law, will the beam emerging from the lens be bent inward, or outward, or will it remain parallel to the axis? Explain your reasoning. As part of your answer, make a huge drawing of one small part of the lens, and apply Snell's law at both interfaces. Recall that rays are bent more if they come to the interface at a larger angle with respect to the normal.

(b) What will happen with lenses 2, 3, and 4? Explain. Drawings are not necessary.



g / Problem 34.

34. The drawing shows the anatomy of the human eye, at twice life size. Find the radius of curvature of the outer surface of the cornea by measurements on the figure, and then derive the focal length of the air-cornea interface, where almost all the focusing of light occurs. You will need to use physical reasoning to modify the lensmaker's equation for the case where there is only a single refracting surface. Assume that the index of refraction of the cornea is essentially that of water.

35. (answer check available at lightandmatter.com) An object is less than one focal length from a converging lens. (a) Draw a ray diagram. (b) Using reasoning like that developed in section 12.3, determine the positive and negative signs in the equation $1/f = \pm 1/d_i \pm 1/d_o$. (c) The images of the rose in section 4.2 were made using a lens with a focal length of 23 cm. If the lens is placed 10 cm from the rose, locate the image.

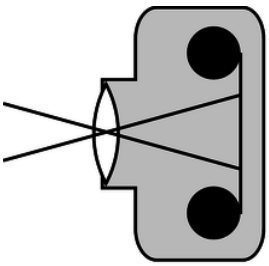
36. (answer check available at lightandmatter.com) Nearsighted people wear glasses whose lenses are diverging. (a) Draw a ray diagram. For simplicity pretend that there is no eye behind the glasses. (b) Using reasoning like that developed in section 12.3, determine the positive and negative signs in the equation $1/f = \pm 1/d_i \pm 1/d_o$. (c) If the focal length of the lens is 50.0 cm, and the person is looking at an object at a distance of 80.0 cm, locate the image.

37. (a) Light is being reflected diffusely from an object 1.000 m underwater. The light that comes up to the surface is refracted at the water-air interface. If the refracted rays all appear to come from the same point, then there will be a virtual image of the object in the water, above the object's actual position, which will be visible to an observer above the water. Consider three rays, A, B and C, whose angles in the water with respect to the normal are $\theta_i = 0.000^\circ$, 1.000° and 20.000° respectively. Find the depth of the point at which the refracted parts of A and B appear to have intersected, and do the same for A and C. Show that the intersections are at nearly the same depth, but not quite. [Check: The difference in depth should be about 4 cm.]

(b) Since all the refracted rays do not quite appear to have come from the same point, this is technically not a virtual image. In practical terms, what effect would this have on what you see?

(c) In the case where the angles are all small, use algebra and trig to show that the refracted rays do appear to come from the same point, and find an equation for the depth of the virtual image. Do not put in any numerical values for the angles or for the indices of refraction --- just keep them as symbols. You will need the approximation $\sin\theta \approx \tan\theta \approx \theta$, which is valid for small angles measured in radians.

38. Prove that the principle of least time leads to Snell's law.



h / Problem 39.

39. (solution in the pdf version of the book) Two standard focal lengths for camera lenses are 50 mm (standard) and 28 mm (wide-angle). To see how the focal lengths relate to the angular size of the field of view, it is helpful to visualize things as represented in the figure. Instead of showing many rays coming from the same point on the same object, as we normally do, the figure shows two rays from two different objects. Although the lens will intercept infinitely many rays from each of these points, we have shown only the ones that pass through the center of the lens, so that they suffer no angular deflection. (Any angular deflection at the front surface of the lens is canceled by an opposite deflection at the back, since the front and back surfaces are parallel at the lens's center.) What is special about these two rays is that they are aimed at the edges of one 35-mm-wide frame of film; that is, they show the limits of the field of view. Throughout this problem, we assume that d_o is much greater than d_i. (a) Compute the angular width of the camera's field of view when these two lenses are used. (b) Use small-angle approximations to find a simplified equation for the angular width of the field of view, θ, in terms of the focal length, f, and the width of the film, w. Your equation should not have any trig functions in it. Compare the results of this approximation with your answers from part a. (c) Suppose that we are holding constant the aperture (amount of surface area of the lens being used to collect light). When switching from a 50-mm lens to a 28-mm lens, how many times longer or shorter must the exposure be in order to make a properly developed picture, i.e., one that is not under- or overexposed? [Based on a problem by Arnold Arons.]

40. A nearsighted person is one whose eyes focus light too strongly, and who is therefore unable to relax the lens inside her eye sufficiently to form an image on her retina of an object that is too far away.

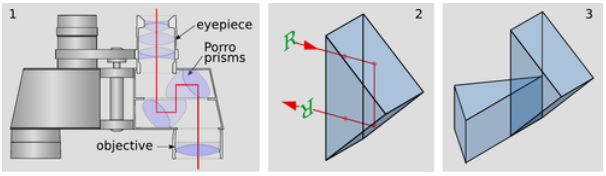
(a) Draw a ray diagram showing what happens when the person tries, with uncorrected vision, to focus at infinity.

(b) What type of lenses do her glasses have? Explain.

(c) Draw a ray diagram showing what happens when she wears glasses. Locate both the image formed by the glasses and the final image.

(d) Suppose she sometimes uses contact lenses instead of her glasses. Does the focal length of her contacts have to be less than, equal to, or greater than that of her glasses? Explain.

41. Fred's eyes are able to focus on things as close as 5.0 cm. Fred holds a magnifying glass with a focal length of 3.0 cm at a height of 2.0 cm above a flatworm. (a) Locate the image, and find the magnification. (b) Without the magnifying glass, from what distance would Fred want to view the flatworm to see its details as well as possible? With the magnifying glass? (c) Compute the angular magnification.



i / Problem 42.

42. Panel 1 of the figure shows the optics inside a pair of binoculars. They are essentially a pair of telescopes, one for each eye. But to make them more compact, and allow the eyepieces to be the right distance apart for a human face, they incorporate a set of eight prisms, which fold the light path. In addition, the prisms make the image upright. Panel 2 shows one of these prisms, known as a Porro prism. The light enters along a normal, undergoes two total internal reflections at angles of 45 degrees with respect to the back surfaces, and exits along a normal. The image of the letter R has been flipped across the horizontal. Panel 3 shows a pair of these prisms glued together. The image will be flipped across both the horizontal and the vertical, which makes it oriented the right way for the user of the binoculars.

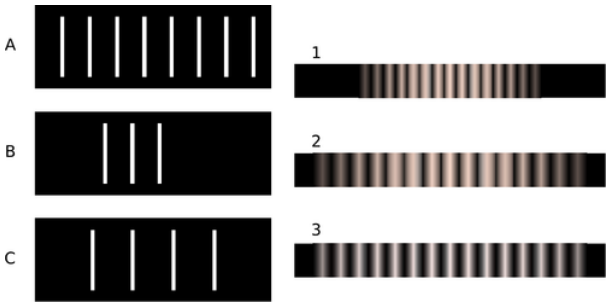
(a) Find the minimum possible index of refraction for the glass used in the prisms.
(b) For a material of this minimal index of refraction, find the fraction of the incoming light that will be lost to reflection in the four Porro prisms on a each side of a pair of binoculars. (See section 6.2.) In real, high-quality binoculars, the optical surfaces of the prisms have antireflective coatings, but carry out your calculation for the case where there is no such coating.
(c) Discuss the reasons why a designer of binoculars might or might not want to use a material with exactly the index of refraction found in part a.

43. It would be annoying if your eyeglasses produced a magnified or reduced image. Prove that when the eye is very close to a lens, and the lens produces a virtual image, the angular magnification is always approximately equal to 1 (regardless of whether the

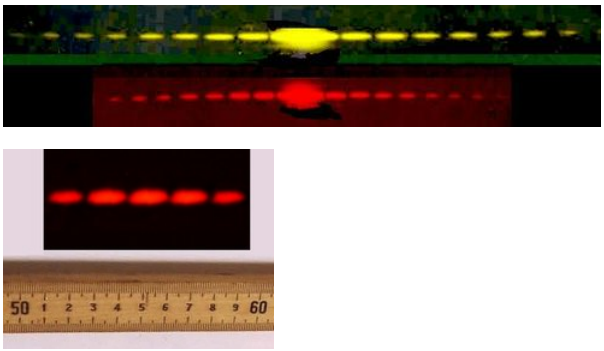
lens is diverging or converging).

44. The figure shows a diffraction pattern made by a double slit, along with an image of a meter stick to show the scale. Sketch the diffraction pattern from the figure on your paper. Now consider the four variables in the equation $\lambda/d = \sin\theta/m$. Which of these are the same for all five fringes, and which are different for each fringe? Which variable would you naturally use in order to label which fringe was which? Label the fringes on your sketch using the values of that variable.

45. Match gratings A-C with the diffraction patterns 1-3 that they produce. Explain.



46. The figure below shows two diffraction patterns. The top one was made with yellow light, and the bottom one with red. Could the slits used to make the two patterns have been the same?

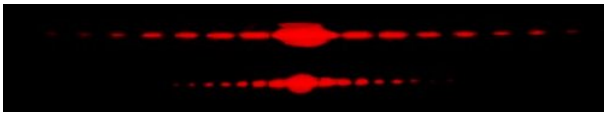


j / Problems 44 and 47.

47. The figure on p. 805 shows a diffraction pattern made by a double slit, along with an image of a meter stick to show the scale. The slits were 146 cm away from the screen on which the diffraction pattern was projected. The spacing of the slits was 0.050 mm. What was the wavelength of the light?(answer check available at lightandmatter.com)

48. Why would blue or violet light be the best for microscopy?

49. The figure below shows two diffraction patterns, both made with the same wavelength of red light. (a) What type of slits made the patterns? Is it a single slit, double slits, or something else? Explain. (b) Compare the dimensions of the slits used to make the top and bottom pattern. Give a numerical ratio, and state which way the ratio is, i.e., which slit pattern was the larger one. Explain.

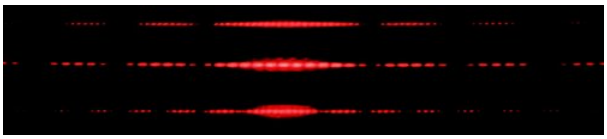


50. When white light passes through a diffraction grating, what is the smallest value of m for which the visible spectrum of order m overlaps the next one, of order $m+1$? (The visible spectrum runs from about 400 nm to about 700 nm.)

k / Problem 51. This image of the Pleiades star cluster shows haloes around the stars due to the wave nature of light.

51. For star images such as the ones in figure y, estimate the angular width of the diffraction spot due to diffraction at the mouth of the telescope. Assume a telescope with a diameter of 10 meters (the largest currently in existence), and light with a wavelength in the middle of the visible range. Compare with the actual angular size of a star of diameter 10^9 m seen from a distance of 10^{17} m. What does this tell you?

52. The figure below shows three diffraction patterns. All were made under identical conditions, except that a different set of double slits was used for each one. The slits used to make the top pattern had a center-to-center separation $d = 0.50$ mm, and each slit was $w = 0.04$ mm wide. (a) Determine d and w for the slits used to make the pattern in the middle. (b) Do the same for the slits used to make the bottom pattern.

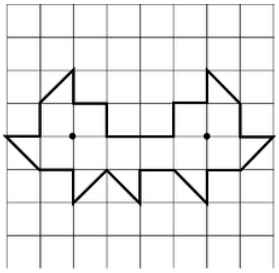


53. (answer check available at lightandmatter.com) The beam of a laser passes through a diffraction grating, fans out, and illuminates a wall that is perpendicular to the original beam, lying at a distance of 2.0 m from the grating. The beam is produced by a helium-neon laser, and has a wavelength of 694.3 nm. The grating has 2000 lines per centimeter. (a) What is the distance on the wall between the central maximum and the maxima immediately to its right and left? (b) How much does your answer change when you use the small-angle approximations $\theta \approx \sin\theta \approx \tan\theta$?

54. Ultrasound, i.e., sound waves with frequencies too high to be audible, can be used for imaging fetuses in the womb or for breaking up kidney stones so that they can be eliminated by the body. Consider the latter application. Lenses can be built to focus sound waves, but because the wavelength of the sound is not all that small compared to the diameter of the lens, the sound will not be concentrated exactly at the geometrical focal point. Instead, a diffraction pattern will be created with an intense central spot surrounded by fainter rings. About 85% of the power is concentrated within the central spot. The angle of the first minimum (surrounding the central spot) is given by $\sin\theta = \lambda/b$, where b is the diameter of the lens. This is similar to the corresponding equation for a single slit, but with a factor of 1.22 in front which arises from the circular shape of the aperture. Let the distance from the lens to the patient's kidney stone be $L = 20$ cm. You will want $f > 20$ kHz, so that the sound is inaudible. Find values of b and f that would result in a usable design, where the central spot is small enough to lie within a kidney stone 1 cm in diameter.

55. Under what circumstances could one get a mathematically undefined result by solving the double-slit diffraction equation for θ? Give a physical interpretation of what would actually be observed.

56. When ultrasound is used for medical imaging, the frequency may be as high as 5-20 MHz. Another medical application of ultrasound is for therapeutic heating of tissues inside the body; here, the frequency is typically 1-3 MHz. What fundamental physical reasons could you suggest for the use of higher frequencies for imaging?



l / Problem 57.

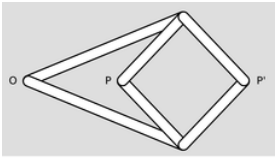
57. Suppose we have a polygonal room whose walls are mirrors, and there a pointlike light source in the room. In most such examples, every point in the room ends up being illuminated by the light source after some finite number of reflections. A difficult mathematical question, first posed in the middle of the last century, is whether it is ever possible to have an example in which the whole room is *not* illuminated. (Rays are assumed to be absorbed if they strike exactly at a vertex of the polygon, or if they pass exactly through the plane of a mirror.)

The problem was finally solved in 1995 by G.W. Tokarsky, who found an example of a room that was not illuminable from a certain point. Figure 57 shows a slightly simpler example found two years later by D. Castro. If a light source is placed at either of the locations shown with dots, the other dot remains unilluminated, although every other point is lit up. It is not straightforward to prove rigorously that Castro's solution has this property. However, the plausibility of the solution can be demonstrated as follows.

Suppose the light source is placed at the right-hand dot. Locate all the images formed by single reflections. Note that they form a regular pattern. Convince yourself that none of these images illuminates the left-hand dot. Because of the regular pattern, it becomes plausible that even if we form images of images, images of images of images, etc., none of them will ever illuminate the other dot.

There are various other versions of the problem, some of which remain unsolved. The book by Klee and Wagon gives a good introduction to the topic, although it predates Tokarsky and Castro's work.

References:
G.W. Tokarsky, “Polygonal Rooms Not Illuminable from Every Point.” Amer. Math. Monthly 102, 867-879, 1995.
D. Castro, “Corrections.” Quantum 7, 42, Jan. 1997.
V. Klee and S. Wagon, *Old and New Unsolved Problems in Plane Geometry and Number Theory*. Mathematical Association of America, 1991.



m / Problem 58.

58. A mechanical linkage is a device that changes one type of motion into another. The most familiar example occurs in a gasoline car's engine, where a connecting rod changes the linear motion of the piston into circular motion of the crankshaft. The top panel of the figure shows a mechanical linkage invented by Peaucellier in 1864, and independently by Lipkin around the same time. It consists of six rods joined by hinges, the four short ones forming a rhombus. Point O is fixed in space, but the apparatus is free to rotate about O. Motion at P is transformed into a different motion at P' (or vice versa).

Geometrically, the linkage is a mechanical implementation of the ancient problem of inversion in a circle. Considering the case in which the rhombus is folded flat, let the k be the distance from O to the point where P and P' coincide. Form the circle of radius k

with its center at O. As P and P' move in and out, points on the inside of the circle are always mapped to points on its outside, such that $rr' = k^2$. That is, the linkage is a type of analog computer that exactly solves the problem of finding the inverse of a number r. Inversion in a circle has many remarkable geometrical properties, discussed in H.S.M. Coxeter, *Introduction to Geometry*, Wiley, 1961. If a pen is inserted through a hole at P, and P' is traced over a geometrical figure, the Peaucellier linkage can be used to draw a kind of image of the figure.

A related problem is the construction of pictures, like the one in the bottom panel of the figure, called anamorphs. The drawing of the column on the paper is highly distorted, but when the reflecting cylinder is placed in the correct spot on top of the page, an undistorted image is produced inside the cylinder. (Wide-format movie technologies such as Cinemascope are based on similar principles.)

Show that the Peaucellier linkage does *not* convert correctly between an image and its anamorph, and design a modified version of the linkage that does. Some knowledge of analytic geometry will be helpful.

n / Problem 59.

59. The figure shows a lens with surfaces that are curved, but whose thickness is constant along any horizontal line. Use the lensmaker's equation to prove that this “lens” is not really a lens at all.(solution in the pdf version of the book)

60. Under ordinary conditions, gases have indices of refraction only a little greater than that of vacuum, i.e., $n = 1 + \epsilon$, where ϵ is some small number. Suppose that a ray crosses a boundary between a region of vacuum and a region in which the index of refraction is $1 + \epsilon$. Find the maximum angle by which such a ray can ever be deflected, in the limit of small ϵ. \hwhint{hwhint:very-weak-refraction}

61. A converging mirror has focal length f. An object is located at a distance $(1 + \epsilon)f$ from the mirror, where ϵ is small. Find the distance of the image from the mirror, simplifying your result as much as possible by using the assumption that ϵ is small. \hwans{hwans:close-to-focal-length}

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

CHAPTER OVERVIEW

14: Quantum Physics

Thumbnail: Schrödinger took the absurd implications of this thought experiment (a cat simultaneously dead and alive) as an argument against the Copenhagen interpretation. However, this interpretation remains the most commonly taught view of quantum mechanics.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

14.1: Rules of Randomness

- Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective positions of the things which compose it...nothing would be uncertain, and the future as the past would be laid out before its eyes. -- *Pierre Simon de Laplace, 1776*
- The energy produced by the atom is a very poor kind of thing. Anyone who expects a source of power from the transformation of these atoms is talking moonshine. -- *Ernest Rutherford, 1933*
- The Quantum Mechanics is very imposing. But an inner voice tells me that it is still not the final truth. The theory yields much, but it hardly brings us nearer to the secret of the Old One. In any case, I am convinced that He does not play dice. -- *Albert Einstein*

However radical Newton's clockwork universe seemed to his contemporaries, by the early twentieth century it had become a sort of smugly accepted dogma. Luckily for us, this deterministic picture of the universe breaks down at the atomic level. The clearest demonstration that the laws of physics contain elements of randomness is in the behavior of radioactive atoms. Pick two identical atoms of a radioactive isotope, say the naturally occurring uranium 238, and watch them carefully. They will decay at different times, even though there was no difference in their initial behavior.

We would be in big trouble if these atoms' behavior was as predictable as expected in the Newtonian world-view, because radioactivity is an important source of heat for our planet. In reality, each atom chooses a random moment at which to release its energy, resulting in a nice steady heating effect. The earth would be a much colder planet if only sunlight heated it and not radioactivity. Probably there would be no volcanoes, and the oceans would never have been liquid. The deep-sea geothermal vents in which life first evolved would never have existed. But there would be an even worse consequence if radioactivity was deterministic: after a few billion years of peace, all the uranium 238 atoms in our planet would presumably pick the same moment to decay. The huge amount of stored nuclear energy, instead of being spread out over eons, would all be released at one instant, blowing our whole planet to Kingdom Come.[1]

Figure* 14.1.1*: *In 1980, the continental U.S. got its first taste of active volcanism in recent memory with the eruption of Mount St. Helens.*

The new version of physics, incorporating certain kinds of randomness, is called quantum physics (for reasons that will become clear later). It represented such a dramatic break with the previous, deterministic tradition that everything that came before is considered “classical,” even the theory of relativity. This chapter is a basic introduction to quantum physics.

Exercise 14.1.1

I said “Pick two identical atoms of a radioactive isotope.” Are two atoms really identical? If their electrons are orbiting the nucleus, can we distinguish each atom by the particular arrangement of its electrons at some instant in time?

Answer

Add texts here. Do not delete this text first.

13.1.1 Randomness isn't random

Einstein's distaste for randomness, and his association of determinism with divinity, goes back to the Enlightenment conception of the universe as a gigantic piece of clockwork that only had to be set in motion initially by the Builder. Many of the founders of quantum mechanics were interested in possible links between physics and Eastern and Western religious and philosophical thought, but every educated person has a different concept of religion and philosophy. Bertrand Russell remarked, “Sir Arthur Eddington deduces religion from the fact that atoms do not obey the laws of mathematics. Sir James Jeans deduces it from the fact that they do.”

Russell's witticism, which implies incorrectly that mathematics cannot describe randomness, remind us how important it is not to oversimplify this question of randomness. You should not simply surmise, “Well, it's all random, anything can happen.” For one thing, certain things simply cannot happen, either in classical physics or quantum physics. The conservation laws of mass, energy, momentum, and angular momentum are still valid, so for instance processes that create energy out of nothing are not just unlikely according to quantum physics, they are impossible.

A useful analogy can be made with the role of randomness in evolution. Darwin was not the first biologist to suggest that species changed over long periods of time. His two new fundamental ideas were that (1) the changes arose through random genetic variation, and (2) changes that enhanced the organism's ability to survive and reproduce would be preserved, while maladaptive changes would be eliminated by natural selection. Doubters of evolution often consider only the first point, about the randomness of natural variation, but not the second point, about the systematic action of natural selection. They make statements such as, “the development of a complex organism like Homo sapiens via random chance would be like a whirlwind blowing through a junkyard and spontaneously assembling a jumbo jet out of the scrap metal.” The flaw in this type of reasoning is that it ignores the deterministic constraints on the results of random processes. For an atom to violate conservation of energy is no more likely than the conquest of the world by chimpanzees next year.

Exercise 14.1.1

Economists often behave like wannabe physicists, probably because it seems prestigious to make numerical calculations instead of talking about human relationships and organizations like other social scientists. Their striving to make economics work like Newtonian physics extends to a parallel use of mechanical metaphors, as in the concept of a market's supply and demand acting like a self-adjusting machine, and the idealization of people as economic automatons who consistently strive to maximize their own wealth. What evidence is there for randomness rather than mechanical determinism in economics?

Answer

Discussion Question - no answer given.

13.1.2 Calculating Randomness

You should also realize that even if something is random, we can still understand it, and we can still calculate probabilities numerically. In other words, physicists are good bookmakers. A good bookmaker can calculate the odds that a horse will win a race much more accurately that an inexperienced one, but nevertheless cannot predict what will happen in any particular race.

As an illustration of a general technique for calculating odds, suppose you are playing a 25-cent slot machine. Each of the three wheels has one chance in ten of coming up with a cherry. If all three wheels come up cherries, you win $100. Even though the results of any particular trial are random, you can make certain quantitative predictions. First, you can calculate that your odds of winning on any given trial are $1/10 \times 1/10 \times 1/10 = 1/1000 = 0.001$. Here, I am representing the probabilities as numbers from 0 to 1, which is clearer than statements like “The odds are 999 to 1,” and makes the calculations easier. A probability of 0 represents something impossible, and a probability of 1 represents something that will definitely happen.

Also, you can say that any given trial is equally likely to result in a win, and it doesn't matter whether you have won or lost in prior games. Mathematically, we say that each trial is statistically independent, or that separate games are uncorrelated. Most gamblers are mistakenly convinced that, to the contrary, games of chance are correlated. If they have been playing a slot machine all day, they are convinced that it is “getting ready to pay,” and they do not want anyone else playing the machine and “using up” the jackpot that they “have coming.” In other words, they are claiming that a series of trials at the slot machine is negatively correlated, that losing now makes you more likely to win later. Craps players claim that you should go to a table where the person rolling the

dice is “hot,” because she is likely to keep on rolling good numbers. Craps players, then, believe that rolls of the dice are positively correlated, that winning now makes you more likely to win later.

My method of calculating the probability of winning on the slot machine was an example of the following important rule for calculations based on independent probabilities:

The law of independent probabilities

If the probability of one event happening is P_A, and the probability of a second statistically independent event happening is P_B, then the probability that they will both occur is the product of the probabilities, $P_A P_B$. If there are more than two events involved, you simply keep on multiplying.

This can be taken as the definition of statistical independence.

Note that this only applies to independent probabilities. For instance, if you have a nickel and a dime in your pocket, and you randomly pull one out, there is a probability of 0.5 that it will be the nickel. If you then replace the coin and again pull one out randomly, there is again a probability of 0.5 of coming up with the nickel, because the probabilities are independent. Thus, there is a probability of 0.25 that you will get the nickel both times.

Suppose instead that you do not replace the first coin before pulling out the second one. Then you are bound to pull out the other coin the second time, and there is no way you could pull the nickel out twice. In this situation, the two trials are not independent, because the result of the first trial has an effect on the second trial. The law of independent probabilities does not apply, and the probability of getting the nickel twice is zero, not 0.25.

Experiments have shown that in the case of radioactive decay, the probability that any nucleus will decay during a given time interval is unaffected by what is happening to the other nuclei, and is also unrelated to how long it has gone without decaying. The first observation makes sense, because nuclei are isolated from each other at the centers of their respective atoms, and therefore have no physical way of influencing each other. The second fact is also reasonable, since all atoms are identical. Suppose we wanted to believe that certain atoms were “extra tough,” as demonstrated by their history of going an unusually long time without decaying. Those atoms would have to be different in some physical way, but nobody has ever succeeded in detecting differences among atoms. There is no way for an atom to be changed by the experiences it has in its lifetime.

The *Law of Independent Probabilities* tells us to use multiplication to calculate the probability that both A and B will happen, assuming the probabilities are independent. What about the probability of an “or” rather than an “and”? If two events A and B are mutually exclusive, then the probability of one or the other occurring is the sum $P_A + P_B$. For instance, a bowler might have a 30% chance of getting a strike (knocking down all ten pins) and a 20% chance of knocking down nine of them. The bowler's chance of knocking down either nine pins or ten pins is therefore 50%.

It does not make sense to add probabilities of things that are not mutually exclusive, i.e., that could both happen. Say I have a 90% chance of eating lunch on any given day, and a 90% chance of eating dinner. The probability that I will eat either lunch or dinner is not 180%.

Normalization

If I spin a globe and randomly pick a point on it, I have about a 70% chance of picking a point that's in an ocean and a 30% chance of picking a point on land. The probability of picking either water or land is 70. Water and land are mutually exclusive, and there are no other possibilities, so the probabilities had to add up to 100%. It works the same if there are more than two possibilities --- if you can classify all possible outcomes into a list of mutually exclusive results, then all the probabilities have to add up to 1, or 100%. This property of probabilities is known as normalization.

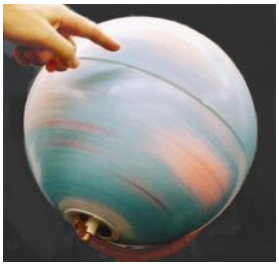


Figure **14.1.2:** Normalization: the probability of picking land plus the probability of picking water adds up to 1.

Averages

Another way of dealing with randomness is to take averages. The casino knows that in the long run, the number of times you win will approximately equal the number of times you play multiplied by the probability of winning. In the slot-machine game described on page 823, where the probability of winning is 0.001, if you spend a week playing, and pay $2500 to play 10,000 times, you are likely to win about 10 times $(10,000 \times 0.001 = 10)$, and collect $1000. On the average, the casino will make a profit of $1500 from you. This is an example of the following rule.

Rule for Calculating Averages

If you conduct N identical, statistically independent trials, and the probability of success in each trial is P, then on the average, the total number of successful trials will be NP. If N is large enough, the relative error in this estimate will become small.

The statement that the rule for calculating averages gets more and more accurate for larger and larger N (known popularly as the "law of averages") often provides a correspondence principle that connects classical and quantum physics. For instance, the amount of power produced by a nuclear power plant is not random at any detectable level, because the number of atoms in the reactor is so large. In general, random behavior at the atomic level tends to average out when we consider large numbers of atoms, which is why physics seemed deterministic before physicists learned techniques for studying atoms individually.

We can achieve great precision with averages in quantum physics because we can use identical atoms to reproduce exactly the same situation many times. If we were betting on horses or dice, we would be much more limited in our precision. After a thousand races, the horse would be ready to retire. After a million rolls, the dice would be worn out.

Exercise 14.1.1

Which of the following things *must* be independent, which *could* be independent, and which definitely are *not* independent? (1) the probability of successfully making two free-throws in a row in basketball; (2) the probability that it will rain in London tomorrow and the probability that it will rain on the same day in a certain city in a distant galaxy; (3) your probability of dying today and of dying tomorrow.

Answer

Answer in the back of the PDF version of the book.

Exercise 14.1.1 Discussion Questions

Figure **14.1.3:** Why are dice random?

- Newtonian physics is an essentially perfect approximation for describing the motion of a pair of dice. If Newtonian physics is deterministic, why do we consider the result of rolling dice to be random?
- Why isn't it valid to define randomness by saying that randomness is when all the outcomes are equally likely?
- The sequence of digits 12121212121212121 seems clearly nonrandom, and 41592653589793 seems random. The latter sequence, however, is the decimal form of pi, starting with the third digit. There is a story about the Indian mathematician Ramanujan, a self-taught prodigy, that a friend came to visit him in a cab, and remarked that the number of the cab, 1729, seemed relatively uninteresting. Ramanujan replied that on the contrary, it was very interesting because it was the smallest number that could be represented in two different ways as the sum of two cubes. The Argentine author Jorge Luis Borges wrote a short story called "The Library of Babel," in which he imagined a library containing every book that could possibly be written using the letters of the alphabet. It would include a book containing only the repeated letter "a;" all the ancient

Greek tragedies known today, all the lost Greek tragedies, and millions of Greek tragedies that were never actually written; your own life story, and various incorrect versions of your own life story; and countless anthologies containing a short story called “The Library of Babel.” Of course, if you picked a book from the shelves of the library, it would almost certainly look like a nonsensical sequence of letters and punctuation, but it's always possible that the seemingly meaningless book would be a science-fiction screenplay written in the language of a Neanderthal tribe, or the lyrics to a set of incomparably beautiful love songs written in a language that never existed. In view of these examples, what does it really mean to say that something is random?

Answer

Add texts here. Do not delete this text first.

13.1.3 Probability Distributions

So far we've discussed random processes having only two possible outcomes: yes or no, win or lose, on or off. More generally, a random process could have a result that is a number. Some processes yield integers, as when you roll a die and get a result from one to six, but some are not restricted to whole numbers, for example the number of seconds that a uranium-238 atom will exist before undergoing radioactive decay.

Consider a throw of a die. If the die is “honest,” then we expect all six values to be equally likely. Since all six probabilities must add up to 1, then probability of any particular value coming up must be 1/6. We can summarize this in a graph (Figure 14.1.4*a*). Areas under the curve can be interpreted as total probabilities. For instance, the area under the curve from 1 to 3 is $1/6+1/6+1/6=1/2$, so the probability of getting a result from 1 to 3 is 1/2. The function shown on the graph is called the probability distribution.

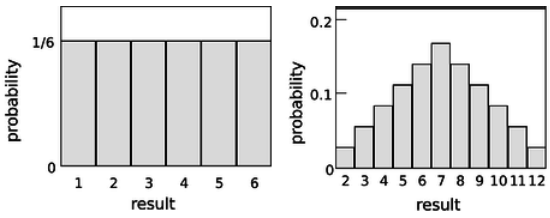


Figure* 14.1.4*: *(a)Probability distribution for the result of rolling a single die. (b) Rolling two dice and adding them up.*

Figure 14.1.4*b*shows the probabilities of various results obtained by rolling two dice and adding them together, as in the game of craps. The probabilities are not all the same. There is a small probability of getting a two, for example, because there is only one way to do it, by rolling a one and then another one. The probability of rolling a seven is high because there are six different ways to do it: 1+6, 2+5, etc.

If the number of possible outcomes is large but finite, for example the number of hairs on a dog, the graph would start to look like a smooth curve rather than a ziggurat.

What about probability distributions for random numbers that are not integers? We can no longer make a graph with probability on the y axis, because the probability of getting a given exact number is typically zero. For instance, there is zero probability that a radioactive atom will last for *exactly* 3 seconds, since there is are infinitely many possible results that are close to 3 but not exactly three: 2.99999999999999999968768765876584654 36, for example. It doesn't usually make sense, therefore, to talk about the probability of a single numerical result, but it does make sense to talk about the probability of a certain range of results. For instance, the probability that an atom will last more than 3 and less than 4 seconds is a perfectly reasonable thing to discuss. We can still summarize the probability information on a graph, and we can still interpret areas under the curve as probabilities.

But the y axis can no longer be a unitless probability scale. In radioactive decay, for example, we want the x axis to have units of time, and we want areas under the curve to be unitless probabilities. The area of a single square on the graph paper is then

$$\begin{aligned}&(\text{unitless area of a square})\\ &=(\text{width of square with time units})\\ &\quad\times(\text{height of square}).\end{aligned}$$

If the units are to cancel out, then the height of the square must evidently be a quantity with units of inverse time. In other words, the y axis of the graph is to be interpreted as probability per unit time, not probability. Figure 14.1.5 shows another example, a probability distribution for people's height. This kind of bell-shaped curve is quite common.\

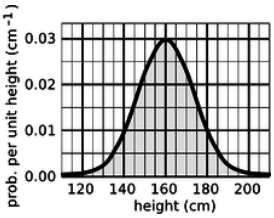


Figure* 14.1.5*: *A probability distribution for height of human adults (not real data).*

Exercise 14.1.1

Compare the number of people with heights in the range of 130-135 cm to the number in the range 135-140

Answer

(answer in the back of the PDF version of the book).

Example 14.1.1: Looking for tall basketball players

A certain country with a large population wants to find very tall people to be on its Olympic basketball team and strike a blow against western imperialism. Out of a pool of 10^8 people who are the right age and gender, how many are they likely to find who are over 225 cm (7 feet 4 inches) in height? Figure g gives a close-up of the "tail" of the distribution shown previously in Figure 14.1.6

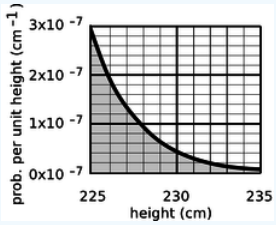


***Figure* 14.1.6**

The shaded area under the curve represents the probability that a given person is tall enough. Each rectangle represents a probability of $0.2 \times 10^{-7}\ \text{cm}^{-1} \times 1\ \text{cm} = 2 \times 10^{-8}$. There are about 35 rectangles covered by the shaded area, so the probability of having a height greater than 225 cm is 7×10^{-7}, or just under one in a million. Using the rule for calculating averages, the average, or expected number of people this tall is $(10^8) \times (7 \times 10^{-7}) = 70$.

Average and width of a probability distribution

If the next Martian you meet asks you, "How tall is an adult human?," you will probably reply with a statement about the average human height, such as "Oh, about 5 feet 6 inches." If you wanted to explain a little more, you could say, "But that's only an average. Most people are somewhere between 5 feet and 6 feet tall." Without bothering to draw the relevant bell curve for your new extraterrestrial acquaintance, you've summarized the relevant information by giving an average and a typical range of variation.

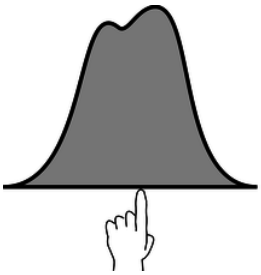


Figure* 14.1.6*: *The average of a probability distribution.*

The average of a probability distribution can be defined geometrically as the horizontal position at which it could be balanced if it was constructed out of cardboard. A convenient numerical measure of the amount of variation about the average, or amount of uncertainty, is the full width at half maximum, or FWHM, shown in Figure 14.1.7.

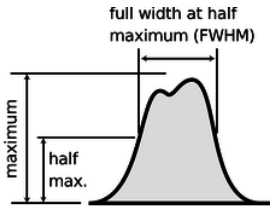


Figure* 14.1.7*: *The full width at half maximum (FWHM) of a probability distribution.*

A great deal more could be said about this topic, and indeed an introductory statistics course could spend months on ways of defining the center and width of a distribution. Rather than force-feeding you on mathematical detail or techniques for calculating these things, it is perhaps more relevant to point out simply that there are various ways of defining them, and to inoculate you against the misuse of certain definitions.

The average is not the only possible way to say what is a typical value for a quantity that can vary randomly; another possible definition is the median, defined as the value that is exceeded with 50% probability. When discussing incomes of people living in a certain town, the average could be very misleading, since it can be affected massively if a single resident of the town is Bill Gates. Nor is the FWHM the only possible way of stating the amount of random variation; another possible way of measuring it is the standard deviation (defined as the square root of the average squared deviation from the average value).

13.1.4 Exponential Decay and Half-life

Most people know that radioactivity “lasts a certain amount of time,” but that simple statement leaves out a lot. As an example, consider the following medical procedure used to diagnose thyroid function. A very small quantity of the isotope $^{131}\mathrm{I}$, produced in a nuclear reactor, is fed to or injected into the patient. The body's biochemical systems treat this artificial, radioactive isotope exactly the same as $^{127}\mathrm{I}$, which is the only naturally occurring type. (Nutritionally, iodine is a necessary trace element. Iodine taken into the body is partly excreted, but the rest becomes concentrated in the thyroid gland. Iodized salt has had iodine added to it to prevent the nutritional deficiency known as goiters, in which the iodine-starved thyroid becomes swollen.) As the $^{131}\mathrm{I}$ undergoes beta decay, it emits electrons, neutrinos, and gamma rays. The gamma rays can be measured by a detector passed over the patient's body. As the radioactive iodine becomes concentrated in the thyroid, the amount of gamma radiation coming from the thyroid becomes greater, and that emitted by the rest of the body is reduced. The rate at which the iodine concentrates in the thyroid tells the doctor about the health of the thyroid.

If you ever undergo this procedure, someone will presumably explain a little about radioactivity to you, to allay your fears that you will turn into the Incredible Hulk, or that your next child will have an unusual number of limbs. Since iodine stays in your thyroid for a long time once it gets there, one thing you'll want to know is whether your thyroid is going to become radioactive forever. They may just tell you that the radioactivity “only lasts a certain amount of time,” but we can now carry out a quantitative derivation of how the radioactivity really will die out.

Let $P_{surv}(t)$ be the probability that an iodine atom will survive without decaying for a period of at least t. It has been experimentally measured that half all $^{131}\mathrm{I}$ atoms decay in 8 hours, so we have

$$P_{surv}(8\text{ hr}) = 0.5. \tag{14.1.1}$$

Now using the law of independent probabilities, the probability of surviving for 16 hours equals the probability of surviving for the first 8 hours multiplied by the probability of surviving for the second 8 hours,

$$\begin{aligned} P_{surv}\,(16\text{ hr}) &= 0.50 \times 0.50 \\ &= 0.25. \end{aligned}$$

Similarly we have

$$\begin{aligned} P_{surv}\,(24\text{ hr}) &= 0.50 \times 0.5 \times 0.5 \\ &= 0.125. \end{aligned}$$

Generalizing from this pattern, the probability of surviving for any time t that is a multiple of 8 hours is

$$P_{surv}\,(t) = 0.5^{t/8\text{ hr}}. \tag{14.1.2}$$

We now know how to find the probability of survival at intervals of 8 hours, but what about the points in time in between? What would be the probability of surviving for 4 hours? Well, using the law of independent probabilities again, we have

$$P_{surv}\,(8\text{ hr}) = P_{surv}\,(4\text{ hr}) \times P_{surv}\,(4\text{ hr}), \tag{14.1.3}$$

which can be rearranged to give

$$\begin{aligned} P_{surv}\,(4\text{ hr}) &= \sqrt{P_{surv}\,(8\text{ hr})} \\ &= \sqrt{0.5} \\ &= 0.707. \end{aligned}$$

This is exactly what we would have found simply by plugging in $P_{surv}\,(t) = 0.5^{t/8\text{ hr}}$ and ignoring the restriction to multiples of 8 hours. Since 8 hours is the amount of time required for half of the atoms to decay, it is known as the half-life, written $t_{1/2}$. The general rule is as follows:

$$\underbrace{P_{surv}\,(t) = 0.5^{t/t_{1/2}}}_{\text{Exponential Decay Equation}} \tag{14.1.4}$$

Using the rule for calculating averages, we can also find the number of atoms, $N(t)$, remaining in a sample at time t:

$$N(t) = N(0) \times 0.5^{t/t_{1/2}} \tag{14.1.5}$$

Both of these equations have graphs that look like dying-out exponentials, as in the example below.

Example 14.1.2: Radioactive contamination at Chernobyl

One of the most dangerous radioactive isotopes released by the Chernobyl disaster in 1986 was ^{90}Sr, whose half-life is 28 years. (a) How long will it be before the contamination is reduced to one tenth of its original level? (b) If a total of 10^{27} atoms was released, about how long would it be before not a single atom was left?

Solution

(a) We want to know the amount of time that a ^{90}Sr nucleus has a probability of 0.1 of surviving. Starting with the exponential decay formula,

$$P_{surv} = 0.5^{t/t_{1/2}},$$

we want to solve for t. Taking natural logarithms of both sides,

$$\ln P = \frac{t}{t_{1/2}} \ln 0.5,$$

so

$$t = \frac{t_{1/2}}{\ln 0.5} \ln P$$

Plugging in $P = 0.1$ and $t_{1/2} = 28$ years, we get $t = 93$ years.

(b) This is just like the first part, but $P = 10^{-27}$. The result is about 2500 years.

Example 14.1.3: ^{14}C Dating

Almost all the carbon on Earth is ^{12}C, but not quite. The isotope ^{14}C, with a half-life of 5600 years, is produced by cosmic rays in the atmosphere. It decays naturally, but is replenished at such a rate that the fraction of ^{14}C in the atmosphere remains constant, at 1.3×10^{-12}. Living plants and animals take in both ^{12}C and ^{14}C from the atmosphere and incorporate both into their bodies. Once the living organism dies, it no longer takes in C atoms from the atmosphere, and the proportion of ^{14}C gradually falls off as it undergoes radioactive decay. This effect can be used to find the age of dead organisms, or human artifacts made from plants or animals. Figure j shows the exponential decay curve of ^{14}C in various objects. Similar methods, using longer-lived isotopes, provided the first firm proof that the earth was billions of years old, not a few thousand as some had claimed on religious grounds.

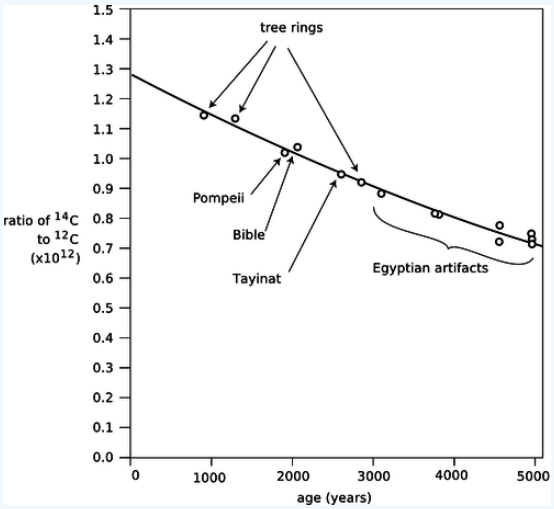


Figure* 14.1.8:** *Calibration of the ^{14}C dating method using tree rings and artifacts whose ages were known from other methods. Redrawn from Emilio Segrè,* ***Nuclei and Particles*, 1965.*

Rate of Decay

If you want to find how many radioactive decays occur within a time interval lasting from time t to time $t + \Delta t$, the most straightforward approach is to calculate it like this:

$$\begin{aligned}&(\text{number of decays between } t \text{ and } t + \Delta t)\\ &= N(t) - N(t + \Delta t)\end{aligned}$$

Usually we're interested in the case where Δt is small compared to $t_{1/2}$, and in this limiting case the calculation starts to look exactly like the limit that goes into the definition of the derivative dN/dt. It is therefore more convenient to talk about the *rate* of decay $-dN/dt$ rather than the *number* of decays in some finite time interval. Doing calculus on the function e^x is also easier than with 0.5^x, so we rewrite the function $N(t)$ as

$$N = N(0)e^{-t/\tau}, \tag{14.1.6}$$

where $\tau = t_{1/2}/\ln 2$ is shown in example 6 on p. 835 to be the average time of survival. The rate of decay is then

$$-\frac{dN}{dt} = \frac{N(0)}{\tau}e^{-t/\tau}. \tag{14.1.7}$$

Mathematically, differentating an exponential just gives back another exponential. Physically, this is telling us that as N falls off exponentially, the rate of decay falls off at the same exponential rate, because a lower N means fewer atoms that remain available to decay.

Exercise 14.1.1

Check that both sides of the equation for the rate of decay have units of s^{-1}, i.e., decays per unit time.

Answer

answer in the back of the PDF version of the book)

Example 14.1.4: The hot potato

A nuclear physicist with a demented sense of humor tosses you a cigar box, yelling "hot potato." The label on the box says "contains 10^{20} atoms of ^{17}F, half-life of 66 s, produced today in our reactor at 1 p.m." It takes you two seconds to read the label, after which you toss it behind some lead bricks and run away. The time is 1:40 p.m. Will you die?

Solution

The time elapsed since the radioactive fluorine was produced in the reactor was 40 minutes, or 2400 s. The number of elapsed half-lives is therefore $t/t_{1/2} = 36$. The initial number of atoms was $N(0) = 10^{20}$. The number of decays per second is now about $10^7\ \text{s}^{-1}$, so it produced about 2×10^7 high-energy electrons while you held it in your hands. Although twenty million electrons sounds like a lot, it is not really enough to be dangerous.

By the way, none of the equations we've derived so far was the actual probability distribution for the time at which a particular radioactive atom will decay. That probability distribution would be found by substituting $N(0) = 1$ into the equation for the rate of decay.

Discussion Questions

◊ In the medical procedure involving ^{131}I, why is it the gamma rays that are detected, not the electrons or neutrinos that are also emitted?

◊ For 1 s, Fred holds in his hands 1 kg of radioactive stuff with a half-life of 1000 years. Ginger holds 1 kg of a different substance, with a half-life of 1 min, for the same amount of time. Did they place themselves in equal danger, or not?

◊ How would you interpret it if you calculated $N(t)$, and found it was less than one?

◊ Does the half-life depend on how much of the substance you have? Does the expected time until the sample decays completely depend on how much of the substance you have?

13.1.5 Applications of Calculus

The area under the probability distribution is of course an integral. If we call the random number x and the probability distribution $D(x)$, then the probability that x lies in a certain range is given by

$$(\text{probability of } a \le x \le b) = \int_a^b D(x)dx. \tag{14.1.8}$$

What about averages? If x had a finite number of equally probable values, we would simply add them up and divide by how many we had. If they weren't equally likely, we'd make the weighted average $x_1P_1 + x_2P_2 + ...$ But we need to generalize this to a variable x that can take on any of a continuum of values. The continuous version of a sum is an integral, so the average is

$$(\text{average value of } x) = \int xD(x)dx, \tag{14.1.9}$$

where the integral is over all possible values of x.

Example 14.1.5: Probability distribution for radioactive decay

Here is a rigorous justification for the statement in subsection 13.1.4 that the probability distribution for radioactive decay is found by substituting $N(0) = 1$ into the equation for the rate of decay. We know that the probability distribution must be of the form

$$D(t) = k0.5^{t/t_{1/2}}, \tag{14.1.10}$$

where k is a constant that we need to determine. The atom is guaranteed to decay eventually, so normalization gives us

$$\begin{aligned}(\text{probability of } 0 \le t < \infty) &= 1 \\ &= \int_0^\infty D(t)dt.\end{aligned}$$

The integral is most easily evaluated by converting the function into an exponential with e as the base

$$\begin{aligned}D(t) &= k\exp\left[\ln\left(0.5^{t/t_{1/2}}\right)\right] \\ &= k\exp\left[\frac{t}{t_{1/2}}\ln 0.5\right] \\ &= k\exp\left(-\frac{\ln 2}{t_{1/2}}t\right),\end{aligned}$$

which gives an integral of the familiar form $\int e^{cx}dx = (1/c)e^{cx}$. We thus have

$$1 = -\frac{kt_{1/2}}{\ln 2}\exp\left(-\frac{\ln 2}{t_{1/2}}t\right)\Bigg]_0^\infty, \tag{14.1.11}$$

which gives the desired result:

$$k = \frac{\ln 2}{t_{1/2}}. \tag{14.1.12}$$

Example 14.1.6: Average lifetime

You might think that the half-life would also be the average lifetime of an atom, since half the atoms' lives are shorter and half longer. But the half whose lives are longer include some that survive for many half-lives, and these rare long-lived atoms skew the average. We can calculate the average lifetime as follows:

$$(\text{average lifetime}) = \int_0^\infty t\,D(t)dt \tag{14.1.13}$$

Using the convenient base-e form again, we have

$$(\text{average lifetime}) = \frac{\ln 2}{t_{1/2}}\int_0^\infty t\exp\left(-\frac{\ln 2}{t_{1/2}}t\right)dt. \tag{14.1.14}$$

This integral is of a form that can either be attacked with integration by parts or by looking it up in a table. The result is $\int xe^{cx}dx = \frac{x}{c}e^{cx} - \frac{1}{c^2}e^{cx}$, and the first term can be ignored for our purposes because it equals zero at both limits of integration. We end up with

$$\begin{aligned}(\text{average lifetime}) &= \frac{\ln 2}{t_{1/2}}\left(\frac{t_{1/2}}{\ln 2}\right)^2 \\ &= \frac{t_{1/2}}{\ln 2} \\ &= 1.443\,t_{1/2},\end{aligned}$$

which is, as expected, longer than one half-life.

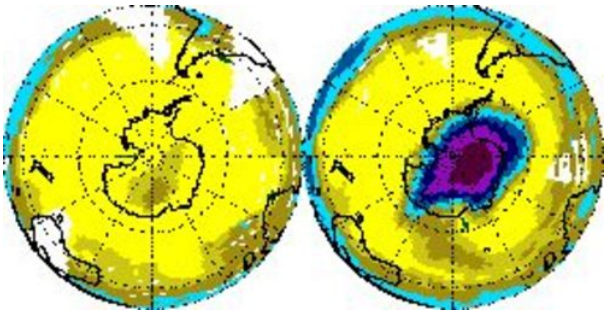


Figure* 14.1.9*: *In recent decades, a huge hole in the ozone layer has spread out from Antarctica. Left: November 1978. Right: November 1992*

Contributors and Attributions

- Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

14.2: Light As a Particle

The only thing that interferes with my learning is my education. -- *Albert Einstein*

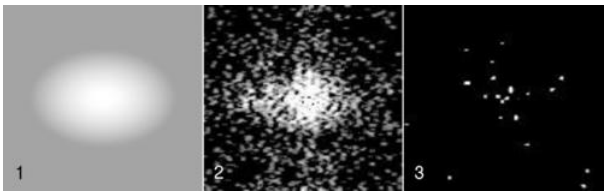
Radioactivity is random, but do the laws of physics exhibit randomness in other contexts besides radioactivity? Yes. Radioactive decay was just a good playpen to get us started with concepts of randomness, because all atoms of a given isotope are identical. By stocking the playpen with an unlimited supply of identical atom-toys, nature helped us to realize that their future behavior could be different regardless of their original identicality. We are now ready to leave the playpen, and see how randomness fits into the structure of physics at the most fundamental level.

The laws of physics describe light and matter, and the quantum revolution rewrote both descriptions. Radioactivity was a good example of matter's behaving in a way that was inconsistent with classical physics, but if we want to get under the hood and understand how nonclassical things happen, it will be easier to focus on light rather than matter. A radioactive atom such as uranium-235 is after all an extremely complex system, consisting of 92 protons, 143 neutrons, and 92 electrons. Light, however, can be a simple sine wave.

However successful the classical wave theory of light had been --- allowing the creation of radio and radar, for example --- it still failed to describe many important phenomena. An example that is currently of great interest is the way the ozone layer protects us from the dangerous short-wavelength ultraviolet part of the sun's spectrum. In the classical description, light is a wave. When a wave passes into and back out of a medium, its frequency is unchanged, and although its wavelength is altered while it is in the medium, it returns to its original value when the wave reemerges. Luckily for us, this is not at all what ultraviolet light does when it passes through the ozone layer, or the layer would offer no protection at all!

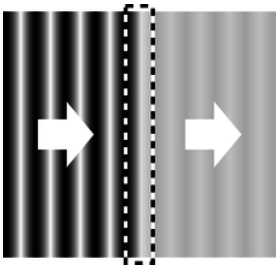
13.2.1 Evidence for light as a particle

For a long time, physicists tried to explain away the problems with the classical theory of light as arising from an imperfect understanding of atoms and the interaction of light with individual atoms and molecules. The ozone paradox, for example, could have been attributed to the incorrect assumption that one could think of the ozone layer as a smooth, continuous substance, when in reality it was made of individual ozone molecules. It wasn't until 1905 that Albert Einstein threw down the gauntlet, proposing that the problem had nothing to do with the details of light's interaction with atoms and everything to do with the fundamental nature of light itself.



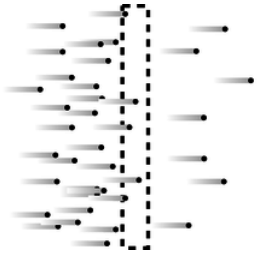
a / Digital camera images of dimmer and dimmer sources of light. The dots are records of individual photons.

In those days the data were sketchy, the ideas vague, and the experiments difficult to interpret; it took a genius like Einstein to cut through the thicket of confusion and find a simple solution. Today, however, we can get right to the heart of the matter with a piece of ordinary consumer electronics, the digital camera. Instead of film, a digital camera has a computer chip with its surface divided up into a grid of light-sensitive squares, called “pixels.” Compared to a grain of the silver compound used to make regular photographic film, a digital camera pixel is activated by an amount of light energy orders of magnitude smaller. We can learn something new about light by using a digital camera to detect smaller and smaller amounts of light, as shown in figure a. Figure a/1 is fake, but a/2 and a/3 are real digital-camera images made by Prof. Lyman Page of Princeton University as a classroom demonstration. Figure a/1 is what we would see if we used the digital camera to take a picture of a fairly dim source of light. In figures a/2 and a/3, the intensity of the light was drastically reduced by inserting semitransparent absorbers like the tinted plastic used in sunglasses. Going from a/1 to a/2 to a/3, more and more light energy is being thrown away by the absorbers.



b / A wave is partially absorbed.

The results are drastically different from what we would expect based on the wave theory of light. If light was a wave and nothing but a wave, b, then the absorbers would simply cut down the wave's amplitude across the whole wavefront. The digital camera's entire chip would be illuminated uniformly, and weakening the wave with an absorber would just mean that every pixel would take a long time to soak up enough energy to register a signal.



c / A stream of particles is partially absorbed.

But figures a/2 and a/3 show that some pixels take strong hits while others pick up no energy at all. Instead of the wave picture, the image that is naturally evoked by the data is something more like a hail of bullets from a machine gun, c. Each “bullet” of light apparently carries only a tiny amount of energy, which is why detecting them individually requires a sensitive digital camera rather than an eye or a piece of film.

Although Einstein was interpreting different observations, this is the conclusion he reached in his 1905 paper: that the pure wave theory of light is an oversimplification, and that the energy of a beam of light comes in finite chunks rather than being spread smoothly throughout a region of space.

We now think of these chunks as particles of light, and call them “photons,” although Einstein avoided the word “particle,” and the word “photon” was invented later. Regardless of words, the trouble was that waves and particles seemed like inconsistent categories. The reaction to Einstein's paper could be kindly described as vigorously skeptical. Even twenty years later, Einstein wrote, “There are therefore now two theories of light, both indispensable, and --- as one must admit today despite twenty years of tremendous effort on the part of theoretical physicists --- without any logical connection.” In the remainder of this chapter we will learn how the seeming paradox was eventually resolved.

d / Einstein and Seurat: twins separated at birth? *Seine Grande Jatte* by Georges Seurat (19th century).

Discussion Questions

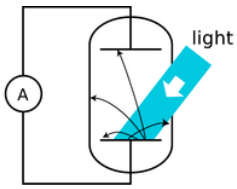
◊ Suppose someone rebuts the digital camera data in figure a, claiming that the random pattern of dots occurs not because of anything fundamental about the nature of light but simply because the camera's pixels are not all exactly the same --- some are just more sensitive than others. How could we test this interpretation?

◊ Discuss how the correspondence principle applies to the observations and concepts discussed in this section.

13.2.2 How much light is one photon?

The photoelectric effect

We have seen evidence that light energy comes in little chunks, so the next question to be asked is naturally how much energy is in one chunk. The most straightforward experimental avenue for addressing this question is a phenomenon known as the photoelectric effect. The photoelectric effect occurs when a photon strikes the surface of a solid object and knocks out an electron. It occurs continually all around you. It is happening right now at the surface of your skin and on the paper or computer screen from which you are reading these words. It does not ordinarily lead to any observable electrical effect, however, because on the average free electrons are wandering back in just as frequently as they are being ejected. (If an object did somehow lose a significant number of electrons, its growing net positive charge would begin attracting the electrons back more and more strongly.)



e / Apparatus for observing the photoelectric effect. A beam of light strikes a capacitor plate inside a vacuum tube, and electrons are ejected (black arrows).

Figure e shows a practical method for detecting the photoelectric effect. Two very clean parallel metal plates (the electrodes of a capacitor) are sealed inside a vacuum tube, and only one plate is exposed to light. Because there is a good vacuum between the plates, any ejected electron that happens to be headed in the right direction will almost certainly reach the other capacitor plate without colliding with any air molecules.

The illuminated (bottom) plate is left with a net positive charge, and the unilluminated (top) plate acquires a negative charge from the electrons deposited on it. There is thus an electric field between the plates, and it is because of this field that the electrons' paths are curved, as shown in the diagram. However, since vacuum is a good insulator, any electrons that reach the top plate are prevented from responding to the electrical attraction by jumping back across the gap. Instead they are forced to make their way around the circuit, passing through an ammeter. The ammeter allows a measurement of the strength of the photoelectric effect.

An unexpected dependence on frequency

The photoelectric effect was discovered serendipitously by Heinrich Hertz in 1887, as he was experimenting with radio waves. He was not particularly interested in the phenomenon, but he did notice that the effect was produced strongly by ultraviolet light and more weakly by lower frequencies. Light whose frequency was lower than a certain critical value did not eject any electrons at all. (In fact this was all prior to Thomson's discovery of the electron, so Hertz would not have described the effect in terms of electrons --- we are discussing everything with the benefit of hindsight.) This dependence on frequency didn't make any sense in terms of the classical wave theory of light. A light wave consists of electric and magnetic fields. The stronger the fields, i.e., the greater the wave's amplitude, the greater the forces that would be exerted on electrons that found themselves bathed in the light. It should have been amplitude (brightness) that was relevant, not frequency. The dependence on frequency not only proves that the wave model of light needs modifying, but with the proper interpretation it allows us to determine how much energy is in one photon, and it also leads to a connection between the wave and particle models that we need in order to reconcile them.

To make any progress, we need to consider the physical process by which a photon would eject an electron from the metal electrode. A metal contains electrons that are free to move around. Ordinarily, in the interior of the metal, such an electron feels attractive forces from atoms in every direction around it. The forces cancel out. But if the electron happens to find itself at the surface of the metal, the attraction from the interior side is not balanced out by any attraction from outside. In popping out through the surface the electron therefore loses some amount of energy E_s, which depends on the type of metal used.

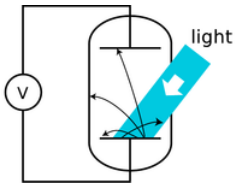
f / The hamster in her hamster ball is like an electron emerging from the metal (tiled kitchen floor) into the surrounding vacuum (wood floor). The wood floor is higher than the tiled floor, so as she rolls up the step, the hamster will lose a certain amount of kinetic energy, analogous to E_s. If her kinetic energy is too small, she won't even make it up the step.

Suppose a photon strikes an electron, annihilating itself and giving up all its energy to the electron. (We now know that this is what always happens in the photoelectric effect, although it had not yet been established in 1905 whether or not the photon was completely annihilated.) The electron will (1) lose kinetic energy through collisions with other electrons as it plows through the metal on its way to the surface; (2) lose an amount of kinetic energy equal to E_s as it emerges through the surface; and (3) lose more energy on its way across the gap between the plates, due to the electric field between the plates. Even if the electron happens to be right at the surface of the metal when it absorbs the photon, and even if the electric field between the plates has not yet built up very much, E_s is the bare minimum amount of energy that it must receive from the photon if it is to contribute to a measurable current. The reason for using very clean electrodes is to minimize E_s and make it have a definite value characteristic of the metal surface, not a mixture of values due to the various types of dirt and crud that are present in tiny amounts on all surfaces in everyday life.

We can now interpret the frequency dependence of the photoelectric effect in a simple way: apparently the amount of energy possessed by a photon is related to its frequency. A low-frequency red or infrared photon has an energy less than E_s, so a beam of them will not produce any current. A high-frequency blue or violet photon, on the other hand, packs enough of a punch to allow an electron to make it to the other plate. At frequencies higher than the minimum, the photoelectric current continues to increase with the frequency of the light because of effects (1) and (3).

Numerical relationship between energy and frequency

Prompted by Einstein's photon paper, Robert Millikan (whom we first encountered in chapter 8) figured out how to use the photoelectric effect to probe precisely the link between frequency and photon energy. Rather than going into the historical details of Millikan's actual experiments (a lengthy experimental program that occupied a large part of his professional career) we will describe a simple version, shown in figure g, that is used sometimes in college laboratory courses.[2] The idea is simply to illuminate one plate of the vacuum tube with light of a single wavelength and monitor the voltage difference between the two plates as they charge up. Since the resistance of a voltmeter is very high (much higher than the resistance of an ammeter), we can assume to a good approximation that electrons reaching the top plate are stuck there permanently, so the voltage will keep on increasing for as long as electrons are making it across the vacuum tube.

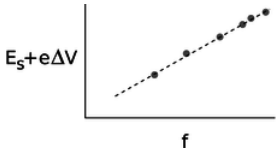


g / A different way of studying the photoelectric effect.

At a moment when the voltage difference has a reached a value ΔV, the minimum energy required by an electron to make it out of the bottom plate and across the gap to the other plate is $E_s + e\Delta$ V. As ΔV increases, we eventually reach a point at which $E_s + e\Delta V$ *equals* the energy of one photon. No more electrons can cross the gap, and the reading on the voltmeter stops rising. The quantity $E_s + e\Delta V$ now tells us the energy of one photon. If we determine this energy for a variety of wavelengths, h, we find the following simple relationship between the energy of a photon and the frequency of the light:

$$E = hf,$$

where h is a constant with the value 6.63×10^{-34} J $\cdot$ s. Note how the equation brings the wave and particle models of light under the same roof: the left side is the energy of one *particle* of light, while the right side is the frequency of the same light, interpreted as a *wave*. The constant h is known as Planck's constant, for historical reasons explained in the footnote beginning on the preceding page.



h / The quantity $E_s + e\Delta V$ indicates the energy of one photon. It is found to be proportional to the frequency of the light.

self-check:

How would you extract h from the graph in figure h? What if you didn't even know E_s in advance, and could only graph $e\Delta V$ versus f?

(answer in the back of the PDF version of the book)

Since the energy of a photon is hf, a beam of light can only have energies of hf, $2hf$, $3hf$, etc. Its energy is quantized --- there is no such thing as a fraction of a photon. Quantum physics gets its name from the fact that it quantizes quantities like energy, momentum, and angular momentum that had previously been thought to be smooth, continuous and infinitely divisible.

Example 7: Number of photons emitted by a lightbulb per second

▷ Roughly how many photons are emitted by a 100-W lightbulb in 1 second?

▷ People tend to remember wavelengths rather than frequencies for visible light. The bulb emits photons with a range of frequencies and wavelengths, but let's take 600 nm as a typical wavelength for purposes of estimation. The energy of a single photon is

$$\begin{aligned} E_{photon} &= hf \\ &= hc/\lambda \end{aligned}$$

A power of 100 W means 100 joules per second, so the number of photons is

$$\begin{aligned} (100\text{ J})/E_{photon} &= (100\text{ J})/(hc/\lambda) \\ &\approx 3\times 10^{20} \end{aligned}$$

This hugeness of this number is consistent with the correspondence principle. The experiments that established the classical theory of optics weren't wrong. They were right, within their domain of applicability, in which the number of photons was so large as to be indistinguishable from a continuous beam.

Example 8: Measuring the wave

When surfers are out on the water waiting for their chance to catch a wave, they're interested in both the height of the waves and when the waves are going to arrive. In other words, they observe both the amplitude and phase of the waves, and it doesn't matter to them that the water is granular at the molecular level. The correspondence principle requires that we be able to do the same thing for electromagnetic waves, since the classical theory of electricity and magnetism was all stated and verified experimentally in terms of the fields $\mathbf{E}$ and $\mathbf{B}$, which are the amplitude of an electromagnetic wave. The phase is also necessary, since the induction effects predicted by Maxwell's equation would flip their signs depending on whether an oscillating field is on its way up or on its way back down.

This is a more demanding application of the correspondence principle than the one in example 7, since amplitudes and phases constitute more detailed information than the over-all intensity of a beam of light. Eyeball measurements can't detect this type of information, since the eye is much bigger than a wavelength, but for example an AM radio receiver can do it with radio waves, since the wavelength for a station at 1000 kHz is about 300 meters, which is much larger than the antenna. The correspondence principle demands that we be able to explain this in terms of the photon theory, and this requires not just that we have a large number of photons emitted by the transmitter per second, as in example 7, but that even by the time they spread out and reach the receiving antenna, there should be many photons overlapping each other within a space of one cubic wavelength. Problem 47 on p. 903 verifies that the number is in fact extremely large.

Example 9: Momentum of a photon

▷ According to the theory of relativity, the momentum of a beam of light is given by $p = E/c$. Apply this to find the momentum of a single photon in terms of its frequency, and in terms of its wavelength.
▷ Combining the equations $p = E/c$ and $E = hf$, we find

$$\begin{aligned} p &= E/c \\ &= \frac{h}{c} f. \end{aligned}$$

To reexpress this in terms of wavelength, we use $c = f\lambda$:

$$\begin{aligned} p &= \frac{h}{c} \cdot \frac{c}{\lambda} \\ &= \frac{h}{\lambda} \end{aligned}$$

The second form turns out to be simpler.

Discussion Questions

◊ The photoelectric effect only ever ejects a very tiny percentage of the electrons available near the surface of an object. How well does this agree with the wave model of light, and how well with the particle model? Consider the two different distance scales involved: the wavelength of the light, and the size of an atom, which is on the order of 10^{-10} or 10^{-9} m.

◊ What is the significance of the fact that Planck's constant is numerically very small? How would our everyday experience of light be different if it was not so small?

◊ How would the experiments described above be affected if a single electron was likely to get hit by more than one photon?

◊ Draw some representative trajectories of electrons for $\Delta V = 0$, ΔV less than the maximum value, and ΔV greater than the maximum value.

◊ Explain based on the photon theory of light why ultraviolet light would be more likely than visible or infrared light to cause cancer by damaging DNA molecules. How does this relate to discussion question C?

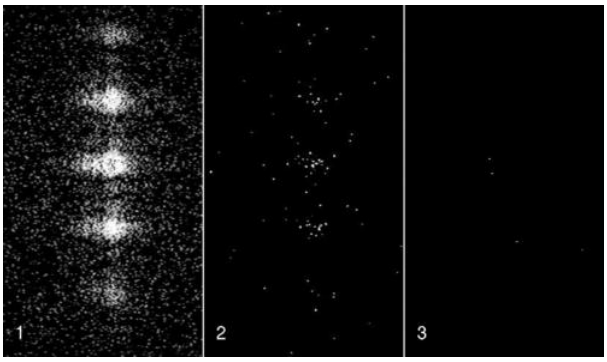
◊ Does $E = hf$ imply that a photon changes its energy when it passes from one transparent material into another substance with a different index of refraction?

13.2.3 Wave-particle duality

How can light be both a particle and a wave? We are now ready to resolve this seeming contradiction. Often in science when something seems paradoxical, it's because we (1) don't define our terms carefully, or (2) don't test our ideas against any specific real-world situation. Let's define particles and waves as follows:

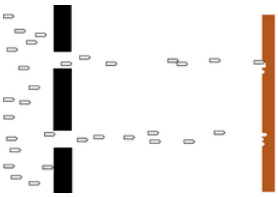
- Waves exhibit superposition, and specifically interference phenomena.
- Particles can only exist in whole numbers, not fractions.

As a real-world check on our philosophizing, there is one particular experiment that works perfectly. We set up a double-slit interference experiment that we know will produce a diffraction pattern if light is an honest-to-goodness wave, but we detect the light with a detector that is capable of sensing individual photons, e.g., a digital camera. To make it possible to pick out individual dots due to individual photons, we must use filters to cut down the intensity of the light to a very low level, just as in the photos by Prof. Page on p. 837. The whole thing is sealed inside a light-tight box. The results are shown in figure i. (In fact, the similar figures in on page 837 are simply cutouts from these figures.)



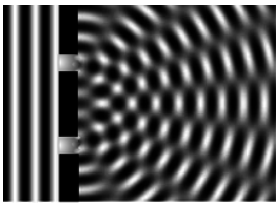
i / Wave interference patterns photographed by Prof. Lyman Page with a digital camera. Laser light with a single well-defined wavelength passed through a series of absorbers to cut down its intensity, then through a set of slits to produce interference, and finally into a digital camera chip. (A triple slit was actually used, but for conceptual simplicity we discuss the results in the main text as if it was a double slit.) In panel 2 the intensity has been reduced relative to 1, and even more so for panel 3.

Neither the pure wave theory nor the pure particle theory can explain the results. If light was only a particle and not a wave, there would be no interference effect. The result of the experiment would be like firing a hail of bullets through a double slit, j. Only two spots directly behind the slits would be hit.



j / Bullets pass through a double slit.

If, on the other hand, light was only a wave and not a particle, we would get the same kind of diffraction pattern that would happen with a water wave, k. There would be no discrete dots in the photo, only a diffraction pattern that shaded smoothly between light and dark.



k / A water wave passes through a double slit.

Applying the definitions to this experiment, light must be both a particle and a wave. It is a wave because it exhibits interference effects. At the same time, the fact that the photographs contain discrete dots is a direct demonstration that light refuses to be split into units of less than a single photon. There can only be whole numbers of photons: four photons in figure i/3, for example.

A wrong interpretation: photons interfering with each other

One possible interpretation of wave-particle duality that occurred to physicists early in the game was that perhaps the interference effects came from photons interacting with each other. By analogy, a water wave consists of moving water molecules, and interference of water waves results ultimately from all the mutual pushes and pulls of the molecules. This interpretation was conclusively disproved by G.I. Taylor, a student at Cambridge. The demonstration by Prof. Page that we've just been discussing is essentially a modernized version of Taylor's work. Taylor reasoned that if interference effects came from photons interacting with each other, a bare minimum of two photons would have to be present at the same time to produce interference. By making the light source extremely dim, we can be virtually certain that there are never two photons in the box at the same time. In figure i/3, however, the intensity of the light has been cut down so much by the absorbers that if it was in the open, the average separation between photons would be on the order of a kilometer! At any given moment, the number of photons in the box is most likely to be zero. It is virtually certain that there were never two photons in the box at once.

l / A single photon can go through both slits.

The concept of a photon's path is undefined.

If a single photon can demonstrate double-slit interference, then which slit did it pass through? The unavoidable answer must be that it passes through both! This might not seem so strange if we think of the photon as a wave, but it is highly counterintuitive if we try to visualize it as a particle. The moral is that we should not think in terms of the path of a photon. Like the fully human and fully divine Jesus of Christian theology, a photon is supposed to be 100% wave and 100% particle. If a photon had a well defined path, then it would not demonstrate wave superposition and interference effects, contradicting its wave nature. (In subsection 13.3.4 we will discuss the Heisenberg uncertainty principle, which gives a numerical way of approaching this issue.)

Another wrong interpretation: the pilot wave hypothesis

A second possible explanation of wave-particle duality was taken seriously in the early history of quantum mechanics. What if the photon *particle* is like a surfer riding on top of its accompanying *wave*? As the wave travels along, the particle is pushed, or "piloted" by it. Imagining the particle and the wave as two separate entities allows us to avoid the seemingly paradoxical idea that a photon is both at once. The wave happily does its wave tricks, like superposition and interference, and the particle acts like a respectable particle, resolutely refusing to be in two different places at once. If the wave, for instance, undergoes destructive interference, becoming nearly zero in a particular region of space, then the particle simply is not guided into that region.

The problem with the pilot wave interpretation is that the only way it can be experimentally tested or verified is if someone manages to detach the particle from the wave, and show that there really are two entities involved, not just one. Part of the scientific method is that hypotheses are supposed to be experimentally testable. Since nobody has ever managed to separate the wavelike part of a photon from the particle part, the interpretation is not useful or meaningful in a scientific sense.

The probability interpretation

The correct interpretation of wave-particle duality is suggested by the random nature of the experiment we've been discussing: even though every photon wave/particle is prepared and released in the same way, the location at which it is eventually detected by the digital camera is different every time. The idea of the probability interpretation of wave-particle duality is that the location of the photon-particle is random, but the probability that it is in a certain location is higher where the photon-wave's amplitude is greater.

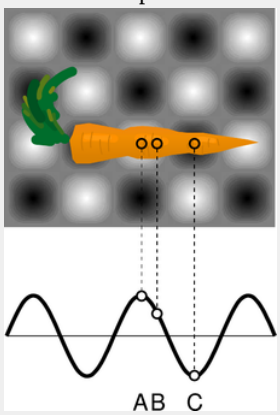
More specifically, the probability distribution of the particle must be proportional to the *square* of the wave's amplitude,

$$(\text{probability distribution}) \propto (\text{amplitude})^2.$$

This follows from the correspondence principle and from the fact that a wave's energy density is proportional to the square of its amplitude. If we run the double-slit experiment for a long enough time, the pattern of dots fills in and becomes very smooth as would have been expected in classical physics. To preserve the correspondence between classical and quantum physics, the amount of energy deposited in a given region of the picture over the long run must be proportional to the square of the wave's amplitude. The amount of energy deposited in a certain area depends on the number of photons picked up, which is proportional to the probability of finding any given photon there.

Example 10: A microwave oven

▷ The figure shows two-dimensional (top) and one-dimensional (bottom) representations of the standing wave inside a microwave oven. Gray represents zero field, and white and black signify the strongest fields, with white being a field that is in the opposite direction compared to black. Compare the probabilities of detecting a microwave photon at points A, B, and C.



m / Example 10.

▷ A and C are both extremes of the wave, so the probabilities of detecting a photon at A and C are equal. It doesn't matter that we have represented C as negative and A as positive, because it is the square of the amplitude that is relevant. The amplitude at B is about 1/2 as much as the others, so the probability of detecting a photon there is about 1/4 as much.

The probability interpretation was disturbing to physicists who had spent their previous careers working in the deterministic world of classical physics, and ironically the most strenuous objections against it were raised by Einstein, who had invented the photon concept in the first place. The probability interpretation has nevertheless passed every experimental test, and is now as well established as any part of physics.

An aspect of the probability interpretation that has made many people uneasy is that the process of detecting and recording the photon's position seems to have a magical ability to get rid of the wavelike side of the photon's personality and force it to decide for once and for all where it really wants to be. But detection or measurement is after all only a physical process like any other, governed by the same laws of physics. We will postpone a detailed discussion of this issue until p. 864, since a measuring device like a digital camera is made of matter, but we have so far only discussed how quantum mechanics relates to light.

Example 11: What is the proportionality constant?

▷ What is the proportionality constant that would make an actual equation out of $(\text{probability distribution}) \propto (\text{amplitude})^2$?

▷ The probability that the photon is in a certain small region of volume v should equal the fraction of the wave's energy that is within that volume. For a sinusoidal wave, which has a single, well-defined frequency f, this gives

$$\begin{aligned} P &= \frac{\text{energy in volume } v}{\text{energy of photon}} \\ &= \frac{\text{energy in volume } v}{hf}. \end{aligned}$$

We assume v is small enough so that the electric and magnetic fields are nearly constant throughout it. We then have

$$P = \frac{\left(\frac{1}{8\pi k}|\mathbf{E}|^2 + \frac{c^2}{8\pi k}|\mathbf{B}|^2\right) v}{hf}.$$

We can simplify this formidable looking expression by recognizing that in a plane wave, $|\mathbf{E}|$ and $|\mathbf{B}|$ are related by $|\mathbf{E}| = c|\mathbf{B}|$. This implies (problem 40, p. 725), that the electric and magnetic fields each contribute half the total energy, so we can simplify the result to

$$\begin{aligned} P &= 2\frac{\left(\frac{1}{8\pi k}|\mathbf{E}|^2\right) v}{hf} \\ &= \frac{v}{4\pi khf}|\mathbf{E}|^2. \end{aligned}$$

The probability is proportional to the square of the wave's amplitude, as advertised.[3]

Discussion Questions

◊ Referring back to the example of the carrot in the microwave oven, show that it would be nonsensical to have probability be proportional to the field itself, rather than the square of the field.

◊ Einstein did not try to reconcile the wave and particle theories of light, and did not say much about their apparent inconsistency. Einstein basically visualized a beam of light as a stream of bullets coming from a machine gun. In the photoelectric effect, a photon "bullet" would only hit one atom, just as a real bullet would only hit one person. Suppose someone reading his 1905 paper wanted to interpret it by saying that Einstein's so-called particles of light are simply short wave-trains that only occupy a small region of space. Comparing the wavelength of visible light (a few hundred nm) to the size of an atom (on the order of 0.1 nm), explain why this poses a difficulty for reconciling the particle and wave theories.

◊ Can a white photon exist?

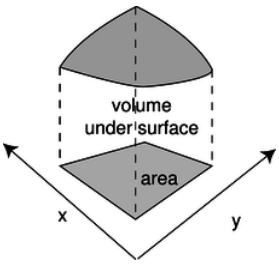
◊ In double-slit diffraction of photons, would you get the same pattern of dots on the digital camera image if you covered one slit? Why should it matter whether you give the photon two choices or only one?

13.2.4 Photons in three dimensions

Up until now I've been sneaky and avoided a full discussion of the three-dimensional aspects of the probability interpretation. The example of the carrot in the microwave oven, for example, reduced to a one-dimensional situation because we were considering three points along the same line and because we were only comparing ratios of probabilities. The purpose of bringing it up now is to head off any feeling that you've been cheated conceptually rather than to prepare you for mathematical problem solving in three dimensions, which would not be appropriate for the level of this course.

A typical example of a probability distribution in section 13.1 was the distribution of heights of human beings. The thing that varied randomly, height, h, had units of meters, and the probability distribution was a graph of a function $D(h)$. The units of the probability distribution had to be m^{-1} (inverse meters) so that areas under the curve, interpreted as probabilities, would be unitless: $(\mathrm{area}) = (\mathrm{height})(\mathrm{width}) = \mathrm{m}^{-1} \cdot \mathrm{m}$.

Now suppose we have a two-dimensional problem, e.g., the probability distribution for the place on the surface of a digital camera chip where a photon will be detected. The point where it is detected would be described with two variables, x and y, each having units of meters. The probability distribution will be a function of both variables, $D(x, y)$. A probability is now visualized as the volume under the surface described by the function $D(x, y)$, as shown in figure n. The units of D must be m^{-2} so that probabilities will be unitless: $(\mathrm{probability}) = (\mathrm{depth})(\mathrm{length})(\mathrm{width}) = \mathrm{m}^{-2} \cdot \mathrm{m} \cdot \mathrm{m}$. In terms of calculus, we have $P = \int D dx dy$.



n / Probability is the volume under a surface defined by $D(x, y)$.

Generalizing finally to three dimensions, we find by analogy that the probability distribution will be a function of all three coordinates, $D(x, y, z)$, and will have units of m^{-3}. It is unfortunately impossible to visualize the graph unless you are a mutant with a natural feel for life in four dimensions. If the probability distribution is nearly constant within a certain volume of space v, the probability that the photon is in that volume is simply vD. If not, then we can use an integral, $P = \int D dx dy dz$.

Contributors

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

14.3: Matter As a Wave

[In] a few minutes I shall be all melted... I have been wicked in my day, but I never thought a little girl like you would ever be able to melt me and end my wicked deeds. Look out --- here I go! -- *The Wicked Witch of the West*

As the Wicked Witch learned the hard way, losing molecular cohesion can be unpleasant. That's why we should be very grateful that the concepts of quantum physics apply to matter as well as light. If matter obeyed the laws of classical physics, molecules wouldn't exist.

Consider, for example, the simplest atom, hydrogen. Why does one hydrogen atom form a chemical bond with another hydrogen atom? Roughly speaking, we'd expect a neighboring pair of hydrogen atoms, A and B, to exert no force on each other at all, attractive or repulsive: there are two repulsive interactions (proton A with proton B and electron A with electron B) and two attractive interactions (proton A with electron B and electron A with proton B). Thinking a little more precisely, we should even expect that once the two atoms got close enough, the interaction would be repulsive. For instance, if you squeezed them so close together that the two protons were almost on top of each other, there would be a tremendously strong repulsion between them due to the $1/r^2$ nature of the electrical force. The repulsion between the electrons would not be as strong, because each electron ranges over a large area, and is not likely to be found right on top of the other electron. Thus hydrogen molecules should not exist according to classical physics.

Quantum physics to the rescue! As we'll see shortly, the whole problem is solved by applying the same quantum concepts to electrons that we have already used for photons.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

14.4: The Atom

You can learn a lot by taking a car engine apart, but you will have learned a lot more if you can put it all back together again and make it run. Half the job of reductionism is to break nature down into its smallest parts and understand the rules those parts obey. The second half is to show how those parts go together, and that is our goal in this chapter. We have seen how certain features of all atoms can be explained on a generic basis in terms of the properties of bound states, but this kind of argument clearly cannot tell us any details of the behavior of an atom or explain why one atom acts differently from another.

The biggest embarrassment for reductionists is that the job of putting things back together job is usually much harder than the taking them apart. Seventy years after the fundamentals of atomic physics were solved, it is only beginning to be possible to calculate accurately the properties of atoms that have many electrons. Systems consisting of many atoms are even harder. Supercomputer manufacturers point to the folding of large protein molecules as a process whose calculation is just barely feasible with their fastest machines. The goal of this chapter is to give a gentle and visually oriented guide to some of the simpler results about atoms.

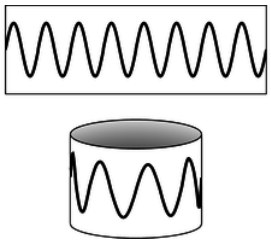
Classifying States

We'll focus our attention first on the simplest atom, hydrogen, with one proton and one electron. We know in advance a little of what we should expect for the structure of this atom. Since the electron is bound to the proton by electrical forces, it should display a set of discrete energy states, each corresponding to a certain standing wave pattern. We need to understand what states there are and what their properties are.

What properties should we use to classify the states? The most sensible approach is to used conserved quantities. Energy is one conserved quantity, and we already know to expect each state to have a specific energy. It turns out, however, that energy alone is not sufficient. Different standing wave patterns of the atom can have the same energy.

Momentum is also a conserved quantity, but it is not particularly appropriate for classifying the states of the electron in a hydrogen atom. The reason is that the force between the electron and the proton results in the continual exchange of momentum between them. (Why wasn't this a problem for energy as well? Kinetic energy and momentum are related by $K = p^2/2m$, so the much more massive proton never has very much kinetic energy. We are making an approximation by assuming all the kinetic energy is in the electron, but it is quite a good approximation.)

Angular momentum does help with classification. There is no transfer of angular momentum between the proton and the electron, since the force between them is a center-to-center force, producing no torque.



a / Eight wavelengths fit around this circle ($\ell = 8$).

Like energy, angular momentum is quantized in quantum physics. As an example, consider a quantum wave-particle confined to a circle, like a wave in a circular moat surrounding a castle. A sine wave in such a “quantum moat” cannot have any old wavelength, because an integer number of wavelengths must fit around the circumference, C, of the moat. The larger this integer is, the shorter the wavelength, and a shorter wavelength relates to greater momentum and angular momentum. Since this integer is related to angular momentum, we use the symbol ℓ for it:

$$\lambda = C/\ell$$

The angular momentum is

$$L = rp.$$

Here, $r = C/2\pi$, and $p = h/\lambda = h\ell/C$, so

$$L = \frac{C}{2\pi} \cdot \frac{h\ell}{C}$$
$$= \frac{h}{2\pi}\ell$$

In the example of the quantum moat, angular momentum is quantized in units of $h/2\pi$. This makes $h/2\pi$ a pretty important number, so we define the abbreviation $\hbar = h/2\pi$. This symbol is read "h-bar."

In fact, this is a completely general fact in quantum physics, not just a fact about the quantum moat:

Quantization of angular momentum

The angular momentum of a particle due to its motion through space is quantized in units of $\hbar$.

Exercise 14.4.1

What is the angular momentum of the wavefunction shown at the beginning of the section?

Answer

(answer in the back of the PDF version of the book)

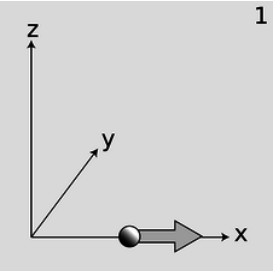
Three dimensions

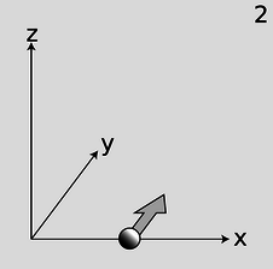
Our discussion of quantum-mechanical angular momentum has so far been limited to rotation in a plane, for which we can simply use positive and negative signs to indicate clockwise and counterclockwise directions of rotation. A hydrogen atom, however, is unavoidably three-dimensional. The classical treatment of angular momentum in three-dimensions has been presented in section 4.3; in general, the angular momentum of a particle is defined as the vector cross product $\mathbf{r} \times \mathbf{p}$.

There is a basic problem here: the angular momentum of the electron in a hydrogen atom depends on both its distance $\mathbf{r}$ from the proton and its momentum $\mathbf{p}$, so in order to know its angular momentum precisely it would seem we would need to know both its position and its momentum simultaneously with good accuracy. This, however, seems forbidden by the Heisenberg uncertainty principle.

Actually the uncertainty principle does place limits on what can be known about a particle's angular momentum vector, but it does not prevent us from knowing its magnitude as an exact integer multiple of $\hbar$. The reason is that in three dimensions, there are really three separate uncertainty principles:

$$\Delta p_x \Delta x \gtrsim h$$
$$\Delta p_y \Delta y \gtrsim h$$
$$\Delta p_z \Delta z \gtrsim h$$





b / Reconciling the uncertainty principle with the definition of angular momentum.

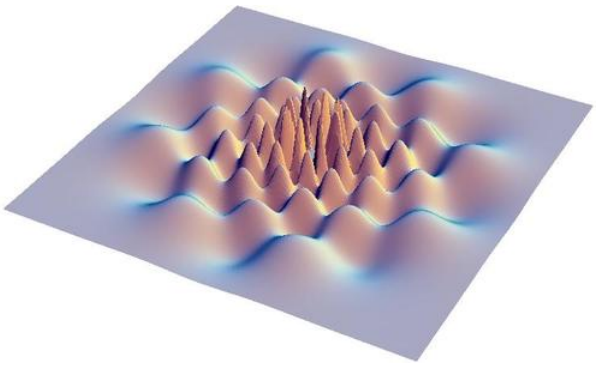
Now consider a particle, b/1, that is moving along the x axis at position x and with momentum p_x. We may not be able to know both x and p_x with unlimited accurately, but we can still know the particle's angular momentum about the origin exactly: it is zero, because the particle is moving directly away from the origin.

Suppose, on the other hand, a particle finds itself, b/2, at a position x along the x axis, and it is moving parallel to the y axis with momentum p_y. It has angular momentum xp_y about the z axis, and again we can know its angular momentum with unlimited accuracy, because the uncertainty principle only relates x to p_x and y to p_y. It does not relate x to p_y.

As shown by these examples, the uncertainty principle does not restrict the accuracy of our knowledge of angular momenta as severely as might be imagined. However, it does prevent us from knowing all three components of an angular momentum vector simultaneously. The most general statement about this is the following theorem, which we present without proof:

The angular momentum vector in quantum physics

The most that can be known about an angular momentum vector is its magnitude and one of its three vector components. Both are quantized in units of $\hbar$.



c / A cross-section of a hydrogen wavefunction.

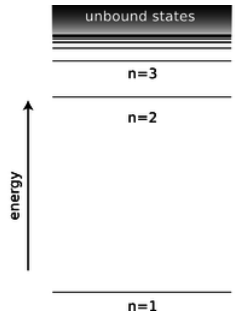
The hydrogen atom

Deriving the wavefunctions of the states of the hydrogen atom from first principles would be mathematically too complex for this book, but it's not hard to understand the logic behind such a wavefunction in visual terms. Consider the wavefunction from the beginning of the section, which is reproduced in figure c. Although the graph looks three-dimensional, it is really only a representation of the part of the wavefunction lying within a two-dimensional plane. The third (up-down) dimension of the plot

represents the value of the wavefunction at a given point, not the third dimension of space. The plane chosen for the graph is the one perpendicular to the angular momentum vector.

Each ring of peaks and valleys has eight wavelengths going around in a circle, so this state has $L = 8\hbar$, i.e., we label it $\ell = 8$. The wavelength is shorter near the center, and this makes sense because when the electron is close to the nucleus it has a lower electrical energy, a higher kinetic energy, and a higher momentum.

Between each ring of peaks in this wavefunction is a nodal circle, i.e., a circle on which the wavefunction is zero. The full three-dimensional wavefunction has nodal spheres: a series of nested spherical surfaces on which it is zero. The number of radii at which nodes occur, including $r = \infty$, is called n, and n turns out to be closely related to energy. The ground state has $n = 1$ (a single node only at $r = \infty$), and higher-energy states have higher n values. There is a simple equation relating n to energy, which we will discuss in subsection 13.4.4.



d / The energy of a state in the hydrogen atom depends only on its n quantum number.

The numbers n and ℓ, which identify the state, are called its quantum numbers. A state of a given n and ℓ can be oriented in a variety of directions in space. We might try to indicate the orientation using the three quantum numbers $\ell_x = L_x/\hbar$, $\ell_y = L_y/\hbar$, and $\ell_z = L_z/\hbar$. But we have already seen that it is impossible to know all three of these simultaneously. To give the most complete possible description of a state, we choose an arbitrary axis, say the z axis, and label the state according to n, ℓ, and ℓ_z.[6]

Angular momentum requires motion, and motion implies kinetic energy. Thus it is not possible to have a given amount of angular momentum without having a certain amount of kinetic energy as well. Since energy relates to the n quantum number, this means that for a given n value there will be a maximum possible . It turns out that this maximum value of equals $n - 1$.

In general, we can list the possible combinations of quantum numbers as follows:

n can equal 1, 2, 3, …
l can range from 0 to*n* − 1, in steps of 1
l_z can range fromell toell, in steps of 1

Applying these rules, we have the following list of states:

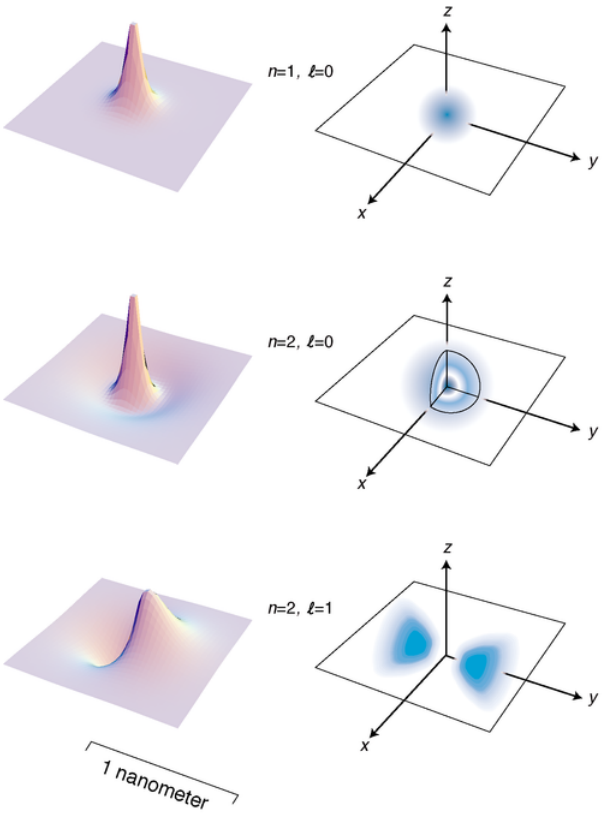
n = 1,	l=0,	l_z=0	one state
n = 2,	l=0,	l_z=0	one state
n = 2,	l=1,	l_z=-1, 0, or 1	three states

self-check:

Continue the list for $n = 3$.

(answer in the back of the PDF version of the book)

Figure e on page 882 shows the lowest-energy states of the hydrogen atom. The left-hand column of graphs displays the wavefunctions in the $x - y$ plane, and the right-hand column shows the probability distribution in a three-dimensional representation.



e / The three states of the hydrogen atom having the lowest energies.

Discussion Questions

◊ The quantum number n is defined as the number of radii at which the wavefunction is zero, including $r = \infty$. Relate this to the features of the figures on the facing page.

◊ Based on the definition of n, why can't there be any such thing as an $n = 0$ state?

◊ Relate the features of the wavefunction plots in figure e to the corresponding features of the probability distribution pictures.

◊ How can you tell from the wavefunction plots in figure e which ones have which angular momenta?

◊ Criticize the following incorrect statement: "The $\ell = 8$ wavefunction in figure c has a shorter wavelength in the center because in the center the electron is in a higher energy level."

◊ Discuss the implications of the fact that the probability cloud in of the $n = 2$, $\ell = 1$ state is split into two parts.

Energies of states in hydrogen

History

The experimental technique for measuring the energy levels of an atom accurately is spectroscopy: the study of the spectrum of light emitted (or absorbed) by the atom. Only photons with certain energies can be emitted or absorbed by a hydrogen atom, for example, since the amount of energy gained or lost by the atom must equal the difference in energy between the atom's initial and final states. Spectroscopy had become a highly developed art several decades before Einstein even proposed the photon, and the Swiss spectroscopist Johann Balmer determined in 1885 that there was a simple equation that gave all the wavelengths emitted by hydrogen. In modern terms, we think of the photon wavelengths merely as indirect evidence about the underlying energy levels of the atom, and we rework Balmer's result into an equation for these atomic energy levels:

$$E_n = -\frac{2.2 \times 10^{-18}\ \text{J}}{n^2},$$

This energy includes both the kinetic energy of the electron and the electrical energy. The zero-level of the electrical energy scale is chosen to be the energy of an electron and a proton that are infinitely far apart. With this choice, negative energies correspond to

bound states and positive energies to unbound ones.

Where does the mysterious numerical factor of 2.2×10^{-18} J come from? In 1913 the Danish theorist Niels Bohr realized that it was exactly numerically equal to a certain combination of fundamental physical constants:

$$E_n = -\frac{mk^2e^4}{2\hbar^2} \cdot \frac{1}{n^2},$$

where m is the mass of the electron, and k is the Coulomb force constant for electric forces.

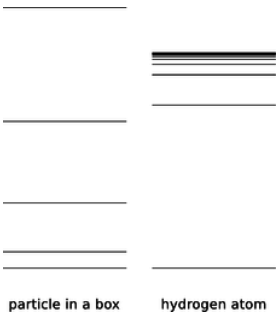
Bohr was able to cook up a derivation of this equation based on the incomplete version of quantum physics that had been developed by that time, but his derivation is today mainly of historical interest. It assumes that the electron follows a circular path, whereas the whole concept of a path for a particle is considered meaningless in our more complete modern version of quantum physics. Although Bohr was able to produce the right equation for the energy levels, his model also gave various wrong results, such as predicting that the atom would be flat, and that the ground state would have $\ell = 1$ rather than the correct $\ell = 0$.

Approximate treatment

Rather than leaping straight into a full mathematical treatment, we'll start by looking for some physical insight, which will lead to an approximate argument that correctly reproduces the form of the Bohr equation.

A typical standing-wave pattern for the electron consists of a central oscillating area surrounded by a region in which the wavefunction tails off. As discussed in subsection 13.3.6, the oscillating type of pattern is typically encountered in the classically allowed region, while the tailing off occurs in the classically forbidden region where the electron has insufficient kinetic energy to penetrate according to classical physics. We use the symbol r for the radius of the spherical boundary between the classically allowed and classically forbidden regions. Classically, r would be the distance from the proton at which the electron would have to stop, turn around, and head back in.

If r had the same value for every standing-wave pattern, then we'd essentially be solving the particle-in-a-box problem in three dimensions, with the box being a spherical cavity. Consider the energy levels of the particle in a box compared to those of the hydrogen atom, f.



f / The energy levels of a particle in a box, contrasted with those of the hydrogen atom.

They're qualitatively different. The energy levels of the particle in a box get farther and farther apart as we go higher in energy, and this feature doesn't even depend on the details of whether the box is two-dimensional or three-dimensional, or its exact shape. The reason for the spreading is that the box is taken to be completely impenetrable, so its size, r, is fixed. A wave pattern with n humps has a wavelength proportional to r/n, and therefore a momentum proportional to n, and an energy proportional to n^2. In the hydrogen atom, however, the force keeping the electron bound isn't an infinite force encountered when it bounces off of a wall, it's the attractive electrical force from the nucleus. If we put more energy into the electron, it's like throwing a ball upward with a higher energy --- it will get farther out before coming back down. This means that in the hydrogen atom, we expect r to increase as we go to states of higher energy. This tends to keep the wavelengths of the high energy states from getting too short, reducing their kinetic energy. The closer and closer crowding of the energy levels in hydrogen also makes sense because we know that there is a certain energy that would be enough to make the electron escape completely, and therefore the sequence of bound states cannot extend above that energy.

When the electron is at the maximum classically allowed distance r from the proton, it has zero kinetic energy. Thus when the electron is at distance r, its energy is purely electrical:

$$E = -\frac{ke^2}{r} \tag{1}$$

Now comes the approximation. In reality, the electron's wavelength cannot be constant in the classically allowed region, but we pretend that it is. Since n is the number of nodes in the wavefunction, we can interpret it approximately as the number of wavelengths that fit across the diameter $2r$. We are not even attempting a derivation that would produce all the correct numerical factors like 2 and π and so on, so we simply make the approximation

$$\lambda \sim \frac{r}{n}. \tag{2}$$

Finally we assume that the typical kinetic energy of the electron is on the same order of magnitude as the absolute value of its total energy. (This is true to within a factor of two for a typical classical system like a planet in a circular orbit around the sun.) We then have

$$\begin{aligned} &\text{absolute value of total energy} \\ &= \frac{ke^2}{r} \\ &\sim K = p^2/2m \\ &= (h/\lambda)^2/2m \\ &\sim h^2n^2/2mr^2 \end{aligned} \tag{3}$$

We now solve the equation $ke^2/r \sim h^2n^2/2mr^2$ for r and throw away numerical factors we can't hope to have gotten right, yielding

$$r \sim \frac{h^2n^2}{mke^2}. \tag{4}$$

Plugging $n = 1$ into this equation gives $r = 2$ nm, which is indeed on the right order of magnitude. Finally we combine equations (4) and (1) to find

$$E \sim -\frac{mk^2e^4}{h^2n^2},$$

which is correct except for the numerical factors we never aimed to find.

Exact treatment of the ground state

The general proof of the Bohr equation for all values of n is beyond the mathematical scope of this book, but it's fairly straightforward to verify it for a particular n, especially given a lucky guess as to what functional form to try for the wavefunction. The form that works for the ground state is

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

where $r = \sqrt{x^2+y^2+z^2}$ is the electron's distance from the proton, and u provides for normalization. In the following, the result $\partial r/\partial x = x/r$ comes in handy. Computing the partial derivatives that occur in the Laplacian, we obtain for the x term

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

The Schrödinger equation gives

$$\begin{aligned} E\cdot\Psi &= -\frac{\hbar^2}{2m}\nabla^2\Psi + U\cdot\Psi \\ &= \frac{\hbar^2}{2m}\left(\frac{2}{ar}-\frac{1}{a^2}\right)\Psi - \frac{ke^2}{r}\cdot\Psi \end{aligned}$$

If we require this equation to hold for all r, then we must have equality for both the terms of the form $(\text{constant})\times\Psi$ and for those of the form $(\text{constant}/r)\times\Psi$. That means

$$E = -\frac{\hbar^2}{2ma^2}$$

and

$$0 = \frac{\hbar^2}{mar} - \frac{ke^2}{r}.$$

These two equations can be solved for the unknowns a and E, giving

$$a = \frac{\hbar^2}{mke^2}$$

and

$$E = -\frac{mk^2e^4}{2\hbar^2},$$

where the result for the energy agrees with the Bohr equation for $n=1$. The calculation of the normalization constant u is relegated to homework problem 36.

Exercise 14.4.1

We've verified that the function $\Psi = he^{-r/a}$ is a solution to the Schrödinger equation, and yet it has a kink in it at $r=0$. What's going on here? Didn't I argue before that kinks are unphysical?

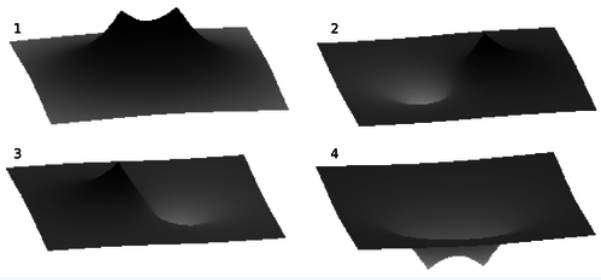
Answer

(answer in the back of the PDF version of the book)

Example 14.4.1: Wave phases in the hydrogen molecule

In example 16 on page 861, I argued that the existence of the H_2 molecule could essentially be explained by a particle-in-a-box argument: the molecule is a bigger box than an individual atom, so each electron's wavelength can be longer, its kinetic energy lower. Now that we're in possession of a mathematical expression for the wavefunction of the hydrogen atom in its ground state, we can make this argument a little more rigorous and detailed. Suppose that two hydrogen atoms are in a relatively cool sample of monoatomic hydrogen gas. Because the gas is cool, we can assume that the atoms are in their ground states. Now suppose that the two atoms approach one another. Making use again of the assumption that the gas is cool, it is reasonable to imagine that the atoms approach one another slowly. Now the atoms come a little closer, but still far enough apart that the region between them is classically forbidden. Each electron can tunnel through this classically forbidden region, but the tunneling probability is small. Each one is now found with, say, 99% probability in its original home, but with 1% probability in the other nucleus. Each electron is now in a state consisting of a superposition of the ground state of its own atom with the ground state of the other atom. There are two peaks in the superposed wavefunction, but one is a much bigger peak than the other.

An interesting question now arises. What are the relative phases of the two electrons? As discussed on page 855, the *absolute* phase of an electron's wavefunction is not really a meaningful concept. Suppose atom A contains electron Alice, and B electron Bob. Just before the collision, Alice may have wondered, “Is my phase positive right now, or is it negative? But of course I shouldn't ask myself such silly questions,” she adds sheepishly.



g / Example 23.

But *relative* phases *are* well defined. As the two atoms draw closer and closer together, the tunneling probability rises, and eventually gets so high that each electron is spending essentially 50% of its time in each atom. It's now reasonable to imagine that either one of two possibilities could obtain. Alice's wavefunction could either look like g/1, with the two peaks in phase with one another, or it could look like g/2, with opposite phases. Because *relative* phases of wavefunctions are well defined, states 1 and 2 are physically distinguishable. In particular, the kinetic energy of state 2 is much higher; roughly speaking, it is like the two-hump wave pattern of the particle in a box, as opposed to 1, which looks roughly like the one-hump pattern with a much longer wavelength. Not only that, but an electron in state 1 has a large probability of being found in the central region, where it has a large negative electrical energy due to its interaction with both protons. State 2, on the other hand, has a low probability of existing in that region. Thus state 1 represents the true ground-state wavefunction of the H_2 molecule, and putting both Alice and Bob in that state results in a lower energy than their total energy when separated, so the molecule is bound, and will not fly apart spontaneously.

State g/3, on the other hand, is not physically distinguishable from g/2, nor is g/4 from g/1. Alice may say to Bob, “Isn't it wonderful that we're in state 1 or 4? I love being stable like this.” But she knows it's not meaningful to ask herself at a given moment which state she's in, 1 or 4.

Solution

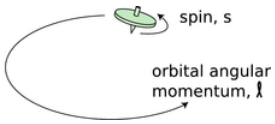
Add text here.

Discussion Questions

- States of hydrogen with n greater than about 10 are never observed in the sun. Why might this be?
- Sketch graphs of r and E versus n for the hydrogen, and compare with analogous graphs for the one-dimensional particle in a box.

Electron spin

It's disconcerting to the novice ping-pong player to encounter for the first time a more skilled player who can put spin on the ball. Even though you can't see that the ball is spinning, you can tell something is going on by the way it interacts with other objects in its environment. In the same way, we can tell from the way electrons interact with other things that they have an intrinsic spin of their own. Experiments show that even when an electron is not moving through space, it still has angular momentum amounting to $\hbar/2$.



h / The top has angular momentum both because of the motion of its center of mass through space and due to its internal rotation. Electron spin is roughly analogous to the intrinsic spin of the top.

This may seem paradoxical because the quantum moat, for instance, gave only angular momenta that were integer multiples of $\hbar$, not half-units, and I claimed that angular momentum was always quantized in units of $\hbar$, not just in the case of the quantum moat. That whole discussion, however, assumed that the angular momentum would come from the motion of a particle through space. The $\hbar/2$ angular momentum of the electron is simply a property of the particle, like its charge or its mass. It has nothing to do with whether the electron is moving or not, and it does not come from any internal motion within the electron. Nobody has ever

succeeded in finding any internal structure inside the electron, and even if there was internal structure, it would be mathematically impossible for it to result in a half-unit of angular momentum.

We simply have to accept this $\hbar/2$ angular momentum, called the "spin" of the electron --- Mother Nature rubs our noses in it as an observed fact.

Protons and neutrons have the same $\hbar/2$ spin, while photons have an intrinsic spin of $\hbar$. In general, half-integer spins are typical of material particles. Integral values are found for the particles that carry forces: photons, which embody the electric and magnetic fields of force, as well as the more exotic messengers of the nuclear and gravitational forces.

As was the case with ordinary angular momentum, we can describe spin angular momentum in terms of its magnitude, and its component along a given axis. We write s and s_z for these quantities, expressed in units of $\hbar$, so an electron has $s = 1/2$ and $s_z = +1/2$ or $-1/2$.

Taking electron spin into account, we need a total of four quantum numbers to label a state of an electron in the hydrogen atom: n, ℓ, ℓ_z, and s_z. (We omit s because it always has the same value.) The symbols and include only the angular momentum the electron has because it is moving through space, not its spin angular momentum. The availability of two possible spin states of the electron leads to a doubling of the numbers of states:

n = 1,	l=0,	l_z=0,	s_z = + 1 / 2 or − 1 / 2	two states
n = 2,	l=0,	l_z=0,	s_z = + 1 / 2 or − 1 / 2	two states
n = 2,	l=1,	l_z=-1, 0, or 1	s_z = + 1 / 2 or − 1 / 2	six states

A note about notation

There are unfortunately two inconsistent systems of notation for the quantum numbers we've been discussing. The notation I've been using is the one that is used in nuclear physics, but there is a different one that is used in atomic physics.

nuclear physics	atomic physics
n	same
l	same
l_x	no notation
l_y	no notation
l_z	m
s = 1 / 2	no notation (sometimesσ)
s_x	no notation
s_y	no notation
s_z	s

he nuclear physics notation is more logical (not giving special status to the z axis) and more memorable (ℓ_z rather than the obscure m), which is why I use it consistently in this book, even though nearly all the applications we'll consider are atomic ones.

We are further encumbered with the following historically derived letter labels, which deserve to be eliminated in favor of the simpler numerical ones:

l=0	l=1	l=2	l=3
s	p	d	f

n = 1	n = 2	n = 3	n = 4	n = 5	n = 6	n = 7
K	L	M	N	O	P	Q

The spdf labels are used in both nuclear[7] and atomic physics, while the KLMNOPQ letters are used only to refer to states of electrons.

And finally, there is a piece of notation that is good and useful, but which I simply haven't mentioned yet. The vector $\mathbf{j} = \mathbf{v}c\ell + \mathbf{s}$ stands for the total angular momentum of a particle in units of $\hbar$, including both orbital and spin parts. This quantum number turns out to be very useful in nuclear physics, because nuclear forces tend to exchange orbital and spin angular momentum, so a given energy level often contains a mixture of ℓ and s values, while remaining fairly pure in terms of j.

13.4.6 Atoms with more than one electron

What about other atoms besides hydrogen? It would seem that things would get much more complex with the addition of a second electron. A hydrogen atom only has one particle that moves around much, since the nucleus is so heavy and nearly immobile. Helium, with two, would be a mess. Instead of a wavefunction whose square tells us the probability of finding a single electron at any given location in space, a helium atom would need to have a wavefunction whose square would tell us the probability of finding two electrons at any given combination of points. Ouch! In addition, we would have the extra complication of the electrical interaction between the two electrons, rather than being able to imagine everything in terms of an electron moving in a static field of force created by the nucleus alone.

Despite all this, it turns out that we can get a surprisingly good description of many-electron atoms simply by assuming the electrons can occupy the same standing-wave patterns that exist in a hydrogen atom. The ground state of helium, for example, would have both electrons in states that are very similar to the $n = 1$ states of hydrogen. The second-lowest-energy state of helium would have one electron in an $n = 1$ state, and the other in an $n = 2$ states. The relatively complex spectra of elements heavier than hydrogen can be understood as arising from the great number of possible combinations of states for the electrons.

A surprising thing happens, however, with lithium, the three-electron atom. We would expect the ground state of this atom to be one in which all three electrons settle down into $n = 1$ states. What really happens is that two electrons go into $n = 1$ states, but the third stays up in an $n = 2$ state. This is a consequence of a new principle of physics:

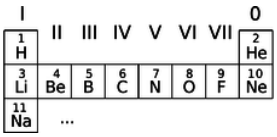
The Pauli Exclusion Principle

Only one electron can ever occupy a given state.

There are two $n = 1$ states, one with $s_z = +1/2$ and one with $s_z = -1/2$, but there is no third $n = 1$ state for lithium's third electron to occupy, so it is forced to go into an $n = 2$ state.

It can be proved mathematically that the Pauli exclusion principle applies to any type of particle that has half-integer spin. Thus two neutrons can never occupy the same state, and likewise for two protons. Photons, however, are immune to the exclusion principle because their spin is an integer.

Deriving the periodic table



i / The beginning of the periodic table.

We can now account for the structure of the periodic table, which seemed so mysterious even to its inventor Mendeleev. The first row consists of atoms with electrons only in the $n = 1$ states:

H	1 electron in an *n* = 1 state
He	2 electrons in the two *n* = 1 states

The next row is built by filling the $n = 2$ energy levels:

Li	2 electrons in *n* = 1 states, 1 electron in an *n* = 2 state
Be	2 electrons in *n* = 1 states, 2 electrons in*n* = 2 states

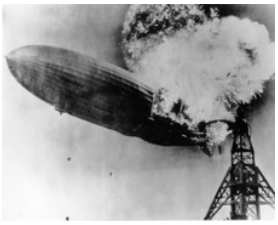
…	
O	2 electrons in n = 1 states, 6 electrons in n = 2 states
F	2 electrons in n = 1 states, 7 electrons in n = 2 states
Ne	2 electrons in n = 1 states, 8 electrons in n = 2 states

In the third row we start in on the $n = 3$ levels:

Na	2 electrons in n = 1 states, 8 electrons in n = 2 states, 1 electron in an n = 3 state

We can now see a logical link between the filling of the energy levels and the structure of the periodic table. Column 0, for example, consists of atoms with the right number of electrons to fill all the available states up to a certain value of n. Column I contains atoms like lithium that have just one electron more than that.

This shows that the columns relate to the filling of energy levels, but why does that have anything to do with chemistry? Why, for example, are the elements in columns I and VII dangerously reactive?



j / Hydrogen is highly reactive.

Consider, for example, the element sodium (Na), which is so reactive that it may burst into flames when exposed to air. The electron in the $n = 3$ state has an unusually high energy. If we let a sodium atom come in contact with an oxygen atom, energy can be released by transferring the $n = 3$ electron from the sodium to one of the vacant lower-energy $n = 2$ states in the oxygen. This energy is transformed into heat. Any atom in column I is highly reactive for the same reason: it can release energy by giving away the electron that has an unusually high energy.

Column VII is spectacularly reactive for the opposite reason: these atoms have a single vacancy in a low-energy state, so energy is released when these atoms steal an electron from another atom.

It might seem as though these arguments would only explain reactions of atoms that are in different rows of the periodic table, because only in these reactions can a transferred electron move from a higher-n state to a lower-n state. This is incorrect. An $n = 2$ electron in fluorine (F), for example, would have a different energy than an $n = 2$ electron in lithium (Li), due to the different number of protons and electrons with which it is interacting. Roughly speaking, the $n = 2$ electron in fluorine is more tightly bound (lower in energy) because of the larger number of protons attracting it. The effect of the increased number of attracting protons is only partly counteracted by the increase in the number of repelling electrons, because the forces exerted on an electron by the other electrons are in many different directions and cancel out partially.

Contributors and Attributions

Benjamin Crowell (Fullerton College).

14.5: Footnotes

1. This is under the assumption that all the uranium atoms were created at the same time. In reality, we have only a general idea of the processes that might have created the heavy elements in the nebula from which our solar system condensed. Some portion of them may have come from nuclear reactions in supernova explosions in that particular nebula, but some may have come from previous supernova explosions throughout our galaxy, or from exotic events like collisions of white dwarf stars.
2. What I'm presenting in this chapter is a simplified explanation of how the photon could have been discovered. The actual history is more complex. Max Planck (1858-1947) began the photon saga with a theoretical investigation of the spectrum of light emitted by a hot, glowing object. He introduced quantization of the energy of light waves, in multiples of hf, purely as a mathematical trick that happened to produce the right results. Planck did not believe that his procedure could have any physical significance. In his 1905 paper Einstein took Planck's quantization as a description of reality, and applied it to various theoretical and experimental puzzles, including the photoelectric effect. Millikan then subjected Einstein's ideas to a series of rigorous experimental tests. Although his results matched Einstein's predictions perfectly, Millikan was skeptical about photons, and his papers conspicuously omit any reference to them. Only in his autobiography did Millikan rewrite history and claim that he had given experimental proof for photons.
3. But note that along the way, we had to make two crucial assumptions: that the wave was sinusoidal, and that it was a plane wave. These assumptions will not prevent us from describing examples such as double-slit diffraction, in which the wave is approximately sinusoidal within some sufficiently small region such as one pixel of a camera's imaging chip. Nevertheless, these issues turn out to be symptoms of deeper problems, beyond the scope of this book, involving the way in which relativity and quantum mechanics should be combined. As a taste of the ideas involved, consider what happens when a photon is reflected from a conducting surface, as in example 23 on p. 699, so that the electric field at the surface is zero, but the magnetic field isn't. The superposition is a standing wave, not a plane wave, so $|\mathbf{E}| = c|\mathbf{B}|$ need not hold, and doesn't. A detector's probability of detecting a photon near the surface could be zero if the detector sensed electric fields, but nonzero if it sensed magnetism. It doesn't make sense to say that either of these is the probability that the photon “was really there.”
4. This interpretation of quantum mechanics is called the Copenhagen interpretation, because it was originally developed by a school of physicists centered in Copenhagen and led by Niels Bohr.
5. This interpretation, known as the many-worlds interpretation, was developed by Hugh Everett in 1957.
6. See page 889 for a note about the two different systems of notations that are used for quantum numbers.
7. After f, the series continues in alphabetical order. In nuclei that are spinning rapidly enough that they are almost breaking apart, individual protons and neutrons can be stirred up to ℓ values as high as 7, which is j.
8. See Barnes et al., “The XYZs of Charmonium at BES,” arxiv.org/abs/hep-ph/0608103. To avoid complication, the levels shown are only those in the group known for historical reasons as the Ψ and J/Ψ.

14.6: Problems

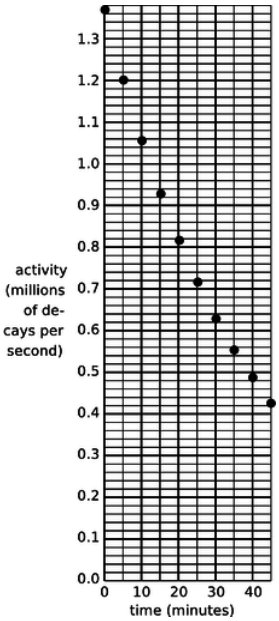
1. If a radioactive substance has a half-life of one year, does this mean that it will be completely decayed after two years? Explain.

2. What is the probability of rolling a pair of dice and getting “snake eyes,” i.e., both dice come up with ones?

3. Problem 3 has been deleted.

4. Problem 4 has been deleted.

5. Refer to the probability distribution for people's heights in figure f on page 828.
(a) Show that the graph is properly normalized.
(b) Estimate the fraction of the population having heights between 140 and 150 cm.(answer check available at lightandmatter.com)



a / Problem 6.

6. (a) A nuclear physicist is studying a nuclear reaction caused in an accelerator experiment, with a beam of ions from the accelerator striking a thin metal foil and causing nuclear reactions when a nucleus from one of the beam ions happens to hit one of the nuclei in the target. After the experiment has been running for a few hours, a few billion radioactive atoms have been produced, embedded in the target. She does not know what nuclei are being produced, but she suspects they are an isotope of some heavy element such as Pb, Bi, Fr or U. Following one such experiment, she takes the target foil out of the accelerator, sticks it in front of a detector, measures the activity every 5 min, and makes a graph (figure). The isotopes she thinks may have been produced are:

isotope	half-life (minutes)
^{211}Pb	36.1
^{214}Pb	26.8
^{214}Bi	19.7
^{223}Fr	21.8
^{239}U	23.5

Which one is it?
(b) Having decided that the original experimental conditions produced one specific isotope, she now tries using beams of ions traveling at several different speeds, which may cause different reactions. The following table gives the activity of the target 10, 20 and 30 minutes after the end of the experiment, for three different ion speeds.

	activity (millions of decays/s) after…		
	10 min	20 min	30 min
first ion speed	1.933	0.832	0.382
second ion speed	1.200	0.545	0.248
third ion speed	7.211	1.296	0.248

Since such a large number of decays is being counted, assume that the data are only inaccurate due to rounding off when writing down the table. Which are consistent with the production of a single isotope, and which imply that more than one isotope was being created?

7. Devise a method for testing experimentally the hypothesis that a gambler's chance of winning at craps is independent of her previous record of wins and losses. If you don't invoke the definition of statistical independence, then you haven't proposed a test.

8. A blindfolded person fires a gun at a circular target of radius b, and is allowed to continue firing until a shot actually hits it. Any part of the target is equally likely to get hit. We measure the random distance r from the center of the circle to where the bullet went in.
(a) Show that the probability distribution of r must be of the form $D(r) = kr$, where k is some constant. (Of course we have $D(r) = 0$ for $r > b$.)
(b) Determine k by requiring D to be properly normalized.(answer check available at lightandmatter.com)

(c) Find the average value of r.(answer check available at lightandmatter.com)

(d) Interpreting your result from part c, how does it compare with $b/2$? Does this make sense? Explain.

9. We are given some atoms of a certain radioactive isotope, with half-life $t_{1/2}$. We pick one atom at random, and observe it for one half-life, starting at time zero. If it decays during that one-half-life period, we record the time t at which the decay occurred. If it doesn't, we reset our clock to zero and keep trying until we get an atom that cooperates. The final result is a time $0 \le t \le t_{1/2}$, with a distribution that looks like the usual exponential decay curve, but with its tail chopped off.
(a) Find the distribution $D(t)$, with the proper normalization.(answer check available at lightandmatter.com)

(b) Find the average value of t.(answer check available at lightandmatter.com)

(c) Interpreting your result from part b, how does it compare with $t_{1/2}/2$? Does this make sense? Explain.

10. The speed, v, of an atom in an ideal gas has a probability distribution of the form $D(v) = bve^{-cv^2}$, where $0 \le v < \infty$, c relates to the temperature, and b is determined by normalization.
(a) Sketch the distribution.
(b) Find b in terms of c.(answer check available at lightandmatter.com)

(c) Find the average speed in terms of c, eliminating b. (Don't try to do the indefinite integral, because it can't be done in closed form. The relevant definite integral can be found in tables or done with computer software.)(answer check available at lightandmatter.com)

11. All helium on earth is from the decay of naturally occurring heavy radioactive elements such as uranium. Each alpha particle that is emitted ends up claiming two electrons, which makes it a helium atom. If the original $^{238}\mathrm{U}$ atom is in solid rock (as opposed to the earth's molten regions), the He atoms are unable to diffuse out of the rock. This problem involves dating a rock using the known decay properties of uranium 238. Suppose a geologist finds a sample of hardened lava, melts it in a furnace, and finds that it contains 1230 mg of uranium and 2.3 mg of helium. $^{238}\mathrm{U}$ decays by alpha emission, with a half-life of 4.5×10^9 years. The subsequent chain of alpha and electron (beta) decays involves much shorter half-lives, and terminates in the stable nucleus $^{206}\mathrm{Pb}$. Almost all natural uranium is $^{238}\mathrm{U}$, and the chemical composition of this rock indicates that there were no decay chains involved other than that of $^{238}\mathrm{U}$.
(a) How many alphas are emitted per decay chain? [Hint: Use conservation of mass.]

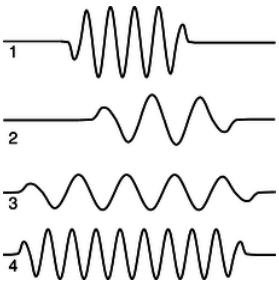
(b) How many electrons are emitted per decay chain? [Hint: Use conservation of charge.]
(c) How long has it been since the lava originally hardened?(answer check available at lightandmatter.com)

12. When light is reflected from a mirror, perhaps only 80% of the energy comes back. One could try to explain this in two different ways: (1) 80% of the photons are reflected, or (2) all the photons are reflected, but each loses 20% of its energy. Based on your everyday knowledge about mirrors, how can you tell which interpretation is correct? [Based on a problem from PSSC Physics.]

13. Suppose we want to build an electronic light sensor using an apparatus like the one described in the section on the photoelectric effect. How would its ability to detect different parts of the spectrum depend on the type of metal used in the capacitor plates?

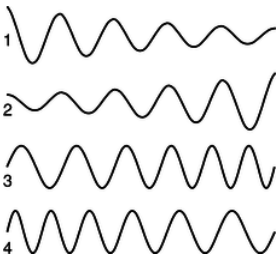
14. The photoelectric effect can occur not just for metal cathodes but for any substance, including living tissue. Ionization of DNA molecules can cause cancer or birth defects. If the energy required to ionize DNA is on the same order of magnitude as the energy required to produce the photoelectric effect in a metal, which of the following types of electromagnetic waves might pose such a hazard? Explain.

- 60 Hz waves from power lines
- 100 MHz FM radio
- microwaves from a microwave oven
- visible light
- ultraviolet light
- x-rays



b / Problem 15.

15. (a) Rank-order the photons according to their wavelengths, frequencies, and energies. If two are equal, say so. Explain all your answers.
(b) Photon 3 was emitted by a xenon atom going from its second-lowest-energy state to its lowest-energy state. Which of photons 1, 2, and 4 are capable of exciting a xenon atom from its lowest-energy state to its second-lowest-energy state? Explain.



c / Problem 16.

16. Which figure could be an electron speeding up as it moves to the right? Explain.

17. The beam of a 100-W overhead projector covers an area of $1\ \text{m} \times 1\ \text{m}$ when it hits the screen 3 m away. Estimate the number of photons that are in flight at any given time. (Since this is only an estimate, we can ignore the fact that the beam is not parallel.) (answer check available at lightandmatter.com)

18. In the photoelectric effect, electrons are observed with virtually no time delay (~ 10 ns), even when the light source is very weak. (A weak light source does however only produce a small number of ejected electrons.) The purpose of this problem is to show that the lack of a significant time delay contradicted the classical wave theory of light, so throughout this problem you should put yourself in the shoes of a classical physicist and pretend you don't know about photons at all. At that time, it was thought that the electron might have a radius on the order of 10^{-15} m. (Recent experiments have shown that if the electron has any finite size at all, it is far smaller.)
(a) Estimate the power that would be soaked up by a single electron in a beam of light with an intensity of 1 $\mathrm{mW/m^2}$.(answer check available at lightandmatter.com)
(b) The energy, E_s, required for the electron to escape through the surface of the cathode is on the order of 10^{-19} J. Find how long it would take the electron to absorb this amount of energy, and explain why your result constitutes strong evidence that there is something wrong with the classical theory.(answer check available at lightandmatter.com)

19. In a television, suppose the electrons are accelerated from rest through a voltage difference of 10^4 V. What is their final wavelength?(answer check available at lightandmatter.com)

20. Use the Heisenberg uncertainty principle to estimate the minimum velocity of a proton or neutron in a $^{208}\mathrm{Pb}$ nucleus, which has a diameter of about 13 fm (1 fm=10^{-15} m). Assume that the speed is nonrelativistic, and then check at the end whether this assumption was warranted.(answer check available at lightandmatter.com)

21. Find the energy of a particle in a one-dimensional box of length L, expressing your result in terms of L, the particle's mass m, the number of peaks and valleys n in the wavefunction, and fundamental constants.(answer check available at lightandmatter.com)

22. A free electron that contributes to the current in an ohmic material typically has a speed of 10^5 m/s (much greater than the drift velocity).
(a) Estimate its de Broglie wavelength, in nm.(answer check available at lightandmatter.com)
(b) If a computer memory chip contains 10^8 electric circuits in a 1 $\mathrm{cm^2}$ area, estimate the linear size, in nm, of one such circuit. (answer check available at lightandmatter.com)
(c) Based on your answers from parts a and b, does an electrical engineer designing such a chip need to worry about wave effects such as diffraction?
(d) Estimate the maximum number of electric circuits that can fit on a 1 $\mathrm{cm^2}$ computer chip before quantum-mechanical effects become important.

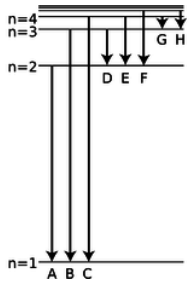
23. In classical mechanics, an interaction energy of the form $U(x) = \frac{1}{2}kx^2$ gives a harmonic oscillator: the particle moves back and forth at a frequency $\omega = \sqrt{k/m}$. This form for $U(x)$ is often a good approximation for an individual atom in a solid, which can vibrate around its equilibrium position at $x = 0$. (For simplicity, we restrict our treatment to one dimension, and we treat the atom as a single particle rather than as a nucleus surrounded by electrons). The atom, however, should be treated quantum-mechanically, not clasically. It will have a wave function. We expect this wave function to have one or more peaks in the classically allowed region, and we expect it to tail off in the classically forbidden regions to the right and left. Since the shape of $U(x)$ is a parabola, not a series of flat steps as in figure m on page 869, the wavy part in the middle will not be a sine wave, and the tails will not be exponentials.
(a) Show that there is a solution to the Schrödinger equation of the form

$$\Psi(x) = e^{-bx^2},$$

and relate b to k, m, and $\hbar$. To do this, calculate the second derivative, plug the result into the Schrödinger equation, and then find what value of b would make the equation valid for *all* values of x. This wavefunction turns out to be the ground state. Note that this wavefunction is not properly normalized --- don't worry about that.(answer check available at lightandmatter.com)
(b) Sketch a graph showing what this wavefunction looks like.
(c) Let's interpret b. If you changed b, how would the wavefunction look different? Demonstrate by sketching two graphs, one for a smaller value of b, and one for a larger value.
(d) Making k greater means making the atom more tightly bound. Mathematically, what happens to the value of b in your result from part a if you make k greater? Does this make sense physically when you compare with part c?

24. (a) A distance scale is shown below the wavefunctions and probability densities illustrated in figure e on page 882. Compare this with the order-of-magnitude estimate derived in subsection 13.4.4 for the radius r at which the wavefunction begins tailing off. Was the estimate on the right order of magnitude?
(b) Although we normally say the moon orbits the earth, actually they both orbit around their common center of mass, which is

below the earth's surface but not at its center. The same is true of the hydrogen atom. Does the center of mass lie inside the proton, or outside it?



d / Problem 25.

25. The figure shows eight of the possible ways in which an electron in a hydrogen atom could drop from a higher energy state to a state of lower energy, releasing the difference in energy as a photon. Of these eight transitions, only D, E, and F produce photons with wavelengths in the visible spectrum.
(a) Which of the visible transitions would be closest to the violet end of the spectrum, and which would be closest to the red end? Explain.
(b) In what part of the electromagnetic spectrum would the photons from transitions A, B, and C lie? What about G and H? Explain.
(c) Is there an upper limit to the wavelengths that could be emitted by a hydrogen atom going from one bound state to another bound state? Is there a lower limit? Explain.

26. Find an equation for the wavelength of the photon emitted when the electron in a hydrogen atom makes a transition from energy level n_1 to level n_2.(answer check available at lightandmatter.com)

27. Estimate the angular momentum of a spinning basketball, in units of $\hbar$. Explain how this result relates to the correspondence principle.

28. Assume that the kinetic energy of an electron the $n = 1$ state of a hydrogen atom is on the same order of magnitude as the absolute value of its total energy, and estimate a typical speed at which it would be moving. (It cannot really have a single, definite speed, because its kinetic and interaction energy trade off at different distances from the proton, but this is just a rough estimate of a typical speed.) Based on this speed, were we justified in assuming that the electron could be described nonrelativistically?

29. Before the quantum theory, experimentalists noted that in many cases, they would find three lines in the spectrum of the same atom that satisfied the following mysterious rule: $1/\lambda_1 = 1/\lambda_2 + 1/\lambda_3$. Explain why this would occur. Do not use reasoning that only works for hydrogen --- such combinations occur in the spectra of all elements. [Hint: Restate the equation in terms of the energies of photons.]

30. The wavefunction of the electron in the ground state of a hydrogen atom is

$$\Psi = \pi^{-1/2} a^{-3/2} e^{-r/a},$$

where r is the distance from the proton, and $a = 5.3 \times 10^{-11}$ m is a constant that sets the size of the wave.
(a) Calculate symbolically, without plugging in numbers, the probability that at any moment, the electron is inside the proton. Assume the proton is a sphere with a radius of $b = 0.5$ fm. [Hint: Does it matter if you plug in $r = 0$ or $r = b$ in the equation for the wavefunction?](answer check available at lightandmatter.com)
(b) Calculate the probability numerically.(answer check available at lightandmatter.com)
(c) Based on the equation for the wavefunction, is it valid to think of a hydrogen atom as having a finite size? Can a be interpreted as the size of the atom, beyond which there is nothing? Or is there any limit on how far the electron can be from the proton?

31. Use physical reasoning to explain how the equation for the energy levels of hydrogen,

$$E_n = -\frac{mk^2e^4}{2\hbar^2} \cdot \frac{1}{n^2},$$

should be generalized to the case of an atom with atomic number Z that has had all its electrons removed except for one.

32. A muon is a subatomic particle that acts exactly like an electron except that its mass is 207 times greater. Muons can be created by cosmic rays, and it can happen that one of an atom's electrons is displaced by a muon, forming a muonic atom. If this happens to a hydrogen atom, the resulting system consists simply of a proton plus a muon.
(a) Based on the results of section 13.4.4, how would the size of a muonic hydrogen atom in its ground state compare with the size of the normal atom?
(b) If you were searching for muonic atoms in the sun or in the earth's atmosphere by spectroscopy, in what part of the electromagnetic spectrum would you expect to find the absorption lines?

33. A photon collides with an electron and rebounds from the collision at 180 degrees, i.e., going back along the path on which it came. The rebounding photon has a different energy, and therefore a different frequency and wavelength. Show that, based on conservation of energy and momentum, the difference between the photon's initial and final wavelengths must be $2h/mc$, where m is the mass of the electron. The experimental verification of this type of "pool-ball" behavior by Arthur Compton in 1923 was taken as definitive proof of the particle nature of light. Note that we're not making any nonrelativistic approximations. To keep the algebra simple, you should use natural units --- in fact, it's a good idea to use even-more-natural-than-natural units, in which we have not just $c = 1$ but also $h = 1$, and $m = 1$ for the mass of the electron. You'll also probably want to use the relativistic relationship $E^2 - p^2 = m^2$, which becomes $E^2 - p^2 = 1$ for the energy and momentum of the electron in these units.

34. Generalize the result of problem 33 to the case where the photon bounces off at an angle other than 180° with respect to its initial direction of motion.

35. On page 869 we derived an expression for the probability that a particle would tunnel through a rectangular barrier, i.e., a region in which the interaction energy $U(x)$ has a graph that looks like a rectangle. Generalize this to a barrier of any shape. [Hints: First try generalizing to two rectangular barriers in a row, and then use a series of rectangular barriers to approximate the actual curve of an arbitrary function $U(x)$. Note that the width and height of the barrier in the original equation occur in such a way that all that matters is the area under the U-versus-x curve. Show that this is still true for a series of rectangular barriers, and generalize using an integral.] If you had done this calculation in the 1930's you could have become a famous physicist.

36. Show that the wavefunction given in problem 30 is properly normalized.

37. Show that a wavefunction of the form $\Psi = e^{by} \sin ax$ is a possible solution of the Schrödinger equation in two dimensions, with a constant potential. Can we tell whether it would apply to a classically allowed region, or a classically forbidden one?

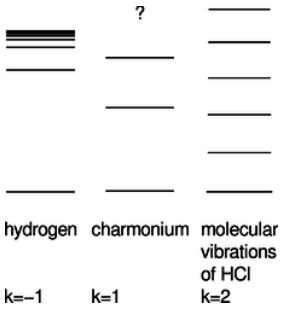
38. Find the energy levels of a particle in a three-dimensional rectangular box with sides of length a, b, and c.(answer check available at lightandmatter.com)

39. Americium-241 is an artificial isotope used in smoke detectors. It undergoes alpha decay, with a half-life of 432 years. As discussed in example 18 on page 870, alpha decay can be understood as a tunneling process, and although the barrier is not rectangular in shape, the equation for the tunneling probability on page 870 can still be used as a rough guide to our thinking. For americium-241, the tunneling probability is about 1×10^{-29}. Suppose that this nucleus were to decay by emitting a tritium (helium-3) nucleus instead of an alpha particle (helium-4). Estimate the relevant tunneling probability, assuming that the total energy E remains the same. This higher probability is contrary to the empirical observation that this nucleus is not observed to decay by tritium emission with any significant probability, and in general tritium emission is almost unknown in nature; this is mainly because the tritium nucleus is far less stable than the helium-4 nucleus, and the difference in binding energy reduces the energy available for the decay.

40. As far as we know, the mass of the photon is zero. However, it's not possible to prove by experiments that anything is zero; all we can do is put an upper limit on the number. As of 2008, the best experimental upper limit on the mass of the photon is about 1×10^{-52} kg. Suppose that the photon's mass really isn't zero, and that the value is at the top of the range that is consistent with the present experimental evidence. In this case, the c occurring in relativity would no longer be interpreted as the speed of light. As with material particles, the speed v of a photon would depend on its energy, and could never be as great as c. Estimate the relative size $(c - v)/c$ of the discrepancy in speed, in the case of a photon with a frequency of 1 kHz, lying in the very low frequency radio range. \hwans{hwans:photon-mass}

41. Hydrogen is the only element whose energy levels can be expressed exactly in an equation. Calculate the ratio λ_E/λ_F of the wavelengths of the transitions labeled E and F in problem 25 on p. 898. Express your answer as an exact fraction, not a decimal approximation. In an experiment in which atomic wavelengths are being measured, this ratio provides a natural, stringent check on the precision of the results.(answer check available at lightandmatter.com)

42. Give a numerical comparison of the number of photons per second emitted by a hundred-watt FM radio transmitter and a hundred-watt lightbulb.(answer check available at lightandmatter.com)



e / Problem 43.

43. On pp. 884-885 of subsection 13.4.4, we used simple algebra to derive an approximate expression for the energies of states in hydrogen, without having to explicitly solve the Schrödinger equation. As input to the calculation, we used the the proportionality $U \propto r^{-1}$, which is a characteristic of the electrical interaction. The result for the energy of the nth standing wave pattern was $E_n \propto n^{-2}$.

There are other systems of physical interest in which we have $U \propto r^k$ for values of k besides -1. Problem 23 discusses the ground state of the harmonic oscillator, with $k = 2$ (and a positive constant of proportionality). In particle physics, systems called charmonium and bottomonium are made out of pairs of subatomic particles called quarks, which interact according to $k = 1$, i.e., a force that is independent of distance. (Here we have a positive constant of proportionality, and $r > 0$ by definition. The motion turns out not to be too relativistic, so the Schrödinger equation is a reasonable approximation.) The figure shows actual energy levels for these three systems, drawn with different energy scales so that they can all be shown side by side. The sequence of energies in hydrogen approaches a limit, which is the energy required to ionize the atom. In charmonium, only the first three levels are known.[8]

Generalize the method used for $k = -1$ to any value of k, and find the exponent j in the resulting proportionality $E_n \propto n^j$. Compare the theoretical calculation with the behavior of the actual energies shown in the figure. Comment on the limit $k \to \infty$. (answer check available at lightandmatter.com)

44. The electron, proton, and neutron were discovered, respectively, in 1897, 1919, and 1932. The neutron was late to the party, and some physicists felt that it was unnecessary to consider it as fundamental. Maybe it could be explained as simply a proton with an electron trapped inside it. The charges would cancel out, giving the composite particle the correct neutral charge, and the masses at least approximately made sense (a neutron is heavier than a proton). (a) Given that the diameter of a proton is on the order of 10^{-15} m, use the Heisenberg uncertainty principle to estimate the trapped electron's minimum momentum.(answer check available at lightandmatter.com)
(b) Find the electron's minimum kinetic energy.(answer check available at lightandmatter.com)
(c) Show via $E = mc^2$ that the proposed explanation fails, because the contribution to the neutron's mass from the electron's kinetic energy would be many orders of magnitude too large.

45. Suppose that an electron, in one dimension, is confined to a certain region of space so that its wavefunction is given by

$$\Psi = \begin{cases} 0 & \text{if } x < 0 \\ A\sin(2\pi x/L) & \text{if } 0 \le x \le L \\ 0 & \text{if } x > L \end{cases}$$

Determine the constant A from normalization.(answer check available at lightandmatter.com)

46. In the following, x and y are variables, while u and v are constants. Compute (a) $\partial(ux\ln(vy))/\partial x$, (b) $\partial(ux\ln(vy))/\partial y$. (answer check available at lightandmatter.com)

47. (a) A radio transmitter radiates power P in all directions, so that the energy spreads out spherically. Find the energy density at a distance r.(answer check available at lightandmatter.com)
(b) Let the wavelength be λ. As described in example 8 on p. 842, find the number of photons in a volume λ^3 at this distance r. (answer check available at lightandmatter.com)

(c) For a 1000 kHz AM radio transmitting station, assuming reasonable values of P and r, verify, as claimed in the example, that the result from part b is very large.

Contributors and Attributions

Benjamin Crowell (Fullerton College). Conceptual Physics is copyrighted with a CC-BY-SA license.

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Conceptual Physics (Crowell)

Webpages: 112

All licenses found:

- CC BY-SA 4.0: 87.5% (98 pages)
- Undeclared: 12.5% (14 pages)

By Page