Simple Nature

An Introduction to Physics for Engineering and Physical Science Students

Benjamin Crowell

Simple Nature

An Introduction to Physics for Engineering and Physical Science Students

Benjamin Crowell

www.lightandmatter.com

Light and Matter
Fullerton, California
www.lightandmatter.com

rev. June 6, 2007

ISBN 0-9704670-7-9

Brief Contents

Contents

11 Electromagnetism

12 Quantum Physics

Chapter 1
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*

The Republic has no need of scientists.

> *Judge Pierre-André Coffinhal's reply to Lavoisier's request for a fifteen-day delay of his execution, so that he could complete some experiments that might be of value to the Republic. Coffinhal was himself executed August 6, 1794. As a scientific experiment, Lavoisier decided to try to determine how long his consciousness would continue after he was guillotined, by blinking his eyes for as long as possible. He blinked twelve times after his head was chopped off.*

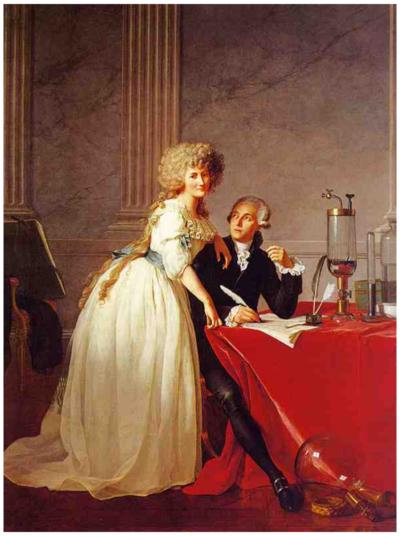
1.1 Mass

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.[1] Numerical measurement of distance and time

[1] If you haven't already, you should now go ahead and memorize the common metric prefixes, which are summarized on page 719.

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.



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.

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.

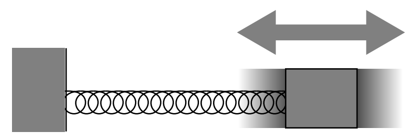
There are, however, at least two approaches to defining mass that lead to its being a conserved quantity, so we consider these definitions to be "right" in the pragmatic sense that what's correct is what's useful.

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*.

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.[2] 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!

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 b and c). They strap themselves to a chair which is attached to a large spring, and measure the time it takes for one cycle of vibration.



b / The time for one cycle of vibration is related to the object's inertial mass.

c / Astronaut Tamara Jernigan measures her inertial mass aboard the Space Shuttle.

[2]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.

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 laser 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.

d / Example 1.

A stream of water — *example 1*

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.

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, p. 705

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.

Lavoisier and chemical reactions in a closed system *example 2*

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!)

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.

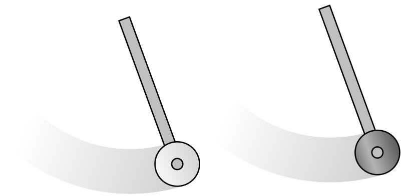
Self-Check
If x represents the location of an object moving in one dimension, then how would positive and negative signs of Δx be interpreted? ▷ Answer, p. 706

Discussion Questions

A 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 $\Delta x \neq 0$?

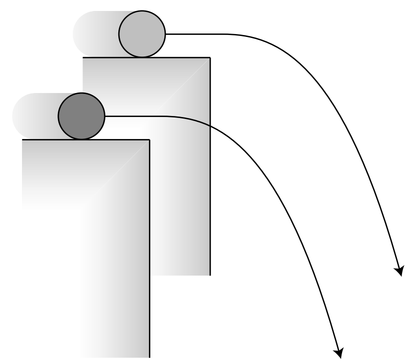
1.2 Equivalence of Gravitational and Inertial Mass

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!



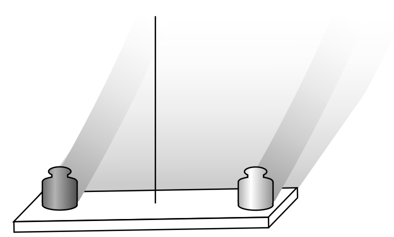
a / The two pendulum bobs are constructed with equal gravitational masses. If their inertial masses are also equal, then each pendulum should take exactly the same amount of time per swing.

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.



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

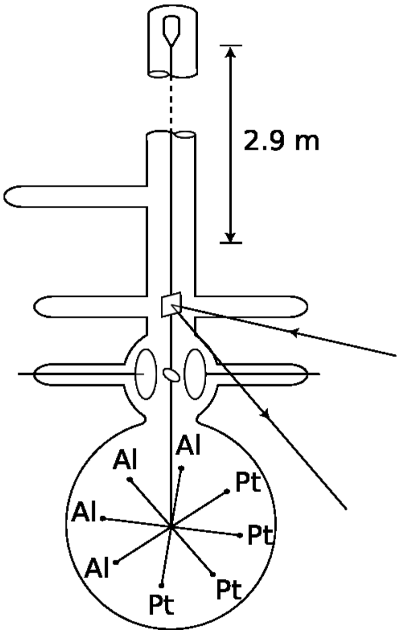
The layperson is surprised, since it would seem logical that heaver objects would always fall faster than light ones. However, it's fairly easy to prove that if air friction is negligible, any two objects made of the same substance must have identical motion when they fall. For instance, a 2-kg copper 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.[3]



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.

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 begs the question of 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.

[3] 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.



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 °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.

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.[4] 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.

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

[4] V.B. Braginskii and V.I. Panov, Soviet Physics JETP 34, 463 (1972).

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.

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 earthshattering. 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.

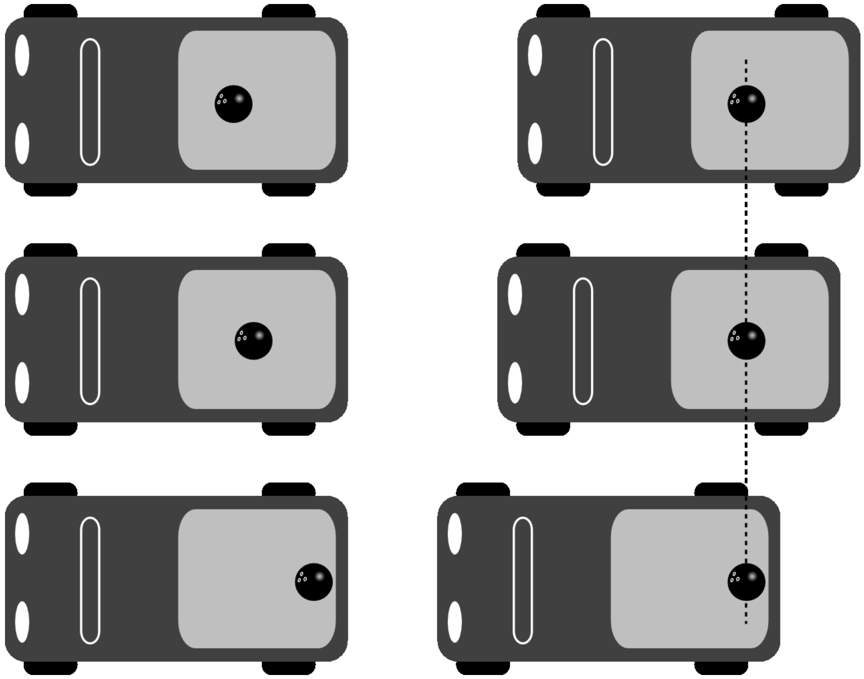
1.3 Galilean Relativity

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!

a / Portrait of Galileo Galilei, by Justus Sustermans, 1636.

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.

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



b / 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.

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 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.

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.[5]

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

[5]The principle of Galilean relativity is extended on page 139.

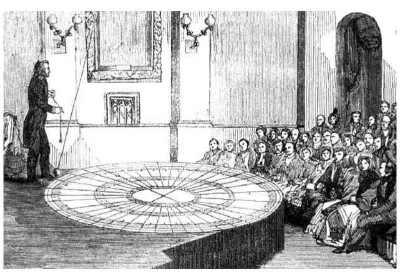
relative to another, but there is no privileged "rest frame." There is no experiment that comes out any different in laboratories in different inertial frames, so there is no experiment that could tell us which inertial frame is really, truly at rest.

The speed of sound *example 3*

▷ 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.

Self-Check

Galileo got in a bet with some rich noblemen about the following experiment. Suppose a ship is sailing across a calm harbor at constant speed in a straight line. A sailor is assigned to carry a rock up to the top of one of the masts and then drop it to the deck. Does the rock land at the base of the mast, or behind it due to the motion of the ship? (Galileo was never able to collect on his bet, because the noblemen didn't think an actual experiment was a valid way of deciding who was right.) ▷ Answer, p. 706



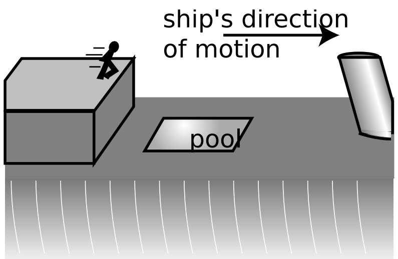
c / Foucault demonstrates his pendulum to an audience at a lecture in 1851.

The Foucault pendulum *example 4*

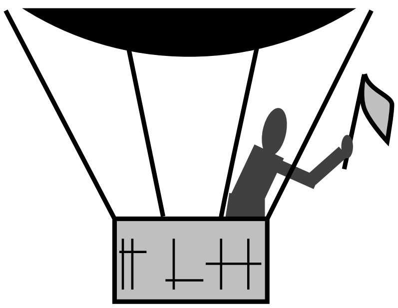
Note that in the example of the bowling ball in the truck, I didn't claim the sidewalk was *exactly* a Galilean frame of reference. This is because the sidewalk is moving in a circle due to the rotation of the Earth, and is therefore changing the direction of its motion continuously on a 24-hour cycle. 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 c) 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.

Although popular belief has Galileo being prosecuted by the

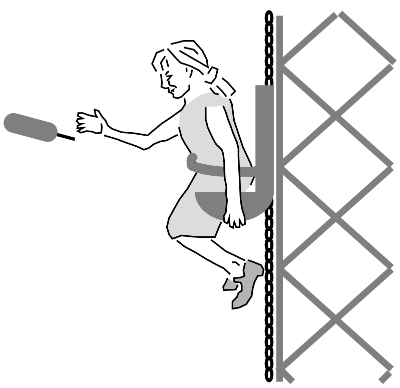
Catholic Church for saying the earth rotated on its axis and also orbited the sun, Foucault's pendulum was still centuries in the future, so Galileo had no hard proof; Galileo's insights into relative versus absolute motion simply made it more plausible that the world could be spinning without producing 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.



d / Discussion question C.



e / Discussion question D.



f / Discussion question E.

Discussion Questions

B Aristotle stated that all objects naturally wanted to come to rest, with the unspoken implication that “rest” would be interpreted relative to the surface of the earth. Suppose we could transport Aristotle to the moon, put him in a space suit, and kick him out the door of the spaceship and into the lunar landscape. What would he expect his fate to be in this situation? If intelligent creatures inhabited the moon, and one of them independently came up with the equivalent of Aristotelian physics, what would they think about objects coming to rest?

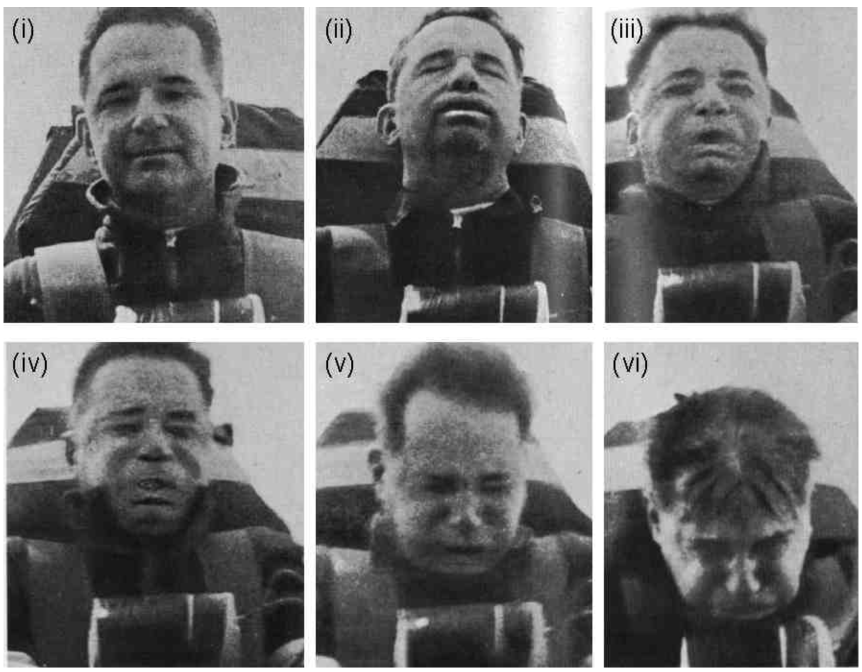
C A passenger on a cruise ship finds, while the ship is docked, that he can leap off of the upper deck and just barely make it into the pool on the lower deck. If the ship leaves dock and is cruising rapidly, will this adrenaline junkie still be able to make it?

D You are a passenger in the open basket hanging under a helium balloon. The balloon is being carried along by the wind at a constant velocity. If you are holding a flag in your hand, will the flag wave? If so, which way? [Based on a question from PSSC Physics.]

E Sally is on an amusement park ride which begins with her chair being hoisted straight up a tower at a constant speed of 60 miles/hour. Despite stern warnings from her father that he’ll take her home the next time she misbehaves, she decides that as a scientific experiment she really needs to release her corndog over the side as she’s on the way up. She does not throw it. She simply sticks it out of the car, lets it go, and watches it against the background of the sky, with no trees or buildings as reference points. What does the corndog’s motion look like as observed by Sally? Does its speed ever appear to her to be zero? What acceleration does she observe it to have: is it ever positive? negative? zero? What would her enraged father answer if asked for a similar description of its motion as it appears to him, standing on the ground?

1.3.1 Applications of calculus

Let's see how this relates to calculus. If an object is moving in one dimension, we can describe its position with a function $x(t)$. The derivative $v = \mathrm{d}x/\mathrm{d}t$ is called the velocity, and the second derivative $a = \mathrm{d}v/\mathrm{d}t = \mathrm{d}^2x/\mathrm{d}t^2$ is the acceleration. Galilean relativity tells us that there is no detectable effect due to an object's absolute velocity, since in some other frame of reference, the object's velocity might be zero. However, an acceleration does have physical consequences.



g / This Air Force doctor volunteered to ride a rocket sled as a medical experiment. The obvious effects on his head and face are not because of the sled's speed but because of its rapid changes in speed: increasing in (ii) and (iii), and decreasing in (v) and (vi). In (iv) his speed is greatest, but because his speed is not increasing or decreasing very much at this moment, there is little effect on him. (*U.S. Air Force*)

Observers in different inertial frames of reference will disagree on velocities, but agree on accelerations. Let's keep it simple by continuing to work in one dimension. One frame of reference uses a coordinate system x_1, and the other we label x_2. If the positive x_1 and x_2 axes point in the same direction, then in general two inertial frames could be related by an equation of the form $x_2 = x_1 + b + ut$, where u is the constant velocity of one frame relative to the other, and the constant b tells us how far apart the origins of the two coordinate systems were at $t = 0$. The velocities are different in the two frames of reference:

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = \frac{\mathrm{d}x_1}{\mathrm{d}t} + u \qquad ,$$

Suppose, for example, frame 1 is defined from the sidewalk, and frame 2 is fixed to a float in a parade that is moving to our left at a velocity $u = 1$ m/s. A dog that is moving to the right with a velocity $v_1 = \mathrm{d}x_1/\mathrm{d}t = 3$ m/s in the sidewalk's frame will appear to be moving at a velocity of $v_2 = \mathrm{d}x_2/\mathrm{d}t = \mathrm{d}x_1/\mathrm{d}t + u = 4$ m/s in the float's frame.

For acceleration, however, we have

$$\frac{\mathrm{d}^2 x_2}{\mathrm{d}t^2} = \frac{\mathrm{d}^2 x_1}{\mathrm{d}t^2} \qquad ,$$

since the derivative of the constant u is zero. Thus an acceleration, unlike a velocity, can have a definite physical significance to all observers in all frames of reference. If this wasn't true, then there would be no particular reason to define a quantity called acceleration in the first place.

Self-Check

Figure h shows a bottle of beer sitting on a table in the dining car of a train. Does the tilting of the surface tell us about the train's velocity, or its acceleration? What would a person in the train say about the bottle's velocity? What about a person standing in a field outside and looking in through the window? What about the acceleration? ▷ Answer, p. 706

h / self-check

1.4 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, p. 706

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 $6x \times 10^{-12}$ 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.

This chapter is summarized on page 722. Notation and terminology are tabulated on pages 718-719.

Problems

The symbols $\surd$, , etc. are explained on page 32.

Problems 1-6 are intended to help you check up on your mathematical skills. In my experience, most students can do most of these problems when they start a physics course, but very few students can do all of them. I've written a complete introduction to these skills in ch. 0 and 1 of my book **Newtonian Physics**, which is available as a free download from www.lightandmatter.com. Rather than duplicating that material in this book, I've decided simply to steer students to it if they need it.

1 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. $\surd$

2 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. $\surd$

3 How many significant figures are there in each of the following measurements? (a) 9.937 m, (b) 4.0 s, (c) 0.0000037 kg.

4 How many cubic mm are there in a cubic meter? The answer is not 1000.

5 Assume that dogs' and cats' brains are the same shape, and that their brain cells are also the same size and shape, but that a dog's brain is twice as large as a cat's in all its linear dimensions, i.e. any two points in a dog's brain are twice as far apart as the corresponding two points in a cat's. How many times more brain cells does a dog have compared to a cat? The answer is not 2.

6 Make an order-of-magnitude estimate of the number of blades of grass on a football field.

7 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?

8 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 is allowed to escape, 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.

9 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:

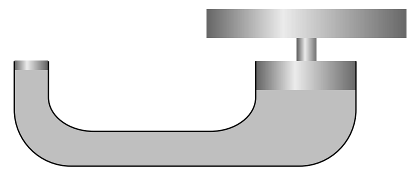
uranium	$3.95292849 \times 10^{-25}$ kg
thorium	$3.88638748 \times 10^{-25}$ kg
helium	6.646481×10^{-27} kg

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?

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

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

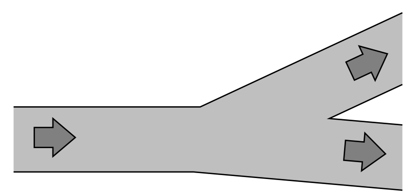
12 The figure shows a hydraulic jack. What is the relationship between the distance traveled by the plunger and the distance traveled by the object being lifted, in terms of the cross-sectional areas?



Problem 12.

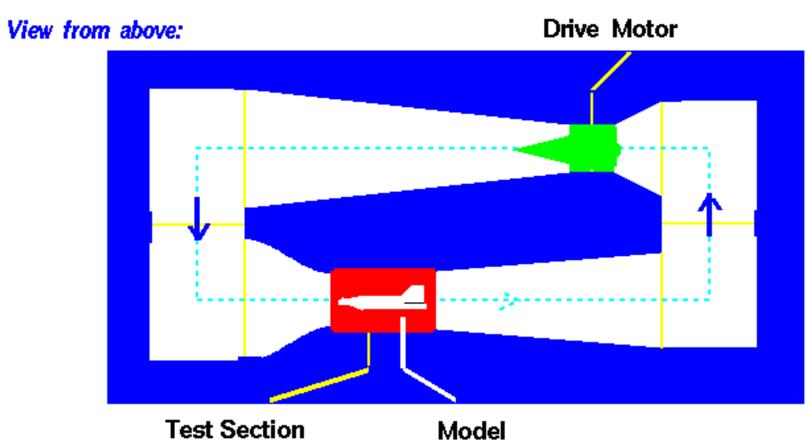
13 In an example in this chapter, I argued that a stream of water must change its cross-sectional area as it rises or falls. Suppose that the stream of water is confined to a constant-diameter pipe. Which assumption breaks down in this situation?

14 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?



Problem 14.

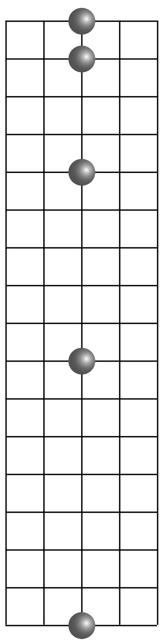
15 The diagram shows a cross-section of a wind tunnel of the kind used, for example, to test designs of airplanes. Under normal conditions of use, the density of the air remains nearly constant throughout the whole wind tunnel. How can the speed of the air be controlled and calculated? (Diagram by NASA,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



16 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

$$V = b \int_0^a y(x) \mathrm{d}x \qquad .$$

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.

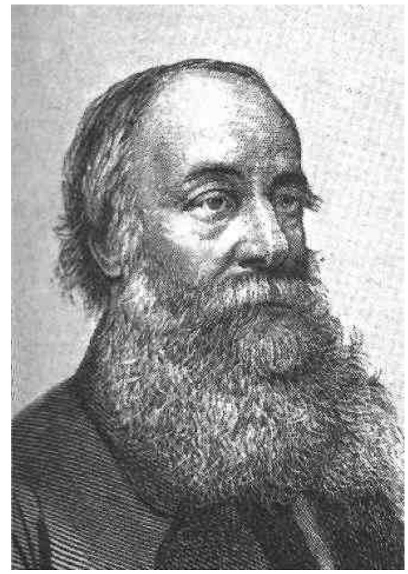


Problem 17.

17 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.

Key to symbols:
easy typical challenging difficult very difficult
√ An answer check is available at www.lightandmatter.com.

Chapter 2
Conservation of Energy



a / James Joule, 1818-1889. The son of a wealthy brewer, Joule was tutored as a young man by the famous scientist John Dalton. Fascinated by electricity, he and his brother experimented by giving electric shocks to each other and to the family's servants. Joule ran the brewery as an adult, and science was merely a serious hobby. His work on energy can be traced to his attempt to build an electric motor that would replace steam engines. His ideas were not accepted at first, partly because they contradicted the widespread belief that heat was a fluid, and partly because they depended on extremely precise measurements, which had not previously been common in physics.

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."*

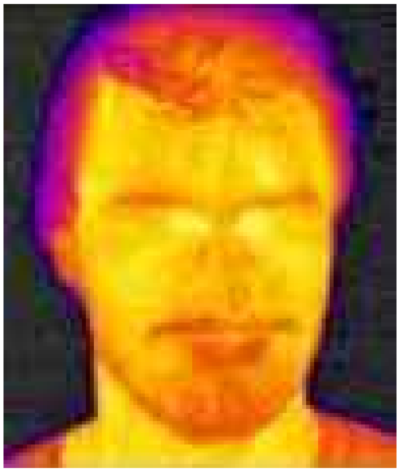
2.1 Energy

2.1.1 The energy concept

You'd probably like to be able to drive your car and light your apartment without having to pay money for gas and electricity, and if you do a little websurfing, you can easily find people who say they have the solution to your problem. This kind of scam has been around for centuries. It used to be known as a perpetual motion machine, but nowadays the con artists' preferred phrase is "free energy."[1] A typical "free-energy" machine would be a sealed box that heats your house without needing to be plugged into a wall socket or a gas pipe. Heat comes out, but nothing goes in, and this can go on indefinitely. But an interesting thing happens if you try to check on the advertised performance of the machine. Typically, you'll find out that either the device is still in development, or it's back-ordered because so many people have already taken advantage of this Fantastic Opportunity! In a few cases, the magic box exists, but the inventor is only willing to demonstrate very small levels of heat output for short periods of time, in which case there's probably a tiny hearing-aid battery hidden in there somewhere, or some other trick.

Since nobody has ever succeeded in building a device that creates heat out of nothing, we might also wonder whether any device exists that can do the opposite, turning heat into nothing. You might think

[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.



b / Heat energy can be converted to light energy. Very hot objects glow visibly, and even objects that aren't so hot give off infrared light, a color of light that lies beyond the red end of the visible rainbow. This photo was made with a special camera that records infrared light. The man's warm skin emits quite a bit of infrared light energy, while his hair, at a lower temperature, emits less.

that a refrigerator was such a device, but actually your refrigerator doesn't destroy the heat in the food. What it really does is to extract some of the heat and bring it out into the room. That's why it has big radiator coils on the back, which get hot when it's in operation.

If it's not possible to destroy or create heat outright, then you might start to suspect that heat was a conserved quantity. This would be a successful rule for explaining certain processes, such as the transfer of heat between a cold Martini and a room-temperature olive: if the olive loses a little heat, then the drink must gain the same amount. It would fail in general, however. Sunlight can heat your skin, for example, and a hot lightbulb filament can cool off by emitting light. Based on these observations, we could revise our proposed conservation law, and say that there is something called heatpluslight, which is conserved. Even this, however, needs to be generalized in order to explain why you can get a painful burn playing baseball when you slide into a base. Now we could call it heatpluslightplusmotion. The word is getting pretty long, and we haven't even finished the list.

Rather than making the word longer and longer, physicists have hijacked the word "energy" from ordinary usage, and give it a new, specific technical meaning. Just as the Parisian platinum-iridium kilogram defines a specific unit of mass, we need to pick something that defines a definite unit of energy. The metric unit of energy is the joule (J), and we'll define it as the amount of energy required to heat 0.24 grams of water from 20 to 21 degrees Celsius. (Don't memorize the numbers.)[2]

Temperature of a mixture *example 1*

▷ If 1.0 kg of water at 20 °C is mixed with 4.0 kg of water at 30 °C, what is the temperature of the mixture?

▷ Let's assume as an approximation that each degree of temperature change corresponds to the same amount of energy. In other words, we assume $\Delta E = mc\Delta T$, regardless of whether, as in the definition of the joule, we have $\Delta T = 21\,^\circ\text{C} - 20\,^\circ\text{C}$ or, as in the present example, some other combination of initial and final temperatures. To be consistent with the definition of the joule, we must have $c = (1\ \text{J})/(0.24\ \text{g})/(1\,^\circ\text{C}) = 4.2\times10^3\ \text{J/kg}\cdot{}^\circ\text{C}$, which is referred to as the specific heat of water.

Conservation of energy tells us $\Delta E = 0$, so

$$m_1 c\Delta T_1 + m_2 c\Delta T_2 = 0$$

[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.

or

$$\begin{aligned}\frac{\Delta T_1}{\Delta T_2} &= -\frac{m_2}{m_1} \\ &= -4.0 \qquad .\end{aligned}$$

If T_1 has to change four times as much as T_2, and the two final temperatures are equal, then the final temperature must be 28 °C.

Note how only *differences* in temperature and energy appeared in the preceding example. In other words, we don't have to make any assumptions about whether there is a temperature at which all an object's heat energy is removed. Historically, the energy and temperature units were invented before it was shown that there is such a temperature, called absolute zero. There is a scale of temperature, the Kelvin scale, in which the unit of temperature is the same as the Celsius degree, but the zero point is defined as absolute zero. But as long as we only deal with temperature differences, it doesn't matter whether we use Kelvin or Celsius. Likewise, as long as we deal with differences in heat energy, we don't normally have to worry about the total amount of heat energy the object has. In standard physics terminology, "heat" is used only to refer to differences, while the total amount is called the object's "thermal energy." This distinction is often ignored by scientists in casual speech, and in this book I'll usually use "heat" for either quantity.

We're defining energy by adding up things from a list, which we lengthen as needed: heat, light, motion, etc. One objection to this approach is aesthetic: physicists tend to regard complication as a synonym for ugliness. If we have to keep on adding more and more forms of energy to our laundry list, then it's starting to sound like energy is distressingly complicated. Luckily it turns out that energy is simpler than it seems. Many forms of energy that are apparently unrelated turn out to be manifestations of a small number of forms at the atomic level, and this is the topic of section 2.4.

Discussion Questions

A The ancient Greek philosopher Aristotle said that objects "naturally" tended to slow down, unless there was something pushing on them to keep them moving. What important insight was he missing?

2.1.2 Logical issues

Another possible objection is that the open-ended approach to defining energy might seem like a kind of cheat, since we keep on inventing new forms whenever we need them. If a certain experiment seems to violate conservation of energy, can't we just invent a new form of invisible "mystery energy" that patches things up? This would be like balancing your checkbook by putting in a fake transaction that makes your calculation of the balance agree with your bank's. If we could fudge this way, then conservation of energy would be untestable — impossible to prove or disprove.

c / As in figure b, an infrared camera distinguishes hot and cold areas. As the bike skids to a stop with its brakes locked, the kinetic energy of the bike and rider is converted into heat in both the floor (top) and the tire (bottom).

Actually all scientific theories are unprovable. A theory can never be proved, because the experiments can only cover a finite number out of the infinitely many situations in which the theory is supposed to apply. Even a million experiments won't suffice to prove it in the same sense of the word "proof" that is used in mathematics. However, even one experiment that contradicts a theory is sufficient to show that the theory is wrong. A theory that is immune to disproof is a bad theory, because there is no way to test it. For instance, if I say that 23 is the maximum number of angels that can dance on the head of a pin, I haven't made a properly falsifiable scientific theory, since there's no method by which anyone could even attempt to prove me wrong based on observations or experiments.

Conservation of energy is testable because new forms of energy are expected to show regular mathematical behavior, and are supposed to be related in a measurable way to observable phenomena. As an example, let's see how to extend the energy concept to include motion.

2.1.3 Kinetic energy

Energy of motion is called *kinetic energy*. (The root of the word is the same as the word "cinema" – in French, kinetic energy is "énergie cinétique.") How does an object's kinetic energy depend on its mass and velocity? Joule attempted a conceptually simple experiment on his honeymoon in the French-Swiss Alps near Mt. Chamonix, in which he measured the difference in temperature between the top and bottom of a waterfall. The water at the top of the falls has some gravitational energy, which isn't our subject right now, but as it drops, that gravitational energy is converted into kinetic energy, and then into heat energy due to internal friction in the churning pool at the bottom:

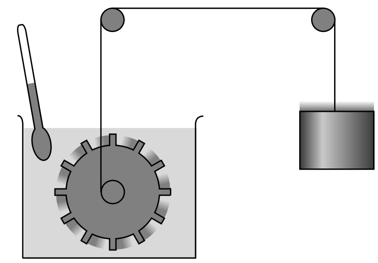
$$\text{gravitational energy} \rightarrow \text{kinetic energy} \rightarrow \text{heat energy}$$

In the logical framework of this book's presentation of energy, the significance of the experiment is that it provides a way to find out how an object's kinetic energy depends on its mass and velocity. The increase in heat energy should equal the kinetic energy of the water just before impact, so in principle we could measure the water's mass, velocity, and kinetic energy, and see how they relate to one another.[3]

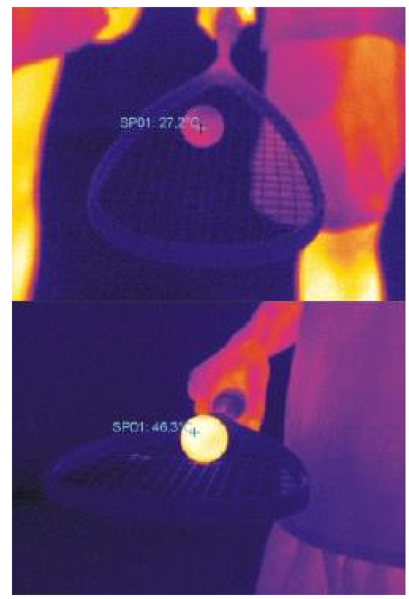
Although the story is picturesque and memorable, most books that mention the experiment fail to note that it was a failure! The problem was that heat wasn't the only form of energy being released. In reality, the situation was more like this:

$$\text{gravitational energy} \rightarrow \text{kinetic energy} \rightarrow \left\{ \text{heat energy} \quad +\text{sound energy} \quad +\text{energy of partial evaporation} \right.$$

The successful version of the experiment, shown in figures d and f, used a paddlewheel spun by a dropping weight. As with the waterfall experiment, this one involves several types of energy, but the difference is that in this case, they can all be determined and taken into account. (Joule even took the precaution of putting a screen between himself and the can of water, so that the infrared

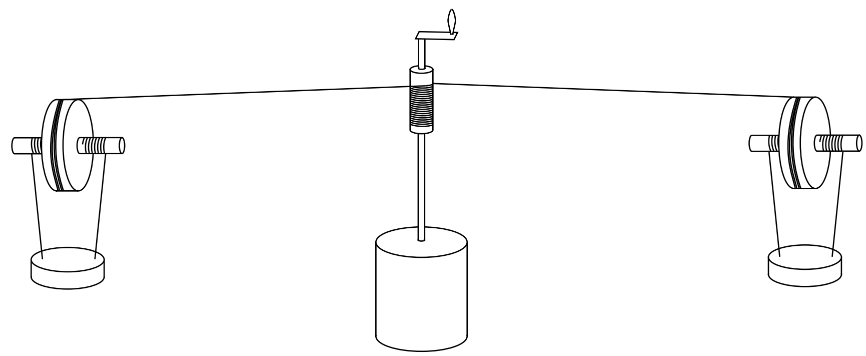


d / A simplified drawing of Joule's paddlewheel experiment.



e / The heating of the tire and floor in figure c is something that the average person might have predicted in advance, but there are other situations where it's not so obvious. When a ball slams into a wall, it doesn't rebound with the same amount of kinetic energy. Was some energy destroyed? No. The ball and the wall heat up. These infrared photos show a squash ball at room temperature (top), and after it has been played with for several minutes (bottom), causing it to heat up detectably.

[3] 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.



f / A realistic drawing of Joule's apparatus, based on the illustration in his original paper. The paddlewheel is sealed inside the can in the middle. Joule wound up the two 13-kg lead weights and dropped them 1.6 meters, repeating this 20 times to produce a temperature change of only about half a degree Fahrenheit in the water inside the sealed can. He claimed in his paper to be able to measure temperatures to an accuracy of 1/200 of a degree.

light emitted by his warm body wouldn't warm it up at all!) The result[4] is

$$K = \frac{1}{2}mv^2 \qquad \text{[kinetic energy]} \qquad .$$

Whenever you encounter an equation like this for the first time, you should get in the habit of interpreting it. First off, we can tell that by making the mass or velocity greater, we'd get more kinetic energy. That makes sense. Notice, however, that we have mass to the first power, but velocity to the second. Having the whole thing proportional to mass to the first power is necessary on theoretical grounds, since energy is supposed to be additive. The dependence on v^2 couldn't have been predicted, but it is sensible. For instance, suppose we reverse the direction of motion. This would reverse the sign of v, because in one dimension we use positive and negative signs to indicate the direction of motion. But since v^2 is what appears in the equation, the resulting kinetic energy is unchanged.

What about the factor of 1/2 in front? It comes out to be exactly 1/2 by the design of the metric system. If we'd been using the old-fashioned British engineering system of units (which is no longer used in the U.K.), the equation would have been $K = (7.44 \times 10^{-2}\ \text{Btu}\cdot\text{s}^2/\text{slug}\cdot\text{ft}^2)mv^2$. The version of the metric system called

[4]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.

the SI,[5] in which everything is based on units of kilograms, meters, and seconds, not only has the numerical constant equal to 1/2, but makes it unitless as well. In other words, we can think of the joule as simply an abbreviation, 1 J=1 $kg \cdot m^2/s^2$. More familiar examples of this type of abbreviation are 1 minute=60 s, and the metric unit of land area, 1 hectare=10000 m^2.

Ergs and joules *example 2*

▷ There used to be two commonly used systems of metric units, referred to as mks and cgs. The mks system, now called the SI, is based on the meter, the kilogram, and the second. The cgs system, which is now obsolete, was based on the centimeter, the gram, and the second. In the cgs system, the unit of energy is not the joule but the erg, 1 erg=1 $g \cdot cm^2/s^2$. How many ergs are in one joule?

▷ The simplest approach is to treat the units as if they were algebra symbols.

$$\begin{aligned} 1\ \text{J} &= 1\ \frac{\text{kg}\cdot\text{m}^2}{\text{s}^2} \\ &= 1\ \frac{\text{kg}\cdot\text{m}^2}{\text{s}^2} \times \frac{1000\ \text{g}}{1\ \text{kg}} \times \left(\frac{100\ \text{cm}}{1\ \text{m}}\right)^2 \\ &= 10^7\ \frac{\text{g}\cdot\text{cm}^2}{\text{s}^2} \\ &= 10^7\ \text{erg} \end{aligned}$$

If you have trouble understanding this example, you should study ch. 0 and 1 of my book **Newtonian Physics**.

Cabin air in a jet airplane *example 3*

▷ A jet airplane typically cruises at a velocity of 270 m/s. Outside air is continuously pumped into the cabin, but must be cooled off first, both because (1) it heats up due to friction as it enters the engines, and (2) it is heated as a side-effect of being compressed to cabin pressure. Calculate the increase in temperature due to the first effect. The specific heat of dry air is about 1.0×10^3 J/kg·°C.

▷ This is easiest to understand in the frame of reference of the plane, in which the air rushing into the engine is stopped, and its kinetic energy converted into heat.[6] Conservation of energy tells us

$$\begin{aligned} 0 &= \Delta E \\ &= \Delta K + \Delta E_{heat} \qquad . \end{aligned}$$

In the plane's frame of reference, the air's initial velocity is v_i=270 m/s, and its final velocity is zero, so the change in its kinetic energy is negative,

$$\begin{aligned} \Delta K &= K_f - K_i \\ &= 0 - (1/2)mv_i^2 \\ &= -(1/2)mv_i^2 \qquad . \end{aligned}$$

[5] Système International

[6] 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.

Assuming that the specific heat of air is roughly independent of temperature (which is why the number was stated with the word "about"), we can substitute into $0 = \Delta K + \Delta E_{heat}$, giving

$$0 = -\frac{1}{2}mv_i^2 + mc\Delta T$$

$$\frac{1}{2}v_i^2 = c\Delta T \qquad .$$

Note how the mass cancels out. This is a big advantage of solving problems algebraically first, and waiting until the end to plug in numbers. With a purely numerical approach, we wouldn't even have known what value of m to pick, or if we'd guessed a value like 1 kg, we wouldn't have known whether our answer depended on that guess.

Solving for ΔT, and writing v instead of v_i for simplicity, we find

$$\begin{aligned}\Delta T &= \frac{v^2}{2c} \\ &\approx 40^\circ\text{C} \qquad .\end{aligned}$$

The passengers would be boiled alive if not for the refrigeration. The first stage of cooling happens via heat exchangers in the engine struts, but a second stage, using a refrigerator under the floor of the cabin, is also necessary. Running this refrigerator uses up energy, cutting into the fuel efficiency of the airplane, which is why typically only 50% of the cabin's air is replaced in each pumping cycle of 2-3 minutes. The airlines emphasize that this is a much faster recirculation rate than in the ventilation systems of most buildings, but people are packed more tightly in an airplane.

2.1.4 Power

Power, P, is defined as the rate of change of energy, $\mathrm{d}E/\mathrm{d}t$. Power thus has units of joules per second, which are usually abbreviated as watts, 1 W=1 J/s. Since energy is conserved, we would have $\mathrm{d}E/\mathrm{d}t = 0$ if E was the total energy of a closed system, and that's not very interesting. What's usually more interesting to discuss is either the power flowing in or out of an open system, or the rate at which energy is being transformed from one form into another. The following is an example of energy flowing into an open system.

Heating by a lightbulb *example 4*

▷ The electric company bills you for energy in units of kilowatt-hours (kilowatts multiplied by hours) rather than in SI units of joules. How many joules is a kilowatt-hour?

▷ 1 kilowatt-hour = (1 kW)(1 hour) = (1000 J/s)(3600 s) = 3.6 MJ.

Now here's an example of energy being transformed from one form into another.

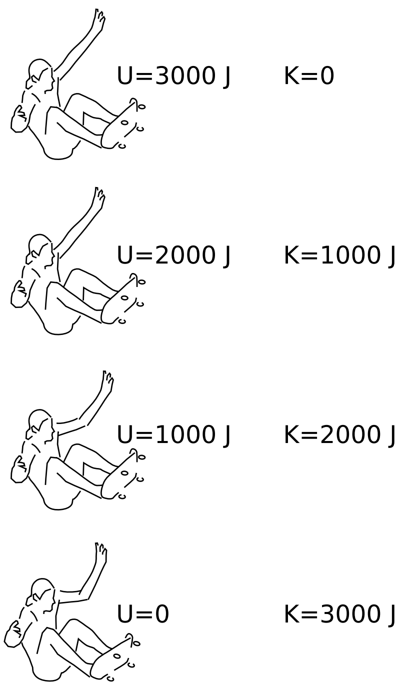
Human wattage *example 5*

▷ Food contains chemical energy (discussed in more detail in section 2.4), and for historical reasons, food energy is normally given in non-SI units of Calories. One Calorie with a capital "C" equals 1000 calories, and 1 calorie is defined as 4.18 J. A typical person consumes 2000 Calories of food in a day, and converts nearly all of that directly to body heat. Compare the person's heat production to the rate of energy consumption of a 100-watt lightbulb.

▷ Strictly speaking, we can't really compute the derivative $\mathrm{d}E/\mathrm{d}t$, since we don't know how the person's metabolism ebbs and flows over the course of a day. What we can really compute is $\Delta E/\Delta t$, which is the power averaged over a one-day period.

Converting to joules, we find $\Delta E = 8 \times 10^6$ J for the amount of energy transformed into heat within our bodies in one day. Converting the time interval likewise into SI units, $\Delta t = 9 \times 10^4$ s. Dividing, we find that our power is 90 J/s = 90 W, about the same as a lightbulb.

g / A skateboarder rises to the edge of an empty pool and then falls back down.



h / The sum of kinetic plus gravitational energy is constant.

2.1.5 Gravitational energy

Gravitational energy, to which I've already alluded, is different from heat and kinetic energy in an important way. Heat and kinetic energy are properties of a single object, whereas gravitational energy describes an interaction between two objects. When the skater in figures g and h is at the top, his distance from the bulk of the planet earth is greater. Since we observe his kinetic energy decreasing on the way up, there must be some other form of energy that is increasing. We invent a new form of energy, called gravitational energy, and written U or U_g, which depends on the distance between his body and the planet. Where is this energy? It's not in the skater's body, and it's not inside the earth, either, since it takes two to tango. If either object didn't exist, there wouldn't be any interaction or any way to measure a distance, so it wouldn't make sense to talk about a distance-dependent energy. Just as marriage is a relationship between two people, gravitational energy is a relationship between two objects.

There is no precise way to define the distance between the skater and the earth, since both are objects that have finite size. As discussed in more detail in section 2.3, gravity is one of the fundamental forces of nature, a universal attraction between any two particles that have mass. Each atom in the skater's body is at a definite distance from each atom in the earth, but each of these distances is different. An atom in his foot is only a few centimeters from some of the atoms in the plaster side of the pool, but most of the earth's atoms are thousands of kilometers away from him. In theory, we might have to add up the contribution to the gravitational energy for every interaction between an atom in the skater's body and an atom in the earth.

For our present purposes, however, there is a far simpler and more practical way to solve problems. In any region of the earth's surface, there is a direction called "down," which we can establish by dropping a rock or hanging a plumb bob. In figure h, the skater is moving up and down in one dimension, and if we did measurements of his kinetic energy, like the made-up data in the figure, we could infer his gravitational energy. As long as we stay within a relatively small range of heights, we find that an object's gravitational energy increases at a steady rate with height. In other words, the strength of gravity doesn't change much if you only move up or down a few meters. We also find that the gravitational energy is proportional to the mass of the object we're testing. Writing y for the height, and g for the overall constant of proportionality, we have

$$U_g = mgy \qquad .$$ [gravitational energy; y=height; only accurate within a small range of heights]

The number g, with units of joules per kilogram per meter, is called the *gravitational field*. It tells us the strength of gravity in a certain region of space. Near the surface of our planet, it has a value of about 9.8 J/kg·m, which is conveniently close to 10 J/kg·m for rough calculations.

Velocity at the bottom of a drop *example 6*

▷ If the skater in figure g drops 3 meters from rest, what is his velocity at the bottom of the pool?

▷ Starting from conservation of energy, we have

$$\begin{aligned} 0 &= \Delta E \\ &= \Delta K + \Delta U \\ &= K_f - K_i + U_f - U_i \\ &= \frac{1}{2}mv_f^2 + mgy_f - mgy_i && \text{(because } K_i\text{=0)} \\ &= \frac{1}{2}mv_f^2 + mg\Delta y \quad , && (\Delta y < 0) \end{aligned}$$

so

$$\begin{aligned} v &= \sqrt{-2g\Delta y} \\ &= \sqrt{-(2)(10 \text{ J/kg}\cdot\text{m})(-3 \text{ m})} \\ &= 8 \text{ m/s} && \text{(rounded to one sig. fig.)} \end{aligned}$$

There are a couple of important things to note about this example. First, we were able to massage the equation so that it only involved Δy, rather than y itself. In other words, we don't need to worry about where $y = 0$ is; any coordinate system will work, as long as the positive y axis points up, not down. This is no accident. Gravitational energy can always be changed by adding a constant onto it, with no effect on the final result, as long as you're consistent within a given problem.

The other interesting thing is that the mass canceled out: even if the skater gained weight or strapped lead weights to himself, his velocity at the bottom would still be 8 m/s. This isn't an accident either. This is the same conclusion we reached in section 1.2, based on the equivalence of gravitational and inertial mass. The kinetic energy depends on the inertial mass, while gravitational energy is related to gravitational mass, but since these two quantities are equal, we were able to use a single symbol, m, for them, and cancel them out.

We can see from the equation $v = \sqrt{-2g\Delta y}$ that a falling object's velocity isn't constant. It increases as the object drops farther and farther. What about its acceleration? If we assume that air friction is negligible, the arguments in section 1.2 show that the acceleration can't depend on the object's mass, so there isn't much else the acceleration *can* depend on besides g. In fact, the acceleration of a falling object equals $-g$ (in a coordinate system where the positive

y axis points up), as we can easily show using the chain rule:

$$\begin{aligned}\left(\frac{\mathrm{d}v}{\mathrm{d}t}\right) &= \left(\frac{\mathrm{d}v}{\mathrm{d}K}\right)\left(\frac{\mathrm{d}K}{\mathrm{d}U}\right)\left(\frac{\mathrm{d}U}{\mathrm{d}y}\right)\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)\\ &= \left(\frac{1}{mv}\right)(-1)(mg)(v)\\ &= -g \qquad ,\end{aligned}$$

where I've calculated $\mathrm{d}v/\mathrm{d}K$ as $1/(\mathrm{d}K/\mathrm{d}v)$, and $\mathrm{d}K/\mathrm{d}U = -1$ can be found by differentiating $K+U =$ (constant) to give $\mathrm{d}K+\mathrm{d}U = 0$.

We can also check that the units of g, J/kg·m, are equivalent to the units of acceleration,

$$\frac{\mathrm{J}}{\mathrm{kg}{\cdot}\mathrm{m}} = \frac{\mathrm{kg}{\cdot}\mathrm{m}^2/\mathrm{s}^2}{\mathrm{kg}{\cdot}\mathrm{m}} = \frac{\mathrm{m}}{\mathrm{s}^2} \qquad ,$$

and therefore the strength of the gravitational field near the earth's surface can just as well be stated as 10 $\mathrm{m/s^2}$.

Speed after a given time *example 7*

▷ An object falls from rest. How fast is it moving after two seconds? Assume that the amount of energy converted to heat by air friction is negligible.

▷ Under the stated assumption, we have $a = -g$, which can be integrated to give $v = -gt +$ constant. If we let $t = 0$ be the beginning of the fall, then the constant of integration is zero, so at $t = 2$ s we have $v = -gt = -(10\ \mathrm{m/s^2}) \times (2\ \mathrm{s}) = 20$ m/s.

The Vomit Comet *example 8*

▷ The U.S. Air Force has an airplane, affectionately known as the Vomit Comet, in which astronaut trainees can experience simulated weightlessness. The plane climbs up high, and then drops straight down like a rock, and since the people are falling with the same acceleration as the plane, the sensation is just like what you'd experience if you went out of the earth's gravitational field. If the plane can start from 10 km up, what is the maximum amount of time for which the dive can last?

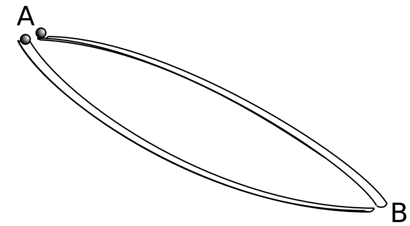
▷ Based on data about acceleration and distance, we want to find time. Acceleration is the second derivative of distance, so if we integrate the acceleration twice with respect to time, we can find how position relates to time. For convenience, let's pick a coordinate system in which the positive y axis is down, so a=g instead of $-g$.

$$\begin{aligned} a &= g \\ v &= gt + \text{constant} && \text{(integrating)} \\ &= gt && \text{(starts from rest)} \\ y &= \frac{1}{2}gt^2 + \text{constant} && \text{(integrating again)} \end{aligned}$$

Choosing our coordinate system to have $y = 0$ at $t = 0$, we can make the second constant of integration equal zero as well, so

$$\begin{aligned} t &= \sqrt{\frac{2y}{g}} \\ &= \sqrt{\frac{2 \cdot 10000\ \text{m}}{10\ \text{m/s}^2}} \\ &= \sqrt{2000\ \text{s}^2} \\ &= 40\ \text{s} \qquad \text{(to one sig. fig.)} \end{aligned}$$

Note that if we hadn't converted the altitude to units of meters, we would have gotten the wrong answer, but we would have been alerted to the problem because the units inside the square root wouldn't have come out to be s^2. In general, it's a good idea to convert all your data into SI (meter-kilogram-second) units before you do anything with them.



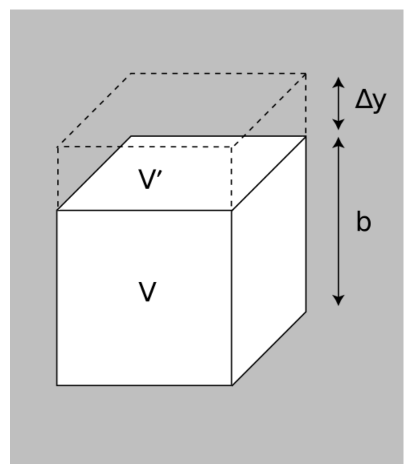
i / Two balls start from rest, and roll from A to B by different paths.

High road, low road — *example 9*

▷ In figure i, what can you say based on conservation of energy about the speeds of the balls when the reach point B? What does conservation of energy tell you about which ball will get there first? Assume friction doesn't convert any mechanical energy to heat or sound energy.

▷ Since friction is assumed to be negligible, there are only two forms of energy involved: kinetic and gravitational. Since both balls start from rest, and both lose the same amount of gravitational energy, they must have the same kinetic energy at the end, and therefore they're rolling at the same speed when they reach B. (A subtle point is that the balls have kinetic energy both because they're moving through space and because they're spinning as they roll. These two types of energy must be in fixed proportion to one another, so this has no effect on the conclusion.)

Conservation of energy does not, however, tell us anything obvious about which ball gets there first. This is a general problem with applying conservation laws: conservation laws don't refer directly to time, since they are statements that something stays the same at all moments in time. We expect on intuitive grounds that the ball that goes by the lower ramp gets to B first, since it builds up speed early on.



j / How much energy is required to raise the submerged box through a height Δy?

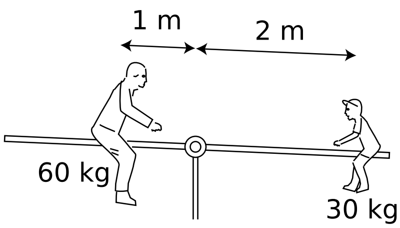
Buoyancy — *example 10*

▷ A cubical box with mass m and volume $V = b^3$ is submerged in a fluid of density ρ. How much energy is required to raise it through a height Δy?

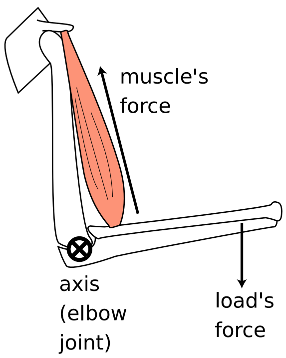
▷ As the box moves up, it invades a volume $V' = b^2\Delta y$ previously occupied by some of the fluid, and fluid flows into an equal volume that it has vacated on the bottom. Lowering this amount of fluid by a height b reduces the fluid's gravitational energy by $\rho V' gb = \rho g b^3 \Delta y$, so the net change in energy is

$$\begin{aligned} \Delta E &= mg\Delta y - \rho g b^3 \Delta y \\ &= (m - \rho V) g \Delta y \qquad . \end{aligned}$$

In other words, it's as if the mass of the box had been reduced by an amount equal to the fluid that otherwise would have occupied that volume. This is known as Archimedes' principle, and it is true even if the box is not a cube, although we'll defer the more general proof until page 149 in the following chapter. If the box is less dense than the fluid, then it will float.



k / A seesaw.



l / The biceps muscle is a reversed lever.

A simple machine — *example 11*

▷ If the father and son on the seesaw in figure k start from rest, what will happen?

▷ Note that although the father is twice as massive, he is at half the distance from the fulcrum. If the seesaw was going to start rotating, it would have to be losing gravitational energy in order to gain some kinetic energy. However, there is no way for it to gain or lose gravitational energy by rotating in either direction. The change in gravitational energy would be

$$\begin{aligned} \Delta U &= \Delta U_1 + \Delta U_2 \\ &= g(m_1 \Delta y_1 + m_2 \Delta y_2) \quad , \end{aligned}$$

but Δy_1 and Δy_2 have opposite signs and are in the proportion of two to one, since the son moves along a circular arc that covers the same angle as the father's but has half the radius. Therefore $\Delta U = 0$, and there is no way for the seesaw to trade gravitational energy for kinetic.

The seesaw example demonstrates the principle of the lever, which is one of the basic mechanical building blocks known as simple machines. As discussed in more detail in chapters 3 and 4, the principle applies even when the forces involved aren't gravitational. (A rigorous definition of "force" is given in chapter 3.)

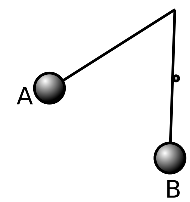
Note that although a lever makes it easier to lift a heavy weight, it also decreases the distance traveled by the load. By reversing the lever, we can make the load travel a greater distance, at the expense of increasing the amount of force required. The human muscular-skeletal system uses reversed levers of this kind, which allows us to move more rapidly, and also makes our bodies more compact, at the expense of brute strength. A piano uses reversed levers so that a small amount of motion of the key produces a longer swing of the hammer. Another interesting example is the hydraulic jack shown in figure n. The analysis in terms of gravitational energy is exactly the same as for the seesaw, except that the relationship between Δy_1 and Δy_2 is now determined not by geometry but by conservation of mass: since water is highly incompressible, conservation of mass is approximately the same as a requirement of constant volume, which can only be satisfied if the distance traveled by each piston is in inverse proportion to its cross-sectional area.

Discussion Questions

A Hydroelectric power (water flowing over a dam to spin turbines) appears to be completely free. Does this violate conservation of energy? If not, then what is the ultimate source of the electrical energy produced by a hydroelectric plant?

B You throw a steel ball up in the air. How can you prove based on conservation of energy that it has the same speed when it falls back into your hand? What if you threw a feather up? Is energy not conserved in this case?

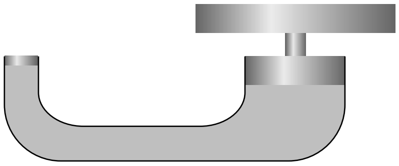
C The figure shows a pendulum that is released at A and caught by a peg as it passes through the vertical, B. To what height will the bob rise on the right?



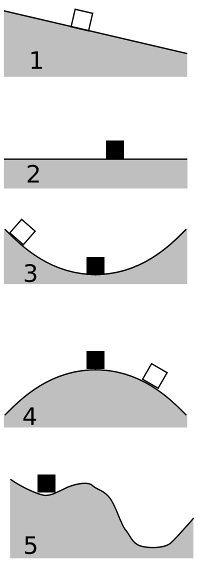
m / Discussion question C.

D What is wrong with the following definitions of g?
(a) "g is gravity."
(b) "g is the speed of a falling object."
(c) "g is how hard gravity pulls on things."

E Two people stand on the edge of a cliff. As they lean over the edge, one person throws a rock down, while the other throws one straight up with an exactly opposite initial velocity. Compare the accelerations of the two rocks, and compare the speeds of the rocks on impact at the bottom of the cliff.



n / A hydraulic jack.



o / The surfaces are frictionless. The black blocks are in equilibrium.

2.1.6 Equilibrium and stability

The seesaw in figure k is in equilibrium, meaning that if it starts out being at rest, it will stay put. This is known as a neutral equilibrium, since the seesaw has no preferred position to which it will return if we disturb it. If we move it to a different position and release it, it will stay at rest there as well. If we put it in motion, it will simply continue in motion until one person's feet hit the ground.

Most objects around you are in stable equilibria, like the black block in figure o/3. Even if the block is moved or set in motion, it will oscillate about the equilibrium position. The pictures are like graphs of y versus x, but since the gravitational energy $U = mgy$ is proportional to y, we can just as well think of them as graphs of U versus x. The block's stable equilibrium position is where the function $U(x)$ has a local minimum. The book you're reading right now is in equilibrium, but gravitational energy isn't the only form of energy involved. To move it upward, we'd have to supply gravitational energy, but downward motion would require a different kind of energy, in order to compress the table more. (As we'll see in section 2.4, this is electrical energy due to interactions between atoms within the table.)

A differentiable function's local extrema occur where its derivative is zero. A position where $\mathrm{d}U/\mathrm{d}x$ is zero can be a stable (3), neutral (2), or unstable equilibrium, (4). An unstable equilibrium is like a pencil balanced on its tip. Although it could theoretically remain balanced there forever, in reality it will topple due to any tiny perturbation, such as an air current or a vibration from a passing truck. This is a technical, mathematical definition of instability, which is more restrictive than the way the word is used in ordinary speech. Most people would describe a domino standing upright as being unstable, but in technical usage it would be considered stable, because a certain finite amount of energy is required to tip it over, and perturbations smaller than that would only cause it to oscillate around its equilibrium position.

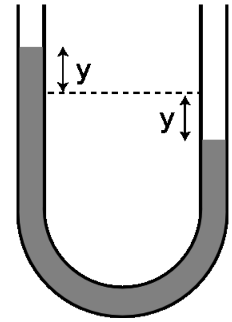
The domino is also an interesting example because it has two local minima, one in which it is upright, and another in which it is lying flat. A local minimum that is not the global minimum, as in figure o/5, is referred to as a metastable equilibrium.

Water in a U-shaped tube — *example 12*

▷ The U-shaped tube in figure p has cross-sectional area A, and the density of the water inside is ρ. Find the gravitational energy as a function of the quantity y shown in the figure, and show that there is an equilibrium at y=0.

▷ The question is a little ambiguous, since gravitational energy is only well defined up to an additive constant. To fix this constant, let's define U to be zero when y=0. The difference between $U(y)$ and $U(0)$ is the energy that would be required to lift a water column of height y out of the right side, and place it above the dashed line, on the left side, raising it through a height y. This water column has height y and cross-sectional area A, so its volume is Ay, its mass is ρAy, and the energy required is mgy=$(\rho Ay)gy$=ρgAy^2. We then have $U(y) = U(0) + \rho gAy^2 = \rho gAy^2$.

To find equilibria, we look for places where the derivative $\mathrm{d}U/\mathrm{d}y = 2\rho gAy$ equals 0. As we'd expect intuitively, the only equilibrium occurs at y=0. The second derivative test shows that this is a local minimum (not a maximum or a point of inflection), so this is a stable equilibrium.



p / Water in a U-shaped tube.

2.1.7 Predicting the direction of motion

Kinetic energy doesn't depend on the direction of motion. Sometimes this is helpful, as in the high road-low road example (p. 45, example 9), where we were able to predict that the balls would have the same final speeds, even though they followed different paths and were moving in different directions at the end. In general, however, the two conservation laws we've encountered so far aren't enough to predict an object's path through space, for which we need conservation of momentum (chapter 3), and the mathematical technique of vectors. Before we develop those ideas in their full generality, however, it will be helpful to do a couple of simple examples, including one that we'll get a lot of mileage out of in section 2.3.

Suppose we observe an air hockey puck gliding frictionlessly to the right at a velocity v, and we want to predict its future motion. Since there is no friction, no kinetic energy is converted to heat. The only form of energy involved is kinetic energy, so conservation of energy, $\Delta E = 0$, becomes simply $\Delta K = 0$. There's no particular reason for the puck to do anything but continue moving to the right at constant speed, but it would be equally consistent with conservation of energy if it spontaneously decided to reverse its direction of motion, changing its velocity to $-v$. Either way, we'd have $\Delta K = 0$. There is, however, a way to tell which motion is physical and which is unphysical. Suppose we consider the whole thing again in the frame of reference that is initially moving right along with the puck. In this frame, the puck starts out with $K = 0$. What we originally described as a reversal of its velocity from v to $-v$ is, in this new frame of reference, a change from zero velocity to $-2v$, which would violate conservation of energy. In other words, the physically possible motion conserves energy in all frames of reference, but the unphysical motion only conserves energy in one special frame of reference.

q / A car drives over a cliff.

For our second example, we consider a car driving off the edge of a cliff (q). For simplicity, we assume that air friction is negligible, so only kinetic and gravitational energy are involved. Does the car follow trajectory 1, familiar from Road Runner cartoons, trajectory 2, a parabola, or 3, a diagonal line? All three could be consistent with conservation of energy, in the ground's frame of reference. For instance, the car would have constant gravitational energy along the initial horizontal segment of trajectory 1, so during that time it would have to maintain constant kinetic energy as well. Only a parabola, however, is consistent with conservation of energy combined with Galilean relativity. Consider the frame of reference that is moving horizontally at the same speed as that with which the car went over the edge. In this frame of reference, the cliff slides out from under the initially motionless car. The car can't just hover for a while, so trajectory 1 is out. Repeating the same math as in example 8 on p. 44, we have

$$x^* = 0\ , \qquad y^* = (1/2)gt^2$$

in this frame of reference, where the stars indicate coordinates measured in the moving frame of reference. These coordinates are related to the ground-fixed coordinates (x, y) by the equations

$$x = x^* + vt \qquad \mathit{and} \qquad y = y^* \qquad ,$$

where v is the velocity of one frame with respect to the other. We therefore have

$$x = vt\ , \qquad y = (1/2)gt^2 \qquad ,$$

in our original frame of reference. Eliminating t, we can see that this has the form of a parabola:

$$y = (g/2v^2)x^2 \qquad .$$

Self-Check

What would the car's motion be like in the * frame of reference if it followed trajectory 3? ▷ Answer, p. 706

2.2 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 691 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,

```
>>> 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:

```
>>> 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 44 that the exact answer, found using calculus, is $\sqrt{2h/g}$. 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 $mgh = (1/2)mv^2 + mgy$ for the velocity v at height y, so $v = -\sqrt{-2g(h-y)}$. (We choose the negative root because the object is moving down, and our coordinate system has the positive y 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 $\sqrt{2 \cdot 1.0\ \text{m}/9.8\ \text{m/s}^2} = 0.45$ s. With the drop split up into only 10 equal height intervals, the numerical technique provides a pretty lousy approximation:

```
>>> 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:

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

A subtle point is that we changed `y` in line 9, and *then* we calculated `v` in line 10, 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_avg = -(math.sqrt(2*g*(h-y_old))+math.sqrt(2*g*(h-y)))/2.
    dt=dy/v_avg
    t=t+dt
  return t
```

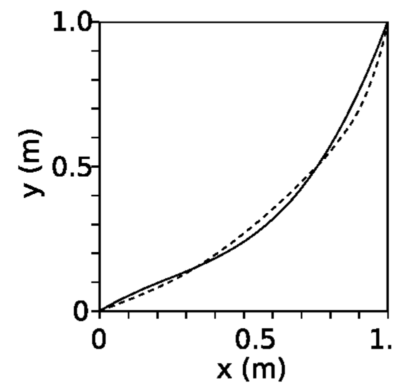
```
>>> time3(1.0,1) # h, n
0.45175395145262565
```

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 (a, b) to the point $(0, 0)$. 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 a=1.0 m and b=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,

$$y = c_1 x + c_2 x^2 + c_3 x^3 \qquad ,$$

where we require $c_3 = (b - c_1 a - c_2 a^2)/a^3$ in order to make the curve pass through the point (a, b). The Python program, below, is not much different from what we've done before. The function only asks for c_1 and c_2, and calculates c_3 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.



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.

```
1   import math
2   def timeb(a,b,c1,c2,n):
3     g=9.8
4     c3 = (b-c1*a-c2*a**2)/(a**3)
5     x=a
6     y=b
7     v=0
8     dx = -a/n
9     t=0
10    for i in range(n):
11      y_old = y
12      x = x+dx
13      y = c1*x+c2*x**2+c3*x**3
14      dy = y-y_old
15      v_avg = (math.sqrt(2*g*(b-y_old))+math.sqrt(2*g*(b-y)))/2.
16      ds = math.sqrt(dx**2+dy**2) # Pythagorean thm.
17      dt=ds/v_avg
18      t=t+dt
19    return t
```

As a first guess, we could try a straight diagonal line, $y = x$, which corresponds to setting $c_1 = 1$, and all the other coefficients to zero. The result is a fairly long time:

```
>>> b=1.
>>> n=10000
>>> c1=1.
>>> c2=0.
>>> 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 $y = x^3$:

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

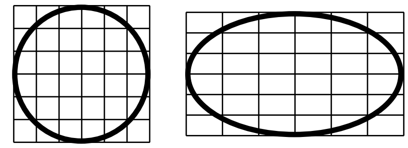
This is a significant improvement, and turns out to be only a hundredth of a second off of the shortest possible time! It's possible, although not very educational or entertaining, to find better approximations to the brachistochrone curve by fiddling around with the coefficients of the polynomial by 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,

$$y = (0.62)x + (-0.93)x^2 + (1.31)x^3$$

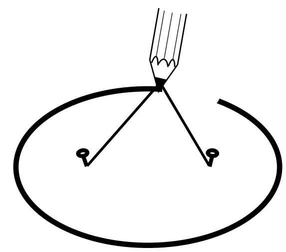
by using the program listed in appendix 2 on page 693 to search automatically for the optimal curve. The seventh-order approximation shown in the figure came from a straightforward extension of the same program.

2.3 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.



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

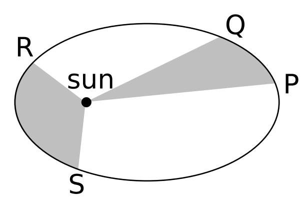


b / An ellipse can be constructed by tying a string to two pins and drawing like this with a pencil stretching the string taut. Each pin constitutes one focus of the ellipse.

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

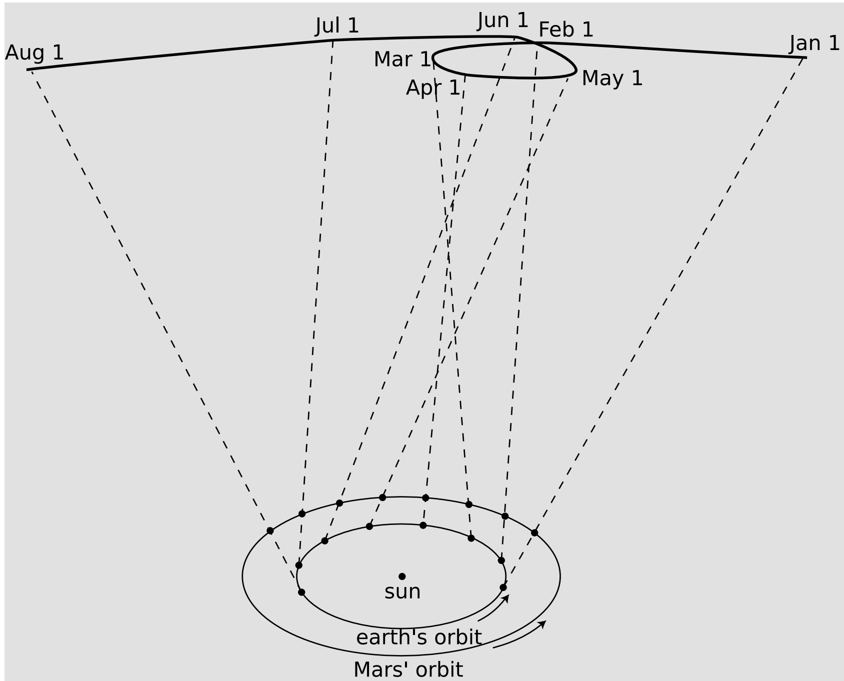
Newton wouldn't have been able to figure out *why* the planets move the way they do if it hadn't been for the astronomer Tycho Brahe (1546-1601) and his protege Johannes Kepler (1571-1630), who together came up with the first simple and accurate description of *how* the planets actually do move. The 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.



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.

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



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.

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:

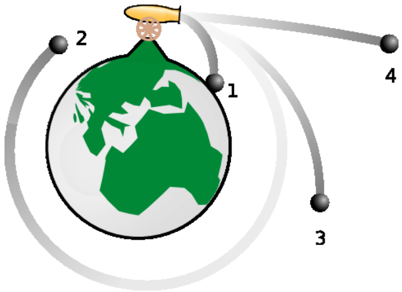
Kepler's elliptical orbit law: The planets orbit the sun in elliptical orbits with the sun at one focus.

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

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.



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**.

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 50 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 A 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 C 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

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

Low-earth orbit *example 13*

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 m/s^2. Taking numerical data from Appendix 5, we have

$$\begin{aligned} v &= \sqrt{gr} \\ &= \sqrt{(10\ \text{m/s}^2)(6.4\times10^3\ \text{km})} \\ &= \sqrt{(10\ \text{m/s}^2)(6.4\times10^6\ \text{m})} \\ &= \sqrt{6.4\times10^7\ \text{m}^2/\text{s}^2} \\ &= 8000\ \text{m/s} \end{aligned}$$

(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.[7] 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.

2.3.4 Gravitational energy in general

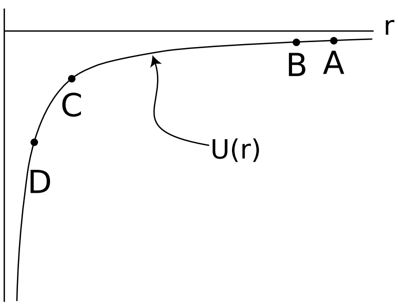
We now want to find an equation for the gravitational energy of any two masses that attract each other from a distance r. We assume that r is large enough compared to the distance between the objects so that we don't really have to worry about whether r is measured from center to center or in some other 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

[7]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. 103.

for all values of r, and g depends on r. We can, however, consider an infinitesimally small change in distance $\mathrm{d}r$, for which we'll have $\mathrm{d}U = m_2 g_1 \mathrm{d}r$, where g_1 is the gravitational field created by m_1. (We could just as well have written this as $\mathrm{d}U = m_1 g_2 \mathrm{d}r$, since we're not assuming either mass is "special" or "active.") Integrating this equation, we have

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



f / The gravitational energy $U = -Gm_1m_2/r$ graphed as a function of r.

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:

$$U = -\frac{Gm_1m_2}{r} \qquad \text{[gravitational energy of two masses separated by a distance } r\text{]}$$

Let's interpret this. 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.

Determining G — *example 14*

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

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

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 65, 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 modern versions of the same experiment, is $6.67 \times 10^{-11}\ \mathrm{J \cdot m/kg^2}$.

Escape velocity — *example 15*

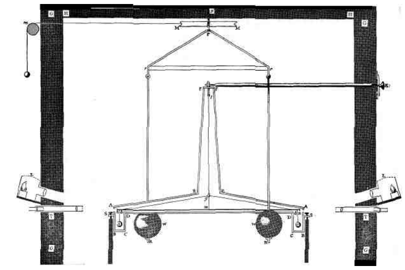
▷ The Pioneer 10 space probe was launched in 1972, and continued sending back signals for 30 years. In the year 2001, not long before contact with the probe was lost, it was about 1.2×10^{13} m from the sun, and was moving almost directly away from the sun at a velocity of 1.21×10^4 m. The mass of the sun is 1.99×10^{30} 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:

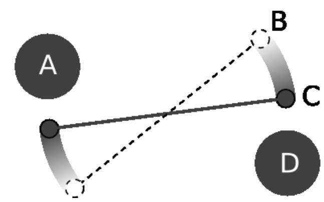
$$\begin{aligned} K_i + U_i &= U_f \\ \frac{1}{2}mv^2 - \frac{GMm}{r_i} &= -\frac{GMm}{r_f} \\ \frac{1}{2}v^2 - \frac{GM}{r_i} &= -\frac{GM}{r_f} \qquad , \end{aligned}$$

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.

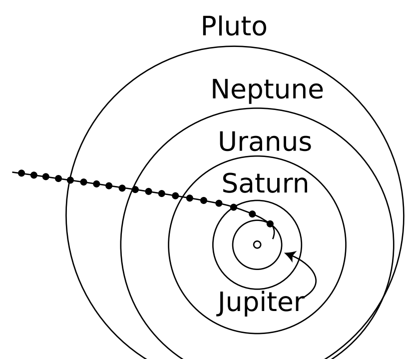
The minimum velocity required for this to happen is called *escape velocity*. For speeds above escape velocity, the orbits are open-ended hyperbolas, rather than repeating elliptical orbits. In figure i, Pioneer's hyperbolic trajectory becomes almost indistinguishable from a line at large distances from 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.



g / Cavendish's original drawing of the apparatus for his experiment, discussed in example 14. 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.



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.

Pioneer 10, by the way, is at the center of a remarkable mystery about the nature of gravity.[8] The probe is still being tracked by the worldwide Deep Space Network of radio dishes, and the precision of these measurements is so good that the acceleration caused by the sun's gravity can be determined to an accuracy of 10^{-11} m/s^2. After accounting for all known nongravitational effects, the measured deceleration is stronger than predicted by 9×10^{-10} m/s^2. The same effect is being observed with Pioneer 10's sister probe, Pioneer 11, which is headed out of the solar system in the opposite direction. Some physicists have interpreted this as evidence that gravitational energy behaves just slightly differently than $-Gm_1m_2/r$, but it may be due instead to some behavior of the probes that is not completely understood — after all, we can't inspect them now!

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:

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

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 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?

[8] J.D. Anderson et al., http://arxiv.org/abs/gr-qc/0104064

2.3.5 The shell theorem

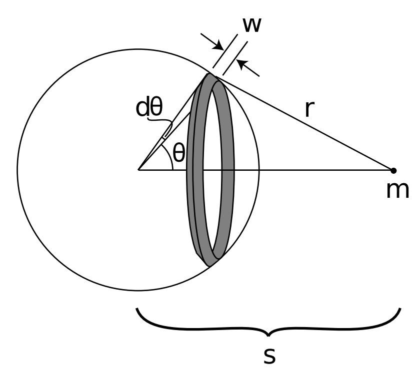
Newton's great insight was that gravity near the earth's surface was the same kind of interaction as the one that kept the planets from flying away from the sun. He told his niece that the idea came to him when he saw an apple fall from a tree, which made him wonder whether the earth might be affecting 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.

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

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.

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 $\theta + \mathrm{d}\theta$. 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 $w = b\mathrm{d}\theta$, and its mass is $\mathrm{d}M = (\rho)$(circumference)(thickness)(width)= $(\rho)(2\pi b\sin\theta)(h)(b\mathrm{d}\theta)$=$2\pi b^2 h\sin\theta\mathrm{d}\theta$. The gravitational energy of the ring interacting with mass m is therefore

$$\begin{aligned}\mathrm{d}U &= -\frac{Gm\mathrm{d}M}{r}\\ &= -2\pi G\rho b^2 hm\frac{\sin\theta\mathrm{d}\theta}{r}\end{aligned} \quad .$$



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

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

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

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

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

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

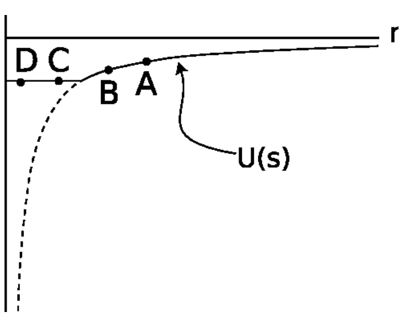
$$2r\mathrm{d}r = 2bs\sin\theta\mathrm{d}\theta \qquad ,$$

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

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

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:

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



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

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.

Newton's apple *example 16*

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, $g = v^2/r = 2.72 \times 10^{-3}$ 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]

Weighing the earth — *example 17*

▷ Once Cavendish had found $G = 6.67\times 10^{-11}$ J·m/kg^2 (p. 63, example 14), 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 $mg\mathrm{d}r = \mathrm{d}U$, or $g = (1/m)\mathrm{d}U/\mathrm{d}r = GM/r^2$. Solving for M, we have

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

Gravity inside the earth — *example 18*

▷ 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

[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.

deeper than r. Under the assumption of constant density, this mass is related to the total mass of the earth by

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

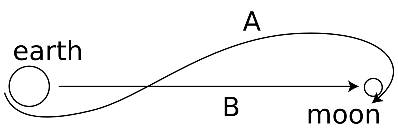
and by the same reasoning as in example 17,

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

so

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

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



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

The following example applies the numerical techniques of section 2.2.

From the earth to the moon *example 19*

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	3.363×10^5 m
initial velocity	1.083×10^4 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

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

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

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

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 54.

```
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
>>> 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
>>> 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.

2.4 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.

a / A vivid demonstration that heat is a form of motion. A small amount of boiling water is poured into the empty can, which rapidly fills up with hot steam. The can is then sealed tightly, and soon crumples.

2.4.1 Heat is kinetic energy.

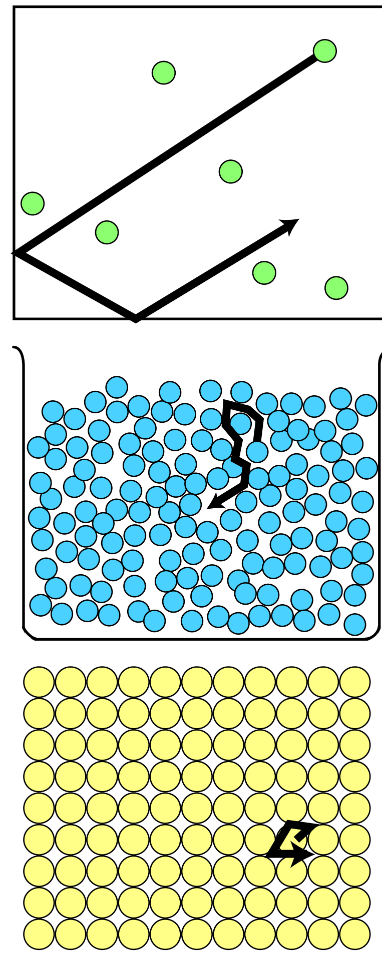
What is heat really? Is it an invisible fluid that your bare feet soak up from a hot sidewalk? Can one ever remove all the heat from an object? Is there a maximum to the temperature scale?

The theory of heat as a fluid seemed to explain why colder objects absorbed heat from hotter ones, but once it became clear that heat was a form of energy, it began to seem unlikely that a material substance could transform itself into and out of all those other forms of energy like motion or light. For instance, a compost pile gets hot, and we describe this as a case where, through the action of bacteria, chemical energy stored in the plant cuttings is transformed into heat energy. The heating occurs even if there is no nearby warmer object that could have been leaking heat fluid into the pile.

An alternative interpretation of heat was suggested by the theory that matter is made of atoms. Since gases are thousands of times less dense than solids or liquids, the atoms (or clusters of atoms called molecules) in a gas must be far apart. In that case, what is keeping all the air molecules from settling into a thin film on the floor of the room in which you are reading this book? The simplest explanation is that they are moving very rapidly, continually ricocheting off of the floor, walls, and ceiling. Though bizarre, the cloud-of-bullets image of a gas did give a natural explanation for the surprising ability of something as tenuous as a gas to exert huge forces.

The experiment shown in figure a, for instance, can be explained as follows. The high temperature of the steam is interpreted as a high average speed of random motions of its molecules. Before the lid was put on the can, the rapidly moving steam molecules pushed their way out of the can, forcing the slower air molecules out of the way. As the steam inside the can thinned out, a stable situation was soon achieved, in which the force from the less dense steam molecules moving at high speed balanced against the force from the more dense but slower air molecules outside. The cap was put on, and after a while the steam inside the can began to cool off. The force from the cooler, thin steam no longer matched the force from the cool, dense air outside, and the imbalance of forces crushed the can.

This type of observation leads naturally to the conclusion that hotter matter differs from colder in that its atoms' random motion is more rapid. In a liquid, the motion could be visualized as people in a milling crowd shoving past each other more quickly. In a solid, where the atoms are packed together, the motion is a random vibration of each atom as it knocks against its neighbors.



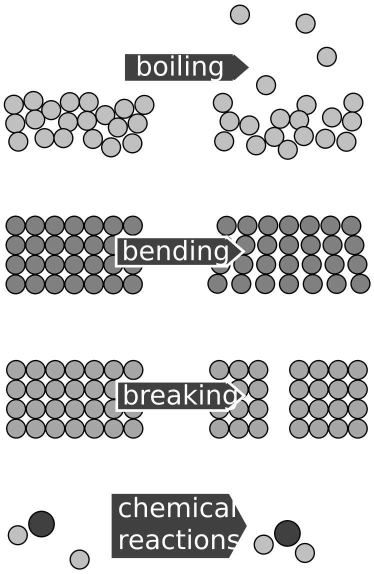
b / Random motion of atoms in a gas, a liquid, and a solid.

We thus achieve a great simplification in the theory of heat. Heat is simply a form of kinetic energy, the total kinetic energy of random motion of all the atoms in an object. With this new understanding, it becomes possible to answer at one stroke the questions posed at the beginning of the section. Yes, it is at least theoretically possible to remove all the heat from an object. The coldest possible temperature, known as absolute zero, is that at which all the atoms have zero velocity, so that their kinetic energies, $K = (1/2)mv^2$, are all zero. No, there is no maximum amount of heat that a certain quantity of matter can have, and no maximum to the temperature scale, since arbitrarily large values of v can create arbitrarily large amounts of kinetic energy per atom.

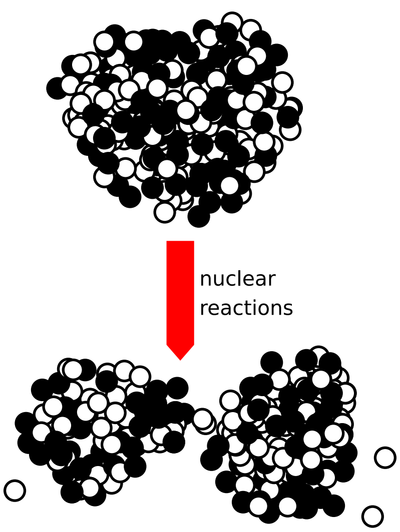
The kinetic theory of heat also provides a simple explanation of the true nature of temperature. Temperature is a measure of the amount of energy per molecule, whereas heat is the total amount of energy possessed by all the molecules in an object.

There is an entire branch of physics, called thermodynamics, that deals with heat and temperature and forms the basis for technologies such as refrigeration. Thermodynamics is discussed in more detail in chapter 5, and I've provided here only a brief overview of the thermodynamic concepts that relate directly to energy.

c / The spring's energy is really due to electrical interactions among atoms.



d / All these energy transformations turn out at the atomic level to be due to changes in the distances between atoms that interact electrically.



e / This figure looks similar to the previous ones, but the scale is a million times smaller. The little balls are the neutrons and protons that make up the tiny nucleus at the center of a uranium atom. When the nucleus splits (fissions), the source of the kinetic energy is partly electrical and partly nuclear.

2.4.2 All energy comes from particles moving or interacting.

If I stretch the spring in figure c and then release it, it snaps taut again. The creation of some kinetic energy shows that there must have been some other form of energy that was destroyed. What was it?

We could just invent a new type of energy called "spring energy," study its behavior, and call it quits, but that would be ugly. Are we going to have to invent a new forms of energy like this, over and over? No: the title of this book doesn't lie, and physics really is fundamentally simple. As shown in figure d, when we bend or stretch an object, we're really changing the distances between the atoms, resulting in a change in electrical energy. Electrical energy isn't really our topic right now — that's what most of the second half of this book is about — but conceptually it's very similar to gravitational energy. Like gravitational energy, it depends on $1/r$, although there are some interesting new phenomena, such as the existence of both attraction and repulsion, which doesn't occur with gravity because gravitational mass can't be negative. The real point is that all the apparently dissimilar forms of energy in figure d turn out to be due to electrical interactions among atoms. Even if we wish to include nuclear reactions (figure e) in the picture, there still turn out to be only four fundamental types of energy:

kinetic energy (including heat)
gravitational energy
electrical and magnetic energy
nuclear energy

Astute students have often asked me how light fits into this picture. This is a very good question, and in fact it could be argued that it is the basic question that led to Einstein's theory of relativity as well as the modern quantum picture of nature. Since these are topics for the second half of the book, we'll have to be content with half an answer at this point. For now, we may think of light energy as a form of kinetic energy, but one calculated not according to $(1/2)mv^2$ but by some other equation. (We know that $(1/2)mv^2$ would not make sense, because light has no mass, and furthermore, high-energy beams of light do not differ in speed from low-energy ones.)

Temperature during boiling *example 20*

▷ If you stick a thermometer in a pan of water, and watch the temperature as you bring the water to a boil, you'll notice an interesting fact. The temperature goes up until the boiling point is reached, but then stays at 100 °C during the whole time the water is being boiled off. The temperature of the steam is also 100 °C. Why does the temperature "stick" like this? What's happening to all the energy that the stove's burner is putting into the pan?

▷ As shown in figure d, boiling requires an increase in electrical energy, because the atoms coming out as gas are moving away from the other atoms, which attract them electrically. It is only this electrical energy that is increasing, not the atoms' kinetic energy, which is what the thermometer can measure.

Diffusion *example 21*

▷ A drop of food coloring in a cup of water will gradually spread out, even if you don't do any mixing with a spoon. This is called diffusion. Why would this happen, and what effect would temperature have? What would happen with solids or gases?

▷ Figure b shows that the atoms in a liquid mingle because of their random thermal motion. Diffusion is slow (typically on the order of a centimeter a minute), despite the *high* speeds of the atoms (typically hundreds of miles per hour). This is due to the randomness of the motion: a particular atom will take a long time to travel any significant distance, because it doesn't travel in a straight line.

Based on this picture, we expect that the speed of diffusion should increase as a function of temperature, and experiments show that this is true.

Diffusion also occurs in gases, which is why you can smell things even when the air is still. The speeds are much faster, because the typical distance between collisions is much longer than in a liquid.

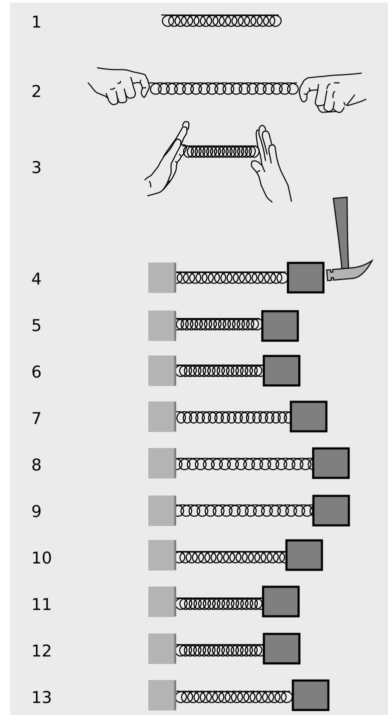
We can see from figure b that diffusion won't occur in solids, because each atom vibrates around an equilibrium position.

Discussion Questions

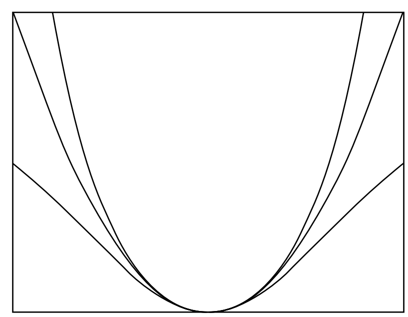
A I'm not making this up. XS Energy Drink has ads that read like this: *All the "Energy" ... Without the Sugar! Only 8 Calories!"* Comment on this.

2.5 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, U, 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 U. 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.



b / Three functions with the same curvature at *x*=0.

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 $x = 0$ corresponds to the position of the mass where the spring is in equilibrium, and since interaction energies like U are only well defined up to an additive constant, we'll simply define it to be zero at equilibrium:

$$U(0) = 0$$

Since $x = 0$ is an equilibrium, $U(x)$ must have a local minimum there, and a differentiable function (which we assume U is) has a zero derivative at a local minimum:

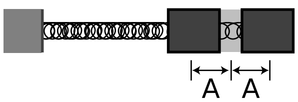
$$\frac{\mathrm{d}U}{\mathrm{d}x}(0) = 0$$

There are still infinitely many functions that could satisfy these criteria, including the three shown in figure b, which are $x^2/2$, $x^2/2(1+x^2)$, and $(e^{3x} + e^{-3x} - 2)/18$. 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 $x = 0$, and the second derivative is a measure of curvature. We write k for the second derivative of the energy at an equilibrium point,

$$\frac{\mathrm{d}^2U}{\mathrm{d}x^2}(0) = k \qquad .$$

Physically, k is a measure of stiffness. For example, the heavy-duty springs in a car's shock absorbers would have a high value of k. 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 $U(0) = 0$, $\mathrm{d}U/\mathrm{d}x = 0$, and $\mathrm{d}^2U/\mathrm{d}x^2 = k$, with the same value of k, are virtually indistinguishable for small values of x, so if we want to analyze small oscillations, it doesn't even matter which function we assume. For simplicity, we'll just use $U(x) = (1/2)kx^2$ 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, T, 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, k, the mass, m, and and the amplitude, A, defined in figure c.[10]



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 54 and modify it for our purposes. For flexibility, we'll define the function $U(x)$ 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.

```
1   import math
2   def u(k,x):
3     return .5*k*x**2
4
5   def osc(m,k,a,n):
6     x=a
7     v=0
8     dx = -2.*a/n
9     t=0
10    e = u(k,x)+.5*m*v**2
11    for i in range(n):
12      x_old = x
13      v_old = v
14      x = x+dx
15      kinetic = e-u(k,x)
```

[10]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.

```
      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
```

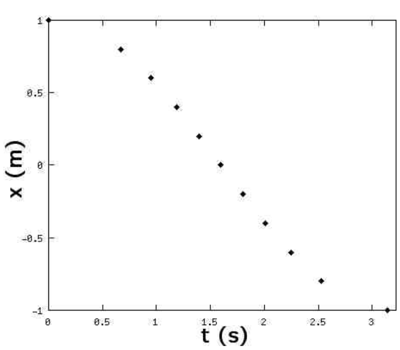
```
>>> osc(1.,1.,1.,100000)
warning, K= -1.43707268307e-12 <0
6.2831854132667919
```

The first thing to notice is that with this particular set of inputs (m=1 kg, $k = 1\ \mathrm{J/m^2}$, and $A = 1$ m), the program has done an excellent job of computing $2\pi = 6.2831853\ldots$. 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:

```
>>> 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 start looking for an algebraic solution. For guidance, here's a graph of x as a function of t, as calculated by the `osc` function with `n`=10.

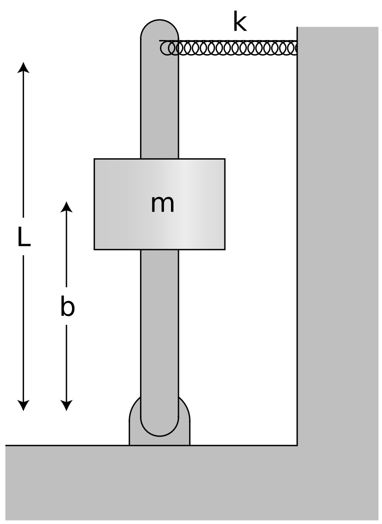


This looks like a cosine function, so let's see if a $x = A\cos(\omega t+\delta)$ is a solution to the conservation of energy equation — it's not uncommon to try to "reverse-engineer" the cryptic results of a numerical calculation like this. The symbol $\omega = 2\pi/T$ (Greek omega) is a standard symbol for the number of radians per second of oscillation, and the phase angle δ is to allow for the possibility that $t = 0$

doesn't coincide with the beginning of the motion. The energy is

$$\begin{aligned}
E &= K + U \\
&= \frac{1}{2}mv^2 + \frac{1}{2}kx^2 \\
&= \frac{1}{2}m\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \frac{1}{2}kx^2 \\
&= \frac{1}{2}m\left[-A\omega\sin\left(\omega t+\delta\right)\right]^2 + \frac{1}{2}k\left[A\cos\left(\omega t+\delta\right)\right]^2 \\
&= \frac{1}{2}A^2\left[m\omega^2\sin^2(\omega t+\delta) + k\cos^2(\omega t+\delta)\right]
\end{aligned}$$

According to conservation of energy, this has to be a constant. Using the identity $\sin^2+\cos^2=1$, we can see that it will be a constant if we have $m\omega^2=k$, or $\omega=\sqrt{k/m}$, i.e. $T=2\pi\sqrt{m/k}$. Note that the period is independent of amplitude.



d / Example 22.

A spring and a lever *example 22*

▷ 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.

▷ This is a little tricky, because the spring constant k, although it is relevant, is *not* the k we should be putting into the equation $T=2\pi\sqrt{m/k}$. The k that goes in there has to be the second derivative of U with respect to the position, x, of the thing that's moving. The energy U stored in the spring depends on how far the *tip* of the lever is from the center. This distance equals $(L/b)x$, so the energy in the spring is

$$\begin{aligned}
U &= \frac{1}{2}k\left(\frac{L}{b}x\right)^2 \\
&= \frac{kL^2}{2b^2}x^2 \qquad ,
\end{aligned}$$

and the k we have to put in $T=2\pi\sqrt{m/k}$ is

$$\frac{\mathrm{d}^2U}{\mathrm{d}x^2} = \frac{kL^2}{b^2} \qquad .$$

The result is

$$\begin{aligned}
T &= 2\pi\sqrt{\frac{mb^2}{kL^2}} \\
&= \frac{2\pi b}{L}\sqrt{\frac{m}{k}}
\end{aligned}$$

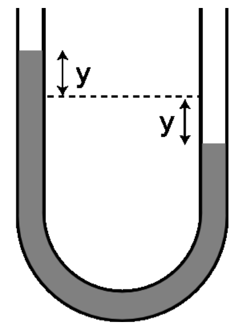
The leverage of the lever makes it as if the spring was stronger, and decreases the period of the oscillations by a factor of b/L.

Water in a U-shaped tube *example 23*

▷ What is the period of oscillation of the water in figure e?

▷ In example 12 on p. 48, we found $U(y) = \rho g A y^2$, so the "spring constant," which really isn't a spring constant here at all, is

$$\begin{aligned} k &= \frac{d^2U}{dy^2} \\ &= 2\rho g A \quad . \end{aligned}$$



e / Water in a U-shaped tube.

This is an interesting example, because k 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, dy/dt, then the mass-on-a-spring analysis applies. Letting L be the total length of the filled part of the tube, the mass is $\rho L A$, and we have

$$\begin{aligned} T &= 2\pi\sqrt{m/k} \\ &= 2\pi\sqrt{\frac{\rho L A}{2\rho g A}} \\ &= 2\pi\sqrt{\frac{L}{2g}} \quad . \end{aligned}$$

This chapter is summarized on page 723. Notation and terminology are tabulated on pages 718-719.

Problems

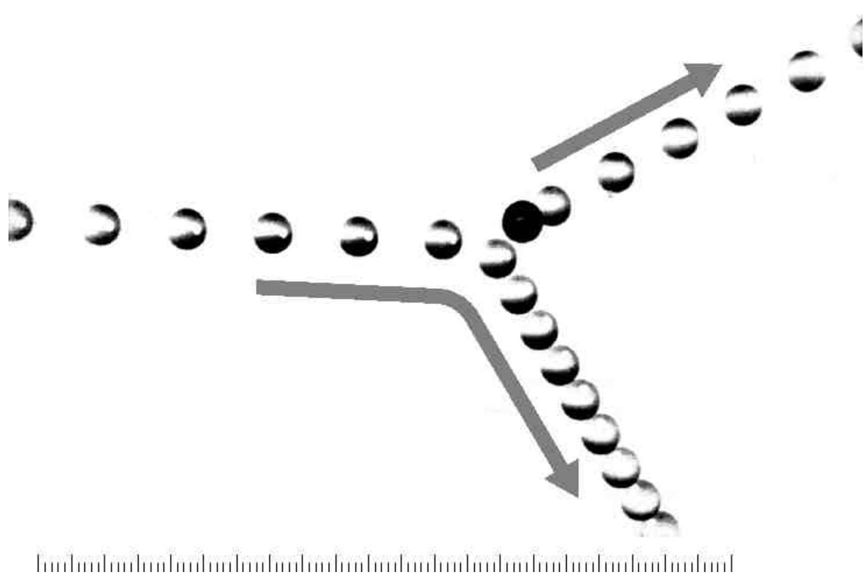
The symbols √, ■, etc. are explained on page 86.

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, p. 712 ■

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, p. 712 ■

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, p. 713 ■

4 The multiflash photograph below 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 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.] ■



Problem 4.

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, p. 713

6 You jump up straight up in the air. When do you 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 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. (c) 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. ✓

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

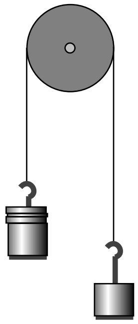
(a) How much energy is required in order to convert 1.00 g of ice at -20 °C into steam at 137 °C? ✓
(b) What is the minimum amount of hot water that could melt 1 g of ice? ✓

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.

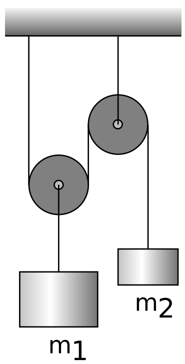
14 (a) A circular hoop of mass m and radius r spins like a wheel while its center remains at rest. Let ω be the number of radians it covers per unit time, i.e. $\omega = 2\pi/T$, where the period, T, is the time for one revolution. Show that its kinetic energy equals $(1/2)m\omega^2r^2$.
(b) Show that 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 v, its kinetic energy equals $(1/2)mv^2$, 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 44, I used the chain rule to prove that the acceleration of a free-falling object is given by $a = -g$. In this problem, you'll use a different technique to prove the same thing. Assume that the acceleration is a constant, a, and then integrate to find v and y, including appropriate constants of integration. Plug your expressions for v and y into the equation for the total energy, and show that $a = -g$ is the only value that results in constant energy.



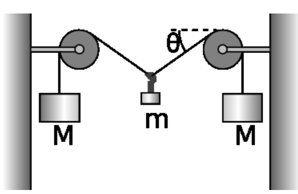
Problem 16.

16 The figure shows two unequal masses, m_1 and m_2, connected by a string running over a pulley. Find the acceleration.
▷ Hint, p. 703



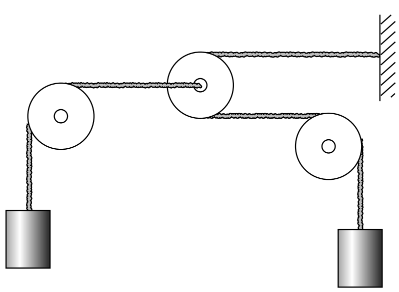
Problem 17.

17 What ratio of masses will balance the pulley system shown in the figure? ▷ Hint, p. 703



Problem 18.

18 (a) For the apparatus shown in the figure, find the equilibrium angle θ in terms of the two masses.
(b) Interpret your result in the case of $M \gg m$ (M much greater than m). Does it make sense physically?
(c) For what combinations of masses would your result give nonsense? Interpret this physically. ▷ Hint, p. 703



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 g only. ▷ Hint, p. 703

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

$$U(r) = k\left[\left(\frac{a}{r}\right)^{12} - 2\left(\frac{a}{r}\right)^{6}\right],$$

where r is the center-to-center distance between the atoms. Show

that (a) there is an equilibrium point at $r = a$,
(b) the equilibrium is stable, and
(c) the energy required to bring the atoms from their equilibrium separation to infinity is k. ▷ Hint, p. 703

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?

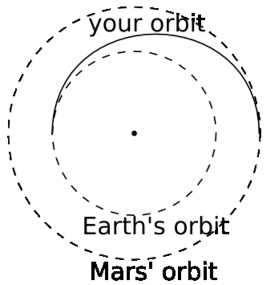
22 (a) A geosynchronous orbit is one in which the satellite orbits above the equator, and has an orbital period of 24 hours, so that it is always above the same point on the spinning earth. Calculate the altitude of such a satellite.
(b) What is the gravitational field experienced by the satellite? Give your answer as a percentage in relation to the gravitational field at the earth's surface. ▷ Hint, p. 703

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, G.

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?

25 A projectile is moving directly away from a planet of mass M at exactly escape velocity. Find r, the distance from the projectile to the center of the planet, as a function of time, t, and also find $v(t)$. Does v show the correct behavior as t approaches infinity? ▷ Hint, p. 703

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 U/m. That is, a thimbleful of water, m, 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 $U/m = -GM_e/r$, and a surface of constant U/m would be a surface of constant r, 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. ▷ Hint, p. 703



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? (Assume the orbits of Earth and Mars are circular.) ✓

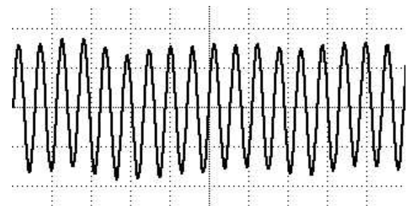
28 When you buy a helium-filled balloon, the seller has to inflate it from a large metal cylinder of the compressed gas. The helium inside the cylinder has energy, as can be demonstrated for example by releasing a little of it into the air: you hear a hissing sound, and that sound energy must have come 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 Energy is consumed in melting and evaporation. Explain in terms of conservation of energy why sweating cools your body, even though the sweat is at the same temperature as your body.

30 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 oven doesn't heat the solid very quickly, although eventually melting begins in one small spot. Once a melted spot forms, it grows rapidly, while the more distant solid parts remain solid. In other words, it appears based on this experiment 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.

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 ice at the polar icecaps to flood all the world's coastal cities. The total sunlight that falls on the ice caps amounts to about 1×10^{16} watts. Presently, 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 3×10^3 J.

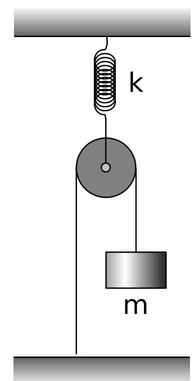


Problem 32.

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 T, the period, f, the frequency, and ω, the angular frequency.

33 (a) A mass m is hung from a spring whose spring constant is k. Write down an expression for the total interaction energy of the system, U, and find its equilibrium position. ▷ Hint, p. 703
(b) Explain how you could use your result from part a to determine an unknown spring constant.

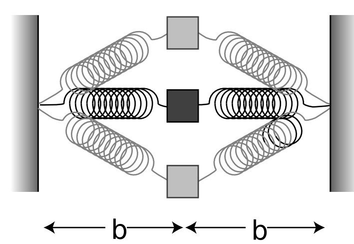
34 A certain mass, when hung from a certain spring, causes the spring to stretch by an amount h compared to its equilibrium length. If the mass is displaced vertically from this equilibrium, it will oscillate up and down with a period T_{osc}. Give a numerical comparison between T_{osc} and T_{fall}, the time required for the mass to fall from rest through a height h, when it isn't attached to the spring. (You will need the result of problem 33).



Problem 35.

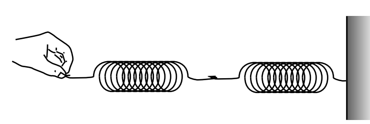
35 Find the period of vertical oscillations of the mass m. The spring, pulley, and ropes have negligible mass. ▷ Hint, p. 703

36 The equilibrium length of each spring in the figure is b, so when the mass m 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 k.
(a) Find the energy, U, stored in the springs, as a function of y, the distance of the mass up or down from the center.
(b) Show that the frequency of small up-down oscillations is infinite.
▷ Answer, p. 711



Problem 36.

37 Two springs with spring constants k_1 and k_2 are put together end-to-end. Let x_1 be the amount by which the first spring is stretched relative to its equilibrium length, and similarly for x_2. If the combined double spring is stretched by an amount b relative to *its* equilibrium length, then $b = x_1 + x_2$. Find the spring constant, K, of the combined spring in terms of k_1 and k_2.
▷ Hint, p. 704 ▷ Answer, p. 711



Problem 37.

38 A mass m on a spring oscillates around an equilibrium at $x = 0$. If the energy is symmetric with respect to positive and negative values of x, then the next level of improvement beyond

$U(x) = (1/2)kx^2$ would be $U(x) = (1/2)kx^2 + bx^4$. Do a numerical simulation with an energy that behaves in this way. 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 x^4 term tend to increase or decrease the period?

39 An idealized pendulum consists of a pointlike mass m on the end of a massless, rigid rod of length L. Its amplitude, θ, 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 m, L, and θ. Examine the effect of changing each variable while manipulating the others.

Key to symbols:
easy typical challenging difficult very difficult
√ An answer check is available at www.lightandmatter.com.

Exercises

Exercise 2A: Reasoning with Ratios and Powers

Equipment:

ping-pong balls and paddles

two-meter sticks

You have probably bounced a ping pong ball straight up and down in the air. The time between hits is related to the height to which you hit the ball. If you take twice as much time between hits, how many times higher do you think you will have to hit the ball? Write down your hypothesis:_____________

Your instructor will first beat out a tempo of 240 beats per minute (four beats per second), which you should try to match with the ping-pong ball. Measure the height to which the ball rises:_______

Now try it at 120 beats per minute:________

Compare your hypothesis and your results with the rest of the class.

Exercise 2B: The Shell Theorem

This exercise is an approximate numerical test of the shell theorem. There are seven masses A-G, each being one kilogram. Masses A-E, each one meter from the center, form a shape like two Egyptian pyramids joined at their bases; this is a rough approximation to a six-kilogram spherical shell of mass. Mass G is five meters from the center of the main group. The class will divide into six groups and split up the work required in order to calculate the vector sum of the six gravitational forces exerted on mass G. Depending on the size of the class, more than one group may be assigned to deal with the contribution of the same mass to the total force, and the redundant groups can check each other's results.

1. Discuss as a class what can be done to simplify the task of calculating the vector sum, and how to organize things so that each group can work in parallel with the others.

2. Each group should write its results on the board in units of piconewtons, retaining six significant figures of precision.

3. The class will determine the vector sum and compare with the result that would be obtained with the shell theorem.

Chapter 3
Conservation of Momentum

I think, therefore I am.

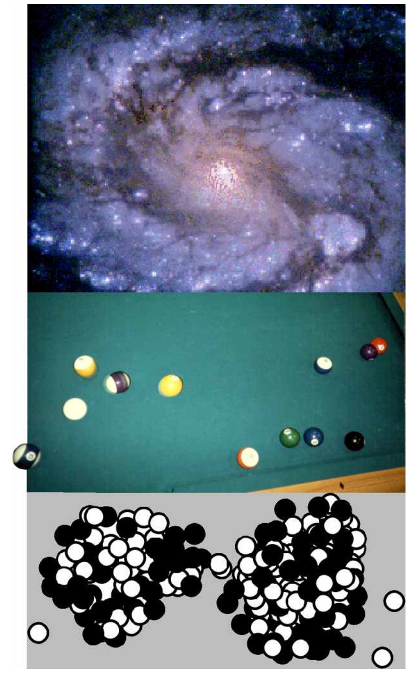
I hope that posterity will judge me kindly, not only as to the things which I have explained, but also to those which I have intentionally omitted so as to leave to others the pleasure of discovery.

René Descartes

3.1 Momentum in One Dimension

3.1.1 Mechanical momentum

In the martial arts movie *Crouching Tiger, Hidden Dragon*, those who had received mystical enlightenment are able to violate the laws of physics. Some of the violations are obvious, such as their ability to fly, but others are a little more subtle. The rebellious young heroine/antiheroine Jen Yu gets into an argument while sitting at a table in a restaurant. A young tough, Iron Arm Lu, comes running toward her at full speed, and she puts up one arm and effortlessly makes him bounce back, without even getting out of her seat or bracing herself against anything. She does all this between bites.



a / Systems consisting of material particles that interact through an energy $U(r)$. *Top:* The galaxy M100. Here the "particles" are stars. *Middle:* The pool balls don't interact until they come together and become compressed; the energy $U(r)$ has a sharp upturn when the center-to-center distance r gets small enough for the balls to be in contact. *Bottom:* A uranium nucleus undergoing fission. The energy $U(r)$ has a repulsive contribution from the electrical interactions of the protons, plus an attractive one due to the strong nuclear interaction. (*M100: Hubble Space Telescope image.*)

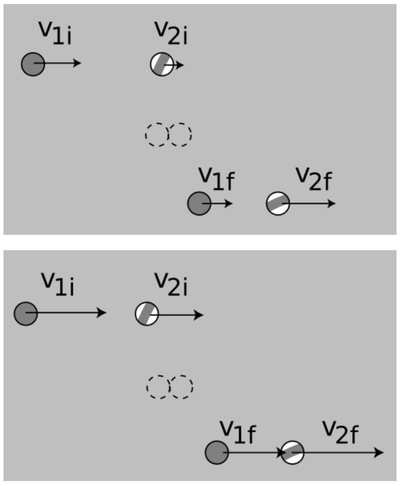
Although kinetic energy doesn't depend on the direction of motion, we've already seen on page 50 how conservation of energy combined with Galilean relativity allows us to make some predictions about the direction of motion. One of the examples was a demonstration that it isn't possible for a hockey puck to spontaneously reverse its direction of motion. In the scene from the movie, however, the woman's assailant isn't just gliding through space. He's interacting with her, so the previous argument doesn't apply here, and we need to generalize it to more than one object. We consider the case of a physical system composed of pointlike material particles, in which every particle interacts with every other particle through an energy $U(r)$ that depends only on the distance r between them. This still allows for a fairly general *mechanical system*, by which I mean roughly a system made of matter, not light. The characters in the movie are made of protons, neutrons, and electrons, so they would constitute such a system if the interactions among all these particles were of the form $U(r)$.[1] We might even be able to get

[1] Electrical and magnetic interactions *don't* quite behave like this, which is a point we'll take up later in the book.

away with thinking of each person as one big particle, if it's a good approximation to say that every part of each person's whole body moves in the same direction at the same speed.

The basic insight can be extracted from the special case where there are only two particles interacting, and they only move in one dimension, as in the example shown in figure b. Conservation of energy says

$$K_{1i} + K_{2i} + U_i = K_{1f} + K_{2f} + U_f \qquad .$$



b / A collision between two pool balls is seen in two different frames of reference. The solid ball catches up with the striped ball. Velocities are shown with arrows. The second observer is moving to the left at velocity u compared to the first observer, so all the velocities in the second frame have u added onto them. The two observers must agree on conservation of energy.

For simplicity, let's assume that the interactions start after the time we're calling initial, and end before the instant we choose as final. This is true in figure b, for example. Then $U_i = U_f$, and we can subtract the interaction energies from both sides, giving.

$$\begin{aligned} K_{1i} + K_{2i} &= K_{1f} + K_{2f} \\ \frac{1}{2}m_1 v_{1i}^2 + \frac{1}{2}m_2 v_{2i}^2 &= \frac{1}{2}m_1 v_{1f}^2 + \frac{1}{2}m_2 v_{2f}^2 \qquad . \end{aligned}$$

As in the one-particle argument on page 50, the trick is to require conservation of energy not just in one particular frame of reference, but in every frame of reference. In a frame of reference moving at velocity u relative to the first one, the velocities all have u added onto them:[2]

$$\frac{1}{2}m_1(v_{1i}+u)^2 + \frac{1}{2}m_2(v_{2i}+u)^2 = \frac{1}{2}m_1(v_{1f}+u)^2 + \frac{1}{2}m_2(v_{2f}+u)^2$$

When we square a quantity like $(v_{1i}+u)^2$, we get the same v_{1i}^2 that occurred in the original frame of reference, plus two u-dependent terms, $2v_{1i}u + u^2$. Subtracting the original conservation of energy equation from the version in the new frame of reference, we have

$$m_1 v_{1i} u + m_2 v_{2i} u = m_1 v_{1f} u + m_2 v_{2f} u \qquad ,$$

or, dividing by u,

$$m_1 v_{1i} + m_2 v_{2i} = m_1 v_{1f} + m_2 v_{2f} \qquad .$$

This is a statement that when you add up mv for the whole system, that total remains constant over time. In other words, this is a conservation law. The quantity mv is called *momentum*, notated p for obscure historical reasons. Its units are $\mathrm{kg}\cdot\mathrm{m/s}$.

Unlike kinetic energy, momentum depends on the direction of motion, since the velocity is not squared. In one dimension, motion in the same direction as the positive x axis is represented with positive values of v and p. Motion in the opposite direction has negative v and p.

[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.

Jen Yu meets Iron Arm Lu *example 1*

▷ Initially, Jen Yu is at rest, and Iron Arm Lu is charging to the left, toward her, at 5 m/s. Jen Yu's mass is 50 kg, and Lu's is 100 kg. After the collision, the movie shows Jen Yu still at rest, and Lu rebounding at 5 m/s to the right. Is this consistent with the laws of physics, or would it be impossible in real life?

▷ This is perfectly consistent with conservation of mass (50 kg+100 kg=50 kg+100 kg), and also with conservation of energy, since neither person's kinetic energy changes, and there is therefore no change in the total energy. (We don't have to worry about interaction energies, because the two points in time we're considering are ones at which the two people aren't interacting.) To analyze whether the scene violates conservation of momentum, we have to pick a coordinate system. Let's define positive as being to the right. The initial momentum is (50 kg)(0 m/s)+(100 kg)($-$5 m/s)=$-$500 kg $\cdot$ m/s, and the final momentum is (50 kg)(0 m/s)+(100 kg)(5 m/s)=500 kg $\cdot$ m/s. This is a change of 1000 kg $\cdot$ m/s, which is impossible if the two people constitute a closed system.

One could argue that they're not a closed system, since Lu might be exchanging momentum with the floor, and Jen Yu might be exchanging momentum with the seat of her chair. This is a reasonable objection, but in the following section we'll see that there are physical reasons why, in this situation, the force of friction would be relatively weak, and would not be able to transfer that much momentum in a fraction of a second.

This example points to an intuitive interpretation of conservation of momentum, which is that interactions are always mutual. That is, Jen Yu can't change Lu's momentum without having her own momentum changed as well.

A cannon *example 2*

▷ A cannon of mass 1000 kg fires a 10-kg shell at a velocity of 200 m/s. At what speed does the cannon recoil?

▷ The law of conservation of momentum tells us that

$$p_{cannon,i} + p_{shell,i} = p_{cannon,f} + p_{shell,f} \quad .$$

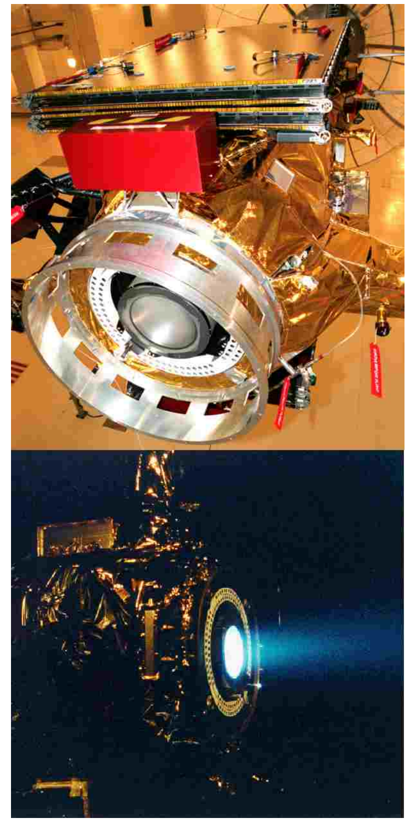
Choosing a coordinate system in which the cannon points in the positive direction, the given information is

$$\begin{aligned} p_{cannon,i} &= 0 \\ p_s hell, i &= 0 \\ p_{shell,f} &= 2000 \text{ kg} \cdot \text{m/s} \quad . \end{aligned}$$

We must have $p_{cannon,f} = -2000$ kg $\cdot$ m/s, so the recoil velocity of the cannon is 2 m/s.

Ion drive *example 3*

▷ The experimental solar-powered ion drive of the Deep Space 1 space probe expels its xenon gas exhaust at a speed of 30,000 m/s, ten times faster than the exhaust velocity for a typical chemical-fuel rocket engine. Roughly how many times greater is the maximum speed this spacecraft can reach, compared with a chemical-fueled probe with the same mass of fuel ("reaction mass") available for pushing out the back as exhaust?



c / The ion drive engine of the NASA Deep Space 1 probe, shown under construction (top) and being tested in a vacuum chamber (bottom) prior to its October 1998 launch. Intended mainly as a test vehicle for new technologies, the craft nevertheless also carried out a scientific program that included a rendezvous with a comet in 2004.(*NASA*)

▷ Momentum equals mass multiplied by velocity. Both spacecraft are assumed to have the same amount of reaction mass, and the ion drive's exhaust has a velocity ten times greater, so the momentum of its exhaust is ten times greater. Before the engine starts firing, neither the probe nor the exhaust has any momentum, so the total momentum of the system is zero. By conservation of momentum, the total momentum must also be zero after all the exhaust has been expelled. If we define the positive direction as the direction the spacecraft is going, then the negative momentum of the exhaust is canceled by the positive momentum of the spacecraft. The ion drive allows a final speed that is ten times greater. (This simplified analysis ignores the fact that the reaction mass expelled later in the burn is not moving backward as fast, because of the forward speed of the already-moving spacecraft.)

3.1.2 Nonmechanical momentum

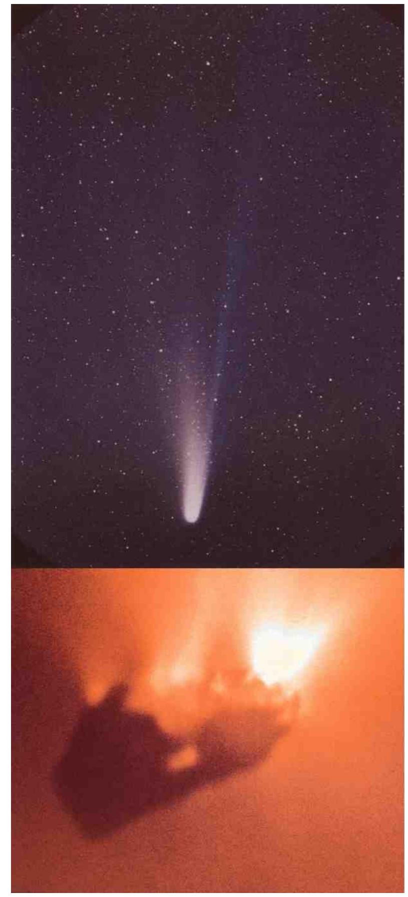
So far, it sounds as though conservation of momentum can be proved mathematically, unlike conservation of mass and energy, which are entirely based on observations. The proof, however, was only for a mechanical system, with interactions of the form $U(r)$. Conservation of momentum can be extended to other systems as well, but this generalization is based on experiments, not mathematical proof. Light is the most important example of momentum that doesn't equal mv — light doesn't have mass at all, but it does have momentum. For example, a flashlight left on for an hour would absorb about $10^{-5}\ \mathrm{kg \cdot m/s}$ of momentum as it recoiled.

Halley's comet *example 4*

Momentum is not always equal to mv. Halley's comet, shown in figure d, has a very elongated elliptical orbit, like those of many other comets. About once per century, its orbit brings it close to the sun. The comet's head, or nucleus, is composed of dirty ice, so the energy deposited by the intense sunlight gradually removes ice from the surface and turns it into water vapor. The bottom photo shows a view of the water coming off of the nucleus from the European Giotto space probe, which passed within 596 km of the comet's head on March 13, 1986.

The sunlight does not just carry energy, however. It also carries momentum. Once the steam comes off, the momentum of the sunlight impacting on it pushes it away from the sun, forming a tail as shown in in the top image. The tail always points away from the sun, so when the comet is receding from the sun, the tail is in front. By analogy with matter, for which momentum equals mv, you would expect that massless light would have zero momentum, but the equation $p = mv$ is not the correct one for light, and light does have momentum. (Some comets also have a second tail, which is propelled by electrical forces rather than by the momentum of sunlight.)

The reason for bringing this up is not so that you can plug numbers into formulas in these exotic situations. The point is that the conservation laws have proven so sturdy exactly because they can easily be amended to fit new circumstances. The momentum of light will be a natural consequence of the discussion of the theory of relativity in chapter 6.



d / Halley's comet. Top: A photograph made from earth. Bottom: A view of the nucleus from the Giotto space probe. (*W. Liller and European Space Agency*)

3.1.3 Momentum compared to kinetic energy

Momentum and kinetic energy are both measures of the quantity of motion, and a sideshow in the Newton-Leibniz controversy over who invented calculus was an argument over whether mv (i.e. momentum) or mv^2 (i.e. kinetic energy without the 1/2 in front) was the "true" measure of motion. The modern student can certainly be excused for wondering why we need both quantities, when their complementary nature was not evident to the greatest minds of the 1700s. The following table highlights their differences.

Kinetic energy...	Momentum...
doesn't depend on direction.	depends on direction.
is always positive, and cannot cancel out.	cancels with momentum in the opposite direction.
can be traded for forms of energy that do not involve motion. Kinetic energy is not a conserved quantity by itself.	is always conserved in a closed system.
is quadrupled if the velocity is doubled.	is doubled if the velocity is doubled.

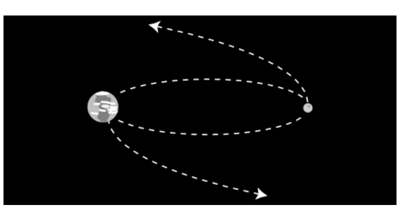
Here are some examples that show the different behaviors of the two quantities.

A spinning top *example 5*

A spinning top has zero total momentum, because for every moving point, there is another point on the opposite side that cancels its momentum. It does, however, have kinetic energy.

Momentum and kinetic energy in firing a rifle *example 6*

The rifle and bullet have zero momentum and zero kinetic energy to start with. When the trigger is pulled, the bullet gains some momentum in the forward direction, but this is canceled by the rifle's backward momentum, so the total momentum is still zero. The kinetic energies of the gun and bullet are both positive numbers, however, and do not cancel. The total kinetic energy is allowed to increase, because kinetic energy is being traded for other forms of energy. Initially there is chemical energy in the gunpowder. This chemical energy is converted into heat, sound, and kinetic energy. The gun's "backward" kinetic energy does not refrigerate the shooter's shoulder!



e / Example 8.

The wobbly earth *example 7*

As the moon completes half a circle around the earth, its motion reverses direction. This does not involve any change in kinetic energy. The reversed velocity does, however, imply a reversed momentum, so conservation of momentum in the closed earth-moon system tells us that the earth must also reverse its momentum. In fact, the earth wobbles in a little "orbit" about a point below its surface on the line connecting it and the moon. The two bodies' momenta always point in opposite directions and cancel each other out.

The earth and moon get a divorce *example 8*

Why can't the moon suddenly decide to fly off one way and the earth the other way? It is not forbidden by conservation of momentum, because the moon's newly acquired momentum in one direction could be canceled out by the change in the momentum of the earth, supposing the earth headed off in the opposite direction at the appropriate, slower speed. The catastrophe is forbidden by conservation of energy, because their energies would have to increase greatly.

Momentum and kinetic energy of a glacier *example 9*

A cubic-kilometer glacier would have a mass of about 10^{12} kg. If it moves at a speed of 10^{-5} m/s, then its momentum is $10^7\ \text{kg}\cdot\text{m/s}$. This is the kind of heroic-scale result we expect, perhaps the equivalent of the space shuttle taking off, or all the cars in LA driving in the same direction at freeway speed. Its kinetic energy, however, is only 50 J, the equivalent of the calories contained in a poppy seed or the energy in a drop of gasoline too small to be seen without a microscope. The surprisingly small kinetic energy is because kinetic energy is proportional to the square of the velocity, and the square of a small number is an even smaller number.

Discussion Questions

A If all the air molecules in the room settled down in a thin film on the floor, would that violate conservation of momentum as well as conservation of energy?

B A refrigerator has coils in back that get hot, and heat is molecular motion. These moving molecules have both energy and momentum. Why doesn't the refrigerator need to be tied to the wall to keep it from recoiling from the momentum it loses out the back?

f / This Hubble Space Telescope photo shows a small galaxy (yellow blob in the lower right) that has collided with a larger galaxy (spiral near the center), producing a wave of star formation (blue track) due to the shock waves passing through the galaxies' clouds of gas. This is considered a collision in the physics sense, even though it is statistically certain that no star in either galaxy ever struck a star in the other — the stars are very small compared to the distances between them.(*NASA*)

3.1.4 Collisions in one dimension

Physicists employ the term "collision" in a broader sense than in ordinary usage, applying it to any situation where objects interact for a certain period of time. A bat hitting a baseball, a cosmic ray damaging DNA, and a gun and a bullet going their separate ways are all examples of collisions in this sense. Physical contact is not even required. A comet swinging past the sun on a hyperbolic orbit is considered to undergo a collision, even though it never touches the sun. All that matters is that the comet and the sun interacted gravitationally with each other.

The reason for broadening the term "collision" in this way is that all of these situations can be attacked mathematically using the same conservation laws in similar ways. In our first example, conservation of momentum is all that is required.

Getting rear-ended *example 10*

▷ Ms. Chang is rear-ended at a stop light by Mr. Nelson, and sues to make him pay her medical bills. He testifies that he was only going 55 km per hour when he hit Ms. Chang. She thinks he was going much faster than that. The cars skidded together after the impact, and measurements of the length of the skid marks and the coefficient of friction show that their joint velocity immediately after the impact was 30 km per hour. Mr. Nelson's Nissan has a mass of 1400 kg, and Ms. Chang 's Cadillac is 2400 kg. Is Mr. Nelson telling the truth?

▷ Since the cars skidded together, we can write down the equation for conservation of momentum using only two velocities, v for Mr. Nelson's velocity before the crash, and v' for their joint velocity afterward:

$$m_N v = m_N v' + m_C v' \qquad .$$

Solving for the unknown, v, we find

$$\begin{aligned} v &= \left(1 + \frac{m_C}{m_N}\right) v' \\ &= 80\ \text{km/hr} \qquad . \end{aligned}$$

He is lying.

The above example was simple because both cars had the same velocity afterward. In many one-dimensional collisions, however, the two objects do not stick. If we wish to predict the result of such a collision, conservation of momentum does not suffice, because both velocities after the collision are unknown, so we have one equation in two unknowns.

Conservation of energy can provide a second equation, but its application is not as straightforward, because kinetic energy is only the particular form of energy that has to do with motion. In many collisions, part of the kinetic energy that was present before the collision is used to create heat or sound, or to break the objects or permanently bend them. Cars, in fact, are carefully designed to crumple in a collision. Crumpling the car uses up energy, and that's good because the goal is to get rid of all that kinetic energy in a relatively safe and controlled way. At the opposite extreme, a superball is "super" because it emerges from a collision with almost all its original kinetic energy, having only stored it briefly as interatomic electrical energy while it was being squashed by the impact.

Collisions of the superball type, in which almost no kinetic energy is converted to other forms of energy, can thus be analyzed more thoroughly, because they have $K_f = K_i$, as opposed to the less useful inequality $K_f < K_i$ for a case like a tennis ball bouncing on grass.

Pool balls colliding head-on — *example 11*

▷ Two pool balls collide head-on, so that the collision is restricted to one dimension. Pool balls are constructed so as to lose as little kinetic energy as possible in a collision, so under the assumption that no kinetic energy is converted to any other form of energy, what can we predict about the results of such a collision?

▷ Pool balls have identical masses, so we use the same symbol m for both. Conservation of energy and no loss of kinetic energy give us the two equations

$$mv_{1i} + mv_{2i} = mv_{1f} + mv_{2f}$$
$$\frac{1}{2}mv_{1i}^2 + \frac{1}{2}mv_{2i}^2 = \frac{1}{2}mv_{1f}^2 + \frac{1}{2}mv_{2f}^2$$

The masses and the factors of 1/2 can be divided out, giving

$$v_{1i} + v_{2i} = v_{1f} + v_{2f}$$
$$v_{1i}^2 + v_{2i}^2 = v_{1f}^2 + v_{2f}^2 \quad .$$

A little experimentation with numbers shows that given values of v_{1i} and v_{2i}, it is impossible to find v_{1f} and v_{1f} that satisfy these equations unless the final velocities equal the initial ones, or the final velocities are the same as the initial ones but swapped around. (An algebraic proof is not difficult, but I won't bother here.) In the special case where ball 2 is initially at rest, this tells us that ball 1 is stopped dead by the collision, and ball 2 heads off at the velocity originally possessed by ball 1. Pool players are familiar with this behavior.

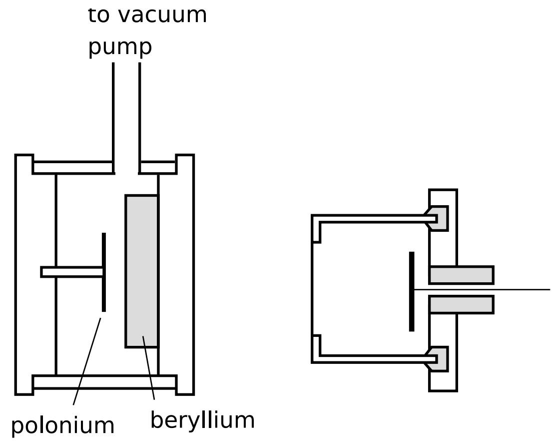
Often, as in the example above, the details of the algebra are the least interesting part of the problem, and considerable physical insight can be gained simply by counting the number of unknowns and comparing to the number of equations. Suppose a beginner at pool notices a case where her cue ball hits an initially stationary ball and stops dead. "Wow, what a good trick," she thinks. "I bet I could never do that again in a million years." But she tries again, and finds that she can't help doing it even if she doesn't want to. Luckily she has just learned about collisions in her physics course. Once she has written down the equations for conservation of energy and no loss of kinetic energy, she really doesn't have to complete the algebra. She knows that she has two equations in two unknowns, so there must be a well-defined solution. Once she has seen the result of one such collision, she knows that the same thing must happen every time. The same thing would happen with colliding marbles or croquet balls. It doesn't matter if the masses or velocities are different, because that just multiplies both equations by some constant factor.

The discovery of the neutron

This was the type of reasoning employed by James Chadwick in his 1932 discovery of the neutron. At the time, the atom was imagined to be made out of two types of fundamental particles, protons and electrons. The protons were far more massive, and clustered together in the atom's core, or nucleus. Electrical attraction caused the electrons to orbit the nucleus in circles, in much the same way that gravity kept the planets from cruising out of the solar system. Experiments showed, for example, that twice as much energy was required to strip the last electron off of a helium atom as was needed to remove the single electron from a hydrogen atom, and this was explained by saying that helium had two protons to hydrogen's one. The trouble was that according to this model, helium would have two electrons and two protons, giving it precisely twice the mass of a hydrogen atom with one of each. In fact, helium has about four times the mass of hydrogen.

Chadwick suspected that the helium nucleus possessed two additional particles of a new type, which did not participate in electrical interactions at all, i.e. were electrically neutral. If these particles had very nearly the same mass as protons, then the four-to-one mass ratio of helium and hydrogen could be explained. In 1930, a new type of radiation was discovered that seemed to fit this description. It was electrically neutral, and seemed to be coming from the nuclei of light elements that had been exposed to other types of radiation. At this time, however, reports of new types of particles were a dime a dozen, and most of them turned out to be either clusters made of previously known particles or else previously known particles with higher energies. Many physicists believed that the "new" particle that had attracted Chadwick's interest was really a previously known particle called a gamma ray, which was electrically neutral. Since gamma rays have no mass, Chadwick decided to try to determine the new particle's mass and see if it was nonzero and approximately equal to the mass of a proton.

Unfortunately a subatomic particle is not something you can just put on a scale and weigh. Chadwick came up with an ingenious solution. The masses of the nuclei of the various chemical elements



g / Chadwick's subatomic pool table. A disk of the naturally occurring metal polonium provides a source of radiation capable of kicking neutrons out of the beryllium nuclei. The type of radiation emitted by the polonium is easily absorbed by a few mm of air, so the air has to be pumped out of the left-hand chamber. The neutrons, Chadwick's mystery particles, penetrate matter far more readily, and fly out through the wall and into the chamber on the right, which is filled with nitrogen or hydrogen gas. When a neutron collides with a nitrogen or hydrogen nucleus, it kicks it out of its atom at high speed, and this recoiling nucleus then rips apart thousands of other atoms of the gas. The result is an electrical pulse that can be detected in the wire on the right. Physicists had already calibrated this type of apparatus so that they could translate the strength of the electrical pulse into the velocity of the recoiling nucleus. The whole apparatus shown in the figure would fit in the palm of your hand, in dramatic contrast to today's giant particle accelerators.

were already known, and techniques had already been developed for measuring the speed of a rapidly moving nucleus. He therefore set out to bombard samples of selected elements with the mysterious new particles. When a direct, head-on collision occurred between a mystery particle and the nucleus of one of the target atoms, the nucleus would be knocked out of the atom, and he would measure its velocity.

Suppose, for instance, that we bombard a sample of hydrogen atoms with the mystery particles. Since the participants in the collision are fundamental particles, there is no way for kinetic energy to be converted into heat or any other form of energy, and Chadwick thus had two equations in three unknowns:

equation #1: conservation of momentum

equation #2: no loss of kinetic energy

unknown #1: mass of the mystery particle

unknown #2: initial velocity of the mystery particle

unknown #3: final velocity of the mystery particle

The number of unknowns is greater than the number of equations, so there is no unique solution. But by creating collisions with nuclei of another element, nitrogen, he gained two more equations at the expense of only one more unknown:

equation #3: conservation of momentum in the new collision

equation #4: no loss of kinetic energy in the new collision

unknown #4: final velocity of the mystery particle in the new collision

He was thus able to solve for all the unknowns, including the mass of the mystery particle, which was indeed within 1% of the mass of a proton. He named the new particle the neutron, since it is electrically neutral.

Discussion Questions

A Good pool players learn to make the cue ball spin, which can cause it not to stop dead in a head-on collision with a stationary ball. If this does not violate the laws of physics, what hidden assumption was there in the example in the text where it was proved that the cue ball must stop?

3.1.5 The center of mass

Figures h and j show two examples where a motion that appears complicated actually has a very simple feature. In both cases, there is a particular point, called the center of mass, whose motion is surprisingly simple. The highjumper flexes his body as he passes over the bar, so his motion is intrinsically very complicated, and yet his center of mass's motion is a simple parabola, just like the parabolic arc of a pointlike particle. The wrench's center of mass travels in a straight line as seen from above, which is what we'd expect for a pointlike particle flying through the air.

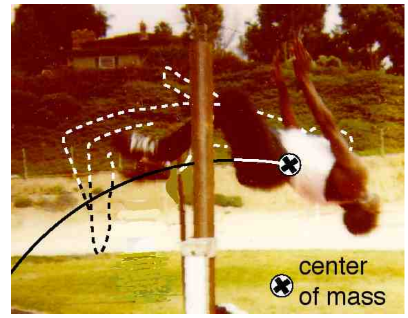
The highjumper and the wrench are both complicated systems, each consisting of zillions of subatomic particles. To understand what's going on, let's instead look at a nice simple system, two pool balls colliding. We assume the balls are a closed system (i.e. their interaction with the felt surface is not important) and that their rotation is unimportant, so that we'll be able to treat each one as a single particle. By symmetry, the only place their center of mass can be is half-way in between, at an x coordinate equal to the average of the two balls' positions, $x_{cm}=(x_1+x_2)/2$.

Figure i makes it appear that the center of mass, marked with an $\times$, moves with constant velocity to the right, regardless of the collision, and we can easily prove this using conservation of momentum:

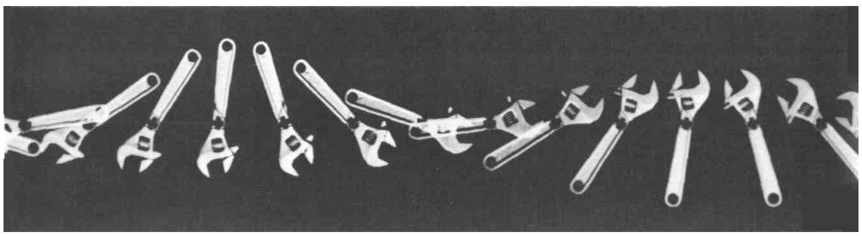
$$\begin{aligned}
v_{cm} &= \mathrm{d}x_{cm}/\mathrm{d}t \\
&= \frac{1}{2}(v_1+v_2) \\
&= \frac{1}{2m}(mv_1+mv_2) \\
&= \frac{p_{total}}{m_{total}}
\end{aligned}$$

Since momentum is conserved, the last expression is constant, which proves that v_{cm} is constant.

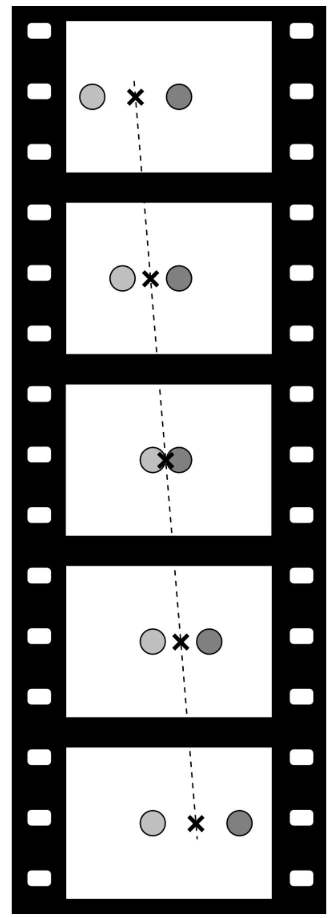
Rearranging this a little, we have $p_{total}=m_{total}v_{cm}$. In other



h / The highjumper's body passes over the bar, but his center of mass passes under it.(*Dunia Young*)



j / In this multiple-flash photograph, we see the wrench from above as it flies through the air, rotating as it goes. Its center of mass, marked with the black cross, travels along a straight line, unlike the other points on the wrench, which execute loops. (*PSSC Physics*)



i / Two pool balls collide.

words, the total momentum of the system is the same as if all its mass was concentrated at the center of mass point.

Sigma notation

When there is a large, potentially unknown number of particles, we can write sums like the ones occurring above using symbols like "$+\ldots$," but that gets awkward. It's more convenient to use the Greek uppercase sigma, Σ, to indicate addition. For example, the sum $1^2+2^2+3^2+4^2=30$ could be written as

$$\sum_{j=1}^{n} j^2 = 30 \qquad ,$$

read "the sum from $j=1$ to n of j^2." The variable j is a dummy variable, just like the dx in an integral that tells you you're integrating with respect to x, but has no significance outside the integral. The j below the sigma tells you what variable is changing from one term of the sum to the next, but j has no significance outside the sum.

As an example, let's generalize the proof of $p_{total}=m_{total}v_{cm}$ to the case of an arbitrary number n of identical particles moving in one dimension, rather than just two particles. The center of mass is at

$$x_{cm} = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad ,$$

where x_1 is the mass of the first particle, and so on. The velocity of the center of mass is

$$\begin{aligned} v_{cm} &= \mathrm{d}x_{cm}/\mathrm{d}t \\ &= \frac{1}{n}\sum_{j=1}^{n} v_j \\ &= \frac{1}{nm}\sum_{j=1}^{n} mv_j \\ &= \frac{p_{total}}{m_{total}} \end{aligned}$$

What about a system containing objects with unequal masses, or containing more than two objects? The reasoning above can be generalized to a weighted average:

$$x_{cm} = \frac{\sum_{j=1}^{n} m_j x_j}{\sum_{j=1}^{n} m_j}$$

The solar system's center of mass *example 12*

In the discussion of the sun's gravitational field on page 61, I mentioned in a footnote that the sun doesn't really stay in one place while the planets orbit around it. Actually, motion is relative, so it's meaningless to ask whether the sun is absolutely at rest, but it is meaningful to ask whether it moves in a straight line at constant velocity. We can now see that since the solar system is a closed system, its total momentum must be constant, and $p_{total} = m_{total}v_{cm}$ then tells us that it's the solar system's center of mass that has constant velocity, not the sun. The sun wobbles around this point irregularly due to its interactions with the planets, Jupiter in particular.

The earth-moon system *example 13*

The earth-moon system is much simpler than the solar system because it contains only two objects. Where is the center of mass of this system? Let x=0 be the earth's center, so that the moon lies at $x = 3.8 \times 10^5$ km. Then

$$\begin{aligned} x_{cm} &= \frac{\sum_{j=1}^{2} m_j x_j}{\sum_{j=1}^{2} m_j} \\ &= \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2} \quad , \end{aligned}$$

and letting 1 be the earth and 2 the moon, we have

$$\begin{aligned} x_{cm} &= \frac{m_{earth} \times 0 + m_{moon} x_{moon}}{m_{earth} + m_{moon}} \\ &= 4600 \text{ km} \quad , \end{aligned}$$

or about three quarters of the way from the earth's center to its surface.

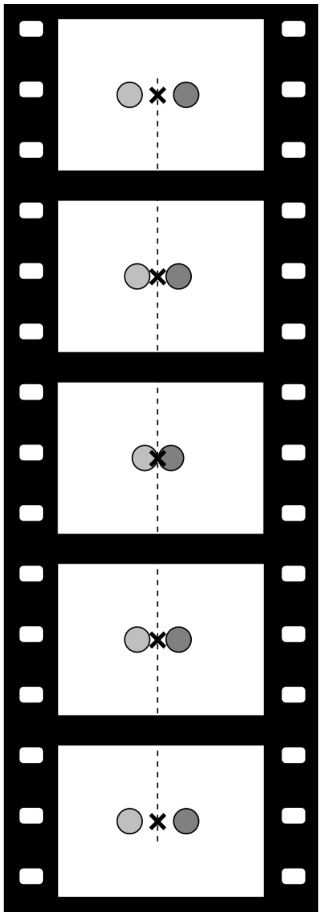
Momentum and Galilean relativity *example 14*

The principle of Galilean relativity states that the laws of physics are supposed to be equally valid in all inertial frames of reference. If we first calculate some momenta in one frame of reference and find that momentum is conserved, and then rework the whole problem in some other frame of reference that is moving with respect to the first, the numerical values of the momenta will all be different. Even so, momentum will still be conserved. All that matters is that we work a single problem in one consistent frame of reference.

One way of proving this is to apply the equation $p_{total} = m_{total}v_{cm}$. If the velocity of one frame relative to the other is u, then the only effect of changing frames of reference is to change v_{cm} from its original value to $v_{cm} + u$. This adds a constant onto the momentum, which has no effect on conservation of momentum.

3.1.6 The center of mass frame of reference

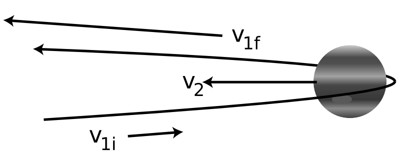
A particularly useful frame of reference in many cases is the frame that moves along with the center of mass, called the center of mass (c.m.) frame. In this frame, the total momentum is zero. The following examples show how the center of mass frame can be a powerful tool for simplifying our understanding of collisions.



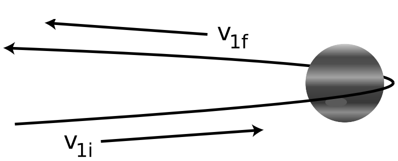
k / The same collision of two pools balls, but now seen in the center of mass frame of reference.

A collision of pool balls viewed in the c.m. frame — *example 15*

If you move your head so that your eye is always above the point halfway in between the two pool balls, as in figure k, you are viewing things in the center of mass frame. In this frame, the balls come toward the center of mass at equal speeds. By symmetry, they must therefore recoil at equal speeds along the lines on which they entered. Since the balls have essentially swapped paths in the center of mass frame, the same must also be true in any other frame. This is the same result that required laborious algebra to prove previously without the concept of the center of mass frame.



l / The sun's frame of reference.



m / The c.m. frame.

The slingshot effect — *example 16*

It is a counterintuitive fact that a spacecraft can pick up speed by swinging around a planet, if it arrives in the opposite direction compared to the planet's motion. Although there is no physical contact, we treat the encounter as a one-dimensional collision, and analyze it in the center of mass frame. Since Jupiter is so much more massive than the spacecraft, the center of mass is essentially fixed at Jupiter's center, and Jupiter has zero velocity in the center of mass frame, as shown in figure 3.1.6. The c.m. frame is moving to the left compared to the sun-fixed frame used in figure 3.1.6, so the spacecraft's initial velocity is greater in this frame than in the sun's frame.

Things are simpler in the center of mass frame, because it is more symmetric. In the sun-fixed frame, the incoming leg of the encounter is rapid, because the two bodies are rushing toward each other, while their separation on the outbound leg is more gradual, because Jupiter is trying to catch up. In the c.m. frame, Jupiter is sitting still, and there is perfect symmetry between the incoming and outgoing legs, so by symmetry we have $v_1 f = -v_{1i}$. Going back to the sun-fixed frame, the spacecraft's final velocity is increased by the frames' motion relative to each other. In the sun-fixed frame, the spacecraft's velocity has increased greatly.

Einstein's motorcycle — *example 17*

We've assumed we were dealing with a system of material objects, for which the equation $p = mv$ was true. What if our system contains only light rays, or a mixture of light and matter? As a college student, Einstein kept worrying about was what a beam of light would look like if you could ride alongside it on a motorcycle. In other words, he imagined putting himself in the light beam's center of mass frame. Chapter 6 discusses Einstein's resolution of this problem, but the basic point is that you *can't* ride the motorcycle alongside the light beam, because material objects can't go as fast as the speed of light. A beam of light has no center of mass frame of reference.

A Make up a numerical example of two unequal masses moving in one dimension at constant velocity, and verify the equation $p_{total} = m_{total} v_{cm}$ over a time interval of one second.

B A more massive tennis racquet or baseball bat makes the ball fly off faster. Explain why this is true, using the center of mass frame. For simplicity, assume that the racquet or bat is simply sitting still before the collision, and that the hitter's hands do not make any force large enough to have a significant effect over the short duration of the impact.

3.2 Force in One Dimension

3.2.1 Momentum transfer

For every conserved quantity, we can define an associated rate of flow. An open system can have mass transferred in or out of it, and we can measure the rate of mass flow, $\mathrm{d}m/\mathrm{d}t$ in units of kg/s. Energy can flow in or out, and the rate of energy transfer is the power, $P = \mathrm{d}E/\mathrm{d}t$, measured in watts.[3] The rate of *momentum* transfer is called force,

$$F = \frac{\mathrm{d}p}{\mathrm{d}t} \qquad \text{[definition of force]} \qquad .$$

The units of force are kg·m/s^2, which can be abbreviated as newtons, 1 N = kg·m/s^2. Newtons are unfortunately not as familiar as watts. A newton is about how much force you'd use to pet a dog. The most powerful rocket engine ever built, the first stage of the Saturn V that sent astronauts to the moon, had a thrust of about 30 million newtons. In one dimension, positive and negative signs indicate the direction of the force — a positive force is one that pushes or pulls in the direction of the positive x axis.

Walking into a lamppost — *example 18*

▷ Starting from rest, you begin walking, bringing your momentum up to 100 kg·m/s. You walk straight into a lamppost. Why is the momentum change of -100 kg·m/s so much more painful than the change of +100 kg·m/s when you started walking?

▷ The forces are not really constant, but for this type of qualitative discussion we can pretend they are, and approximate $\mathrm{d}p/\mathrm{d}t$ as $\Delta p/\Delta t$. It probably takes you about 1 s to speed up initially, so the ground's force on you is $F = \Delta p/\Delta t \approx 100$ N. Your impact with the lamppost, however, is over in the blink of an eye, say 1/10 s or less. Dividing by this much smaller Δt gives a much larger force, perhaps thousands of newtons (with a negative sign).

This is also the principle of airbags in cars. The time required for the airbag to decelerate your head is fairly long: the time it takes your face to travel 20 or 30 cm. Without an airbag, your face would have hit the dashboard, and the time interval would have been the much shorter time taken by your skull to move a couple of centimeters while your face compressed. Note that either way, the same amount of momentum is transferred: the entire momentum of your head.

Force is defined as a derivative, and the derivative of a sum is the sum of the derivatives. Therefore force is additive: when more than one force acts on an object, you add the forces to find out what happens. An important special case is that forces can cancel. Consider your body sitting in a chair as you read this book. Let the positive x axis be upward. The chair's upward force on you

[3] Recall that uppercase P is power, while lowercase p is momentum.

is represented with a positive number, which cancels out with the earth's downward gravitational force, which is negative. The total rate of momentum transfer into your body is zero, and your body doesn't change its momentum.

A Many collisions, like the collision of a bat with a baseball, appear to be instantaneous. Most people also would not imagine the bat and ball as bending or being compressed during the collision. Consider the following possibilities:

(1) The collision is instantaneous.
(2) The collision takes a finite amount of time, during which the ball and bat retain their shapes and remain in contact.
(3) The collision takes a finite amount of time, during which the ball and bat are bending or being compressed.

How can two of these be ruled out based on energy or momentum considerations?

3.2.2 Newton's laws

Although momentum is the third conserved quantity we've encountered, historically it was the first to be discovered. Isaac Newton formulated a complete treatment of mechanical systems in terms of force and momentum. Newton's theory was based on three laws of motion, which we now think of as consequences of conservation of mass, energy, and momentum.

Newton's laws in one dimension:
Newton's first law: If there is no force acting on an object, it stays in the same state of motion.
Newton's second law: $F = \mathrm{d}p/\mathrm{d}t$.
Newton's third law: Forces occur in opposite pairs. If object A interacts with object B, then A's force on B and B's force on A are related by $F_{AB} = -F_{BA}$.

The second law is the definition of force, which we've already encountered.[4] The first law is a special case of the second law[5] — if $\mathrm{d}p/\mathrm{d}t$ is zero, then $p = mv$ is a constant, and since mass is conserved, constant p implies constant v. The third law is a restatement of conservation of momentum: for two objects interacting, we have constant total momentum, so $0 = \frac{\mathrm{d}}{\mathrm{d}t}(p_A + p_B) = F_{BA} + F_{AB}$.

a=F/m *example 19*

Many modern textbooks restate Newton's second law as $a = F/m$, i.e. as an equation that predicts an object's acceleration based on the force exerted on it. This is easily derived from Newton's original form as follows: $a = \mathrm{d}v/\mathrm{d}t = (\mathrm{d}p/\mathrm{d}t)/m = F/m$.

Gravitational force related to g *example 20*

As a special case of the previous example, consider an object in free fall, and let the x axis point down, so that g is positive. Then a=g, and $F = ma = mg$. For example, the gravitational force on a 1 kg mass at the earth's surface is about 9.8 N. Even if other forces act on the object, and it isn't in free fall, the gravitational force on it is still the same, and can still be calculated as mg.

If you've already accepted Galilean relativity in your heart, then there is nothing really difficult about the first and second laws. The

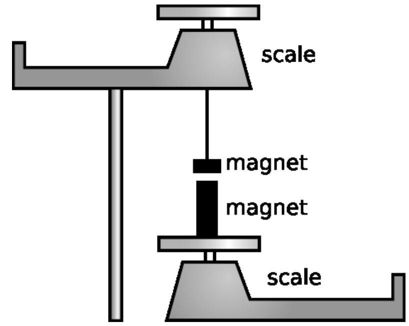
[4]This is with the benefit of hindsight. At the time, the word "force" already had certain connotations, and people thought they understood what it meant and how to measure it, e.g. by using a spring scale. From their point of view, $F = \mathrm{d}p/\mathrm{d}t$ was not a definition but a testable — and controversial! — statement.

[5]My own opinion is that Newton stated this special case as a separate law in order to emphasize his commitment to Galilean relativity, as opposed to the prevalent Aristotelian belief that a force was required to keep an object moving. Many modern textbooks, however, present the first law as a statement that inertial frames of reference exist. Newton wrote in Latin and didn't use modern algebra notation in his published work, so any modernized presentation of his work is necessarily not literal. Newton was a notoriously bad communicator (he lectured to an empty hall once a year to fulfill a requirement of his position at Cambridge), so the modern student should be thankful that textbook authors take so many liberties with Newton's laws!

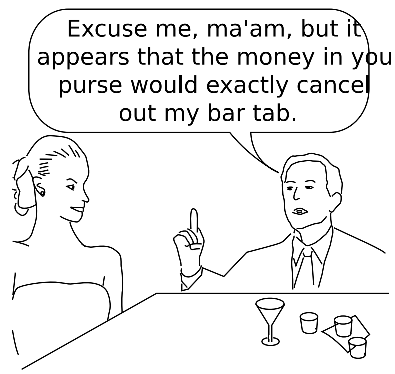
third law, however, is more of a conceptual challenge. The first hurdle is that it is counterintuitive. Is it really true that if a speeding cement truck hits an old lady who is crossing the street, the old lady's force on the cement truck is just as strong as the truck's force on her? Yes, it is true, but it is hard to believe at first. That amount of force simply has more of an effect on the old lady, because she is less massive and not as tough.

A more humane and practical experiment is shown in figure a. A large magnet and a small magnet are weighed separately, and then one magnet is hung from the pan of the top balance so that it is directly above the other magnet. There is an attraction between the two magnets, causing the reading on the top scale to increase and the reading on the bottom scale to decrease. The large magnet is more "powerful" in the sense that it can pick up a heavier paperclip from the same distance, so many people have a strong expectation that one scale's reading will change by a far different amount than the other. Instead, we find that the two changes are equal in magnitude but opposite in direction, so the upward force of the top magnet on the bottom magnet is of the same magnitude as the downward force of the bottom magnet on the top magnet.

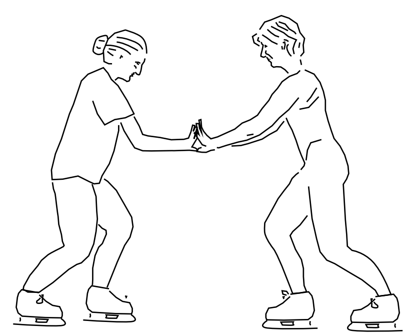
To students, it often sounds as though Newton's third law implies nothing could ever change its motion, since the two equal and opposite forces would always cancel. As illustrated in figure b, the fallacy arises from assuming that we can add things that it doesn't make sense to add. It only makes sense to add up forces that are acting on the same object, whereas two forces related to each other by Newton's third law are always acting on two different objects. If two objects are interacting via a force and no other forces are involved, then *both* objects will accelerate — in opposite directions, as shown in figure c!



a / Two magnets exert forces on each other.



b / It doesn't make sense for the man to talk about the woman's money canceling out his bar tab, because there is no good reason to combine his debts and her assets.



c / Newton's third law does not mean that forces always cancel out so that nothing can ever move. If these two ice skaters, initially at rest, push against each other, they will both move.

Discussion Questions

A Criticize the following incorrect statement:
"If an object is at rest and the total force on it is zero, it stays at rest. There can also be cases where an object is moving and keeps on moving without having any total force on it, but that can only happen when there's no friction, like in outer space."

B The table gives laser timing data for Ben Johnson's 100 m dash at the 1987 World Championship in Rome. (His world record was later revoked because he tested positive for steroids.) How does the total force on him change over the duration of the race?

x (m)	t (s)
10	1.84
20	2.86
30	3.80
40	4.67
50	5.53
60	6.38
70	7.23
80	8.10
90	8.96
100	9.83

C Criticize the following incorrect statement: "If you shove a book across a table, friction takes away more and more of its force, until finally it stops."

D You hit a tennis ball against a wall. Explain any and all incorrect ideas in the following description of the physics involved: "The ball gets some force from you when you hit it, and when it hits the wall, it loses part of that force, so it doesn't bounce back as fast. The muscles in your arm are the only things that a force can come from."

E When you fire a gun, the exploding gases push outward in all directions, causing the bullet to accelerate down the barrel. What Newton's-third-law pairs are involved? [Hint: Remember that the gases themselves are an object.]

F Tam Anh grabs Sarah by the hand and tries to pull her. She tries to remain standing without moving. A student analyzes the situation as follows. "If Tam Anh's force on Sarah is greater than her force on him, he can get her to move. Otherwise, she'll be able to stay where she is." What's wrong with this analysis?

G You hit a tennis ball against a wall. Explain any and all incorrect ideas in the following description of the physics involved: "According to Newton's third law, there has to be a force opposite to your force on the ball. The opposite force is the ball's mass, which resists acceleration, and also air resistance."

H The earth's gravitational force on you, i.e. your weight, is always equal to mg, where m is your mass. So why can you get a shovel to go deeper into the ground by jumping onto it? Just because you're jumping, that doesn't mean your mass or weight is any greater, does it?

3.2.3 Forces between solids

Conservation laws are more fundamental than Newton's laws, and they apply where Newton's laws don't, e.g. to light and to the internal structure of atoms. However, there are certain problems that are much easier to solve using Newton's laws. As a trivial example, if you drop a rock, it could conserve momentum and energy by levitating, or by falling in the usual manner. With Newton's laws, however, we can reason that $a = F/m$, so the rock must respond to the gravitational force by accelerating.

Less trivially, suppose a person is hanging onto a rope, and we want to know if she will slip. Unlike the case of the levitating rock, here the no-motion solution could be perfectly reasonable if her grip is strong enough. We know that her hand's interaction with the rope is fundamentally an electrical interaction between the atoms in the surface of her palm and the nearby atoms in the surface of the rope. For practical problem-solving, however, this is a case where we're better off forgetting the fundamental classification of interactions at the atomic level and working with a more practical, everyday classification of forces. In this practical scheme, we have three types of forces that can occur between solid objects in contact:

A normal force, F_n,	is perpendicular to the surface of contact, and prevents objects from passing through each other by becoming as strong as necessary (up to the point where the objects break). "Normal" means perpendicular.
Static friction, F_s,	is parallel to the surface of contact, and prevents the surfaces from starting to slip by becoming as strong as necessary, up to a maximum value of $F_{s,max}$. "Static" means not moving, i.e. not slipping.
Kinetic friction, F_k,	is parallel to the surface of contact, and tends to slow down any slippage once it starts. "Kinetic" means moving, i.e. slipping.

If you put a coin on this page, which is horizontal, gravity pulls down on the coin, but the atoms in the paper and the coin repel each other electrically, and the atoms are compressed until the repulsion becomes strong enough to stop the downward motion of the coin. We describe this complicated and invisible atomic process by saying that the paper makes an upward normal force on the coin, and the coin makes a downward normal force on the paper. (The two normal forces are related by Newton's third law. In fact, Newton's third law only relates forces that are of the same type.)

If you now tilt the book a little, static friction keeps the coin from slipping. The picture at the microscopic level is even more complicated than the previous description of the normal force. One model is to think of the tiny bumps and depressions in the coin as settling into the similar irregularities in the paper. This model predicts that rougher surfaces should have more friction, which is sometimes true but not always. Two very smooth, clean glass surfaces or very well finished machined metal surfaces may actually stick *better* than rougher surfaces would, the probable explanation being that there is some kind of chemical bonding going on, and the smoother surfaces allow more atoms to be in contact.

Finally, as you tilt the book more and more, there comes a point where static friction reaches its maximum value. The surfaces become unstuck, and the coin begins to slide over the paper. Kinetic friction slows down this slipping motion significantly. In terms of energy, kinetic friction is converting mechanical energy into heat, just like when you rub your hands together to keep warm. One model of kinetic friction is that the tiny irregularities in the two surfaces bump against each other, causing vibrations whose energy rapidly converts to heat and sound — you can hear this sound if you rub your fingers together near your ear.

For *dry* surfaces, experiments show that the following equations usually work fairly well:

$$F_{s,max} \approx \mu_s F_n \qquad ,$$

and

$$F_k \approx \mu_k F_n \qquad ,$$

where μ_s, the coefficient of static friction, and μ_k, the coefficient of kinetic friction, are constants that depend on the properties of the two surfaces, such as what they're made of and how rough they are.

Maximum acceleration of a car *example 21*

▷ Rubber on asphalt gives $\mu_k \approx 0.4$ and $\mu_s \approx 0.6$. What is the upper limit on a car's acceleration on a flat road, assuming that the engine has plenty of power and that air friction is negligible?

▷ The earth makes a downward gravitational force on the car whose absolute value is mg, and since the car doesn't accelerate vertically, the road apparently makes an upward normal force of the same magnitude, $F_n = mg$. As is always true, the coefficient of static friction is greater than the coefficient of kinetic friction, so the maximum acceleration is obtained with static friction, i.e. the driver should try not to burn rubber. The maximum force of static friction is $F_s, max = \mu_s F_n = \mu_s mg$. The maximum acceleration is $a = F_s/m = \mu_s g \approx 6\ \text{m/s}^2$. This is true regardless of how big the tires are, since the experimentally determined relationship $F_{s,max} = \mu_s F_n$ is independent of surface area.

Self-Check

Can a frictionless surface exert a normal force? Can a frictional force exist without a normal force? ▷ Answer, p. 706

3.2.4 Work

Energy transferred to a particle

To change the kinetic energy, $K=(1/2)mv^2$, of a particle moving in one dimension, we must change its velocity. That will entail a change in its momentum, $p=mv$, as well, and since force is the rate of transfer of momentum, we conclude that the only way to change a particle's kinetic energy is to apply a force.[6] A force in the same direction as the motion speeds it up, increasing the kinetic energy, while a force in the opposite direction slows it down.

Consider an infinitesimal time interval during which the particle moves an infinitesimal distance $\mathrm{d}x$, and its kinetic energy changes by $\mathrm{d}K$. In one dimension, we represent the direction of the force and the direction of the motion with positive and negative signs for F and $\mathrm{d}x$, so the relationship among the signs can be summarized as follows:

$F>0$	$\mathrm{d}x>0$	$\mathrm{d}K>0$
$F<0$	$\mathrm{d}x<0$	$\mathrm{d}K>0$
$F>0$	$\mathrm{d}x<0$	$\mathrm{d}K<0$
$F<0$	$\mathrm{d}x>0$	$\mathrm{d}K<0$

This looks exactly like the rule for determining the sign of a product, and we can easily show using the chain rule that this is indeed a multiplicative relationship:

$$\begin{aligned}
\mathrm{d}K &= \frac{\mathrm{d}K}{\mathrm{d}v}\frac{\mathrm{d}v}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}x}\mathrm{d}x \qquad \text{[chain rule]}\\
&= (mv)(a)(1/v)\mathrm{d}x\\
&= m\,a\,\mathrm{d}x\\
&= F\,\mathrm{d}x \qquad \text{[Newton's second law]}
\end{aligned}$$

We can verify that force multiplied by distance has units of energy:

$$\begin{aligned}
\mathrm{N{\cdot}m} &= \frac{\mathrm{kg{\cdot}m/s}}{\mathrm{s}}\times\mathrm{m}\\
&= \mathrm{kg{\cdot}m^2/s^2}\\
&= \mathrm{J}
\end{aligned}$$

A TV picture tube *example 22*

▷ At the back of a typical TV's picture tube, electrical forces accelerate each electron to an energy of 5×10^{-16} J over a distance of about 1 cm. How much force is applied to a single electron? (Assume the force is constant.) What is the corresponding acceleration?

[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.

▷ Integrating

$$dK = F dx \quad ,$$

we find

$$K_f - K_i = F(x_f - x_i)$$

or

$$\Delta K = F \Delta x \quad .$$

The force is

$$\begin{aligned} F &= \Delta K / \Delta x \\ &= \frac{5 \times 10^{-16}\ \text{J}}{0.01\ \text{m}} \\ &= 5 \times 10^{-14}\ \text{N} \quad . \end{aligned}$$

This may not sound like an impressive force, but it's enough to supply an electron with a spectacular acceleration. Looking up the mass of an electron on p. 720, we find

$$\begin{aligned} a &= F/m \\ &= 5 \times 10^{16}\ \text{m/s}^2 \quad . \end{aligned}$$

An air gun *example 23*

▷ An airgun, figure d, uses compressed air to accelerate a pellet. As the pellet moves from x_1 to x_2, the air decompresses, so the force is not constant. Using methods from chapter 5, one can show that the air's force on the pellet is given by $F = bx^{-7/5}$. A typical high-end airgun used for competitive target shooting has

$$\begin{aligned} x_1 &= 0.046\ \text{m} \quad , \\ x_2 &= 0.41\ \text{m} \quad , \end{aligned}$$

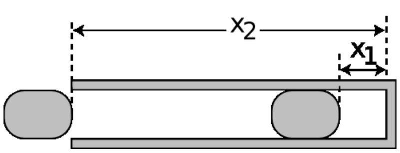
and

$$b = 4.4\ \text{N}\cdot\text{m}^{7/5} \quad .$$

What is the kinetic energy of the pellet when it leaves the muzzle? (Assume friction is negligible.)

▷ Since the force isn't constant, it would be incorrect to do $F = \Delta K/\Delta x$. Integrating both sides of the equation $dK = F dx$, we have

$$\begin{aligned} \Delta K &= \int_{x_1}^{x_2} F dx \\ &= -\frac{5b}{2}\left(x_2^{-2/5} - x_1^{-2/5}\right) \\ &= 22\ \text{J} \end{aligned}$$



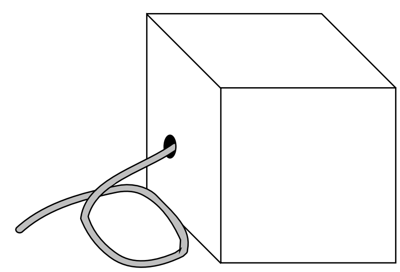
d / A simplified drawing of an airgun.

In general, when energy is transferred by a force,[7] we use the term *work* to refer to the amount of energy transferred. This is different from the way the word is used in ordinary speech. If you stand for a long time holding a bag of cement, you get tired, and everyone will agree that you've worked hard, but you haven't changed the energy of the cement, so according to the definition of the physics term, you haven't done any work on the bag. There has been an energy transformation inside your body, of chemical energy into heat, but this just means that one part of your body did positive work (lost energy) while another part did a corresponding amount of negative work (gained energy).

Work in general

I derived the expression $F\mathrm{d}x$ for one particular type of kinetic-energy transfer, the work done in accelerating a particle, and then defined work as a more general term. Is the equation correct for other types of work as well? For example, if a force lifts a mass m against the resistance of gravity at constant velocity, the increase in the mass's gravitational energy is $\mathrm{d}(mgy) = mg\mathrm{d}y = F\mathrm{d}y$, so again the equation works, but this still doesn't prove that the equation is *always* correct as a way of calculating energy transfers.

Imagine a black box[8], containing a gasoline-powered engine, which is designed to reel in a steel cable, exerting a certain force F. For simplicity, we imagine that this force is always constant, so we can talk about Δx rather than an infinitesimal $\mathrm{d}x$. If this black box is used to accelerate a particle (or any mass without internal structure), and no other forces act on the particle, then the original derivation applies, and the work done by the box is $W = F\Delta x$. Since F is constant, the box will run out of gas after reeling in a certain amount of cable Δx. The chemical energy inside the box has decreased by $-W$, while the mass being accelerated has gained W worth of kinetic energy.[9]



e / The black box does work by reeling in its cable.

Now what if we use the black box to pull a plow? The energy increase in the outside world is of a different type than before; it takes the forms of (1) the gravitational energy of the dirt that has been lifted out to the sides of the furrow, (2) frictional heating of the dirt and the plowshare, and (3) the energy needed to break up the dirt clods (a form of electrical energy involving the attractions among the atoms in the clod). The box, however, only communicates with the outside world via the hole through which its cable passes. The amount of chemical energy lost by the gasoline can therefore only

[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.

depend on F and Δx, so it is the same $-W$ as when the box was being used to accelerate a mass, and thus by conservation of energy, the work done on the outside world is again W.

This is starting to sound like a proof that the force-times-distance method is always correct, but there was one subtle assumption involved, which was that the force was exerted at one point (the end of the cable, in the black box example). Real life often isn't like that. For example, a cyclist exerts forces on both pedals at once. Serious cyclists use toe-clips, and the conventional wisdom is that one should use equal amounts of force on the upstroke and downstroke, to make full use of both sets of muscles. This is a two-dimensional example, since the pedals go in circles. We're only discussing one-dimensional motion right now, so let's just pretend that the upstroke and downstroke are both executed in straight lines. Since the forces are in opposite directions, one is positive and one is negative. The cyclist's *total* force on the crank set is zero, but the work done isn't zero. We have to add the work done by each stroke, $W = F_1\Delta x_1 + F_2\Delta x_2$. (I'm pretending that both forces are constant, so we don't have to do integrals.) Both terms are positive; one is a positive number multiplied by a positive number, while the other is a negative times a negative.

f / The wheel spinning in the air has $K_{cm} = 0$. The space shuttle has all its kinetic energy in the form of center of mass motion, $K = K_{cm}$. The rolling ball has some, but not all, of its energy in the form of center of mass motion, $K_{cm} < K$.(*Space Shuttle photo by NASA*)

This might not seem like a big deal — just remember not to use the total force — but there are many situations where the total force is all we can measure. The ultimate example is heat conduction. Heat conduction is not supposed to be counted as a form of work, since it occurs without a force. But at the atomic level, there are forces, and work is done by one atom on another. When you hold a hot potato in your hand, the transfer of heat energy through your skin takes place with a total force that's extremely close to zero. At the atomic level, atoms in your skin are interacting electrically with atoms in the potato, but the attractions and repulsions add up to zero total force. It's just like the cyclist's feet acting on the pedals, but with zillions of forces involved instead of two. There is no practical way to measure all the individual forces, and therefore we can't calculate the total energy transferred.

To summarize, $\sum F_j \mathrm{d}x_j$ is a correct way of calculating work, where F_j is the individual force acting on particle j, which moves a distance $\mathrm{d}x_j$. However, this is only useful if you can identify all the individual forces and determine the distance moved at each point of contact. For convenience, I'll refer to this as the *work theorem*. (It doesn't have a standard name.)

There is, however, something useful we can do with the total force. We can use it to calculate the part of the work done on an object that consists of a change in the kinetic energy it has due to the motion of its center of mass. The proof is essentially the same as the proof on 113, except that now we don't assume the force is acting on a single particle, so we have to be a little more delicate. Let the object consist of n particles. Its total kinetic energy is $K=\sum_{j=1}^{n}(1/2)m_jv_j^2$, but this is what we've already realized *can't* be calculated using the total force. The kinetic energy it has due to motion of its center of mass is

$$K_{cm}=\frac{1}{2}m_{total}v_{cm}^2 \qquad .$$

Figure f shows some examples of the distinction between K_{cm} and K. Differentiating K_{cm}, we have

$$\begin{aligned}
\mathrm{d}K_{cm} &= m_{total}v_{cm}\mathrm{d}v_{cm} \\
&= m_{total}v_{cm}\frac{\mathrm{d}v_{cm}}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}x_{cm}}\mathrm{d}x_{cm} \qquad \text{[chain rule]} \\
&= m_{total}\frac{\mathrm{d}v_{cm}}{\mathrm{d}t}\mathrm{d}x_{cm} \qquad [\mathrm{d}t/\mathrm{d}x_{cm}=1/v_{cm}] \\
&= \frac{\mathrm{d}p_{total}}{\mathrm{d}t}\mathrm{d}x_{cm} \qquad [p_{total}=m_{total}v_{cm}] \\
&= F_{total}\mathrm{d}x_{cm}
\end{aligned}$$

I'll call this the *kinetic energy theorem* — like the work theorem, it has no standard name.

An ice skater pushing off from a wall *example 24*

The kinetic energy theorem tells us how to calculate the skater's kinetic energy if we know the amount of force and the distance her center of mass travels while she is pushing off.

The work theorem tells us that the wall does no work on the skater, since the point of contact isn't moving. This makes sense, because the wall does not have any source of energy.

Absorbing an impact without recoiling? *example 25*

▷ Is it possible to absorb an impact without recoiling? For instance, if a ping-pong ball hits a brick wall, does the wall "give" at all?

▷ There will always be a recoil. In the example proposed, the wall will surely have some energy transferred to it in the form of heat and vibration. The work theorem tells us that we can only have an energy transfer if the distance traveled by the point of contact is nonzero.

Dragging a refrigerator at constant velocity *example 26*

The fridge's momentum is constant, so there is no net momentum transfer, and the total force on it must be zero: your force is canceling the floor's kinetic frictional force. The kinetic energy theorem is therefore true but useless. It tells us that there is zero total force on the refrigerator, and that the refrigerator's kinetic energy doesn't change.

The work theorem tells us that the work you do equals your hand's force on the refrigerator multiplied by the distance traveled. Since we know the floor has no source of energy, the only way for the floor and refrigerator to gain energy is from the work you do. We can thus calculate the total heat dissipated by friction in the refrigerator and the floor.

Note that there is no way to find how much of the heat is dissipated in the floor and how much in the refrigerator.

Accelerating a cart *example 27*

If you push on a cart and accelerate it, there are two forces acting on the cart: your hand's force, and the static frictional force of the ground pushing on the wheels in the opposite direction.

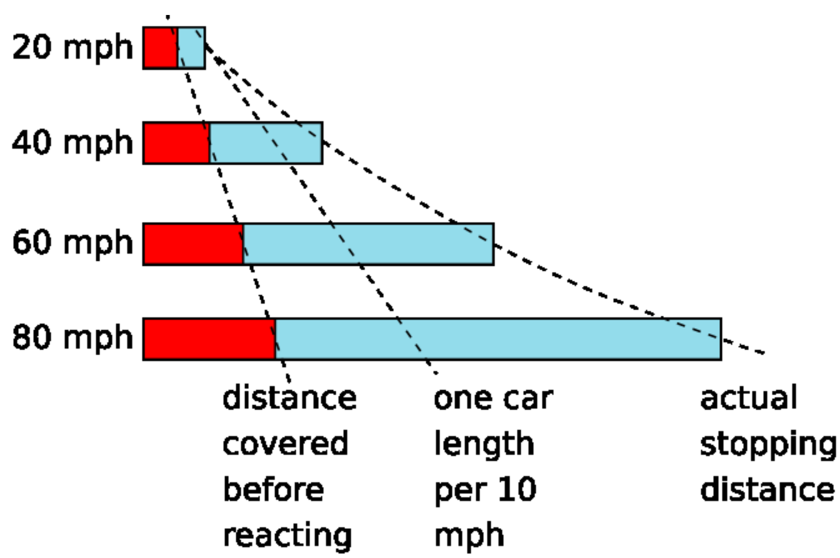
Applying the work theorem to your force tells us how to calculate the work you do.

Applying the work theorem to the floor's force tells us that the floor does no work on the cart. There is no motion at the point of contact, because the atoms in the floor are not moving. (The atoms in the surface of the wheel are also momentarily at rest when they touch the floor.) This makes sense, because the floor does not have any source of energy.

The kinetic energy theorem refers to the total force, and because the floor's backward force cancels part of your force, the total force is less than your force. This tells us that only part of your work goes into the kinetic energy associated with the forward motion of the cart's center of mass. The rest goes into rotation of the wheels.

Discussion Questions

A Criticize the following incorrect statement: "A force doesn't do any work unless it's causing the object to move."

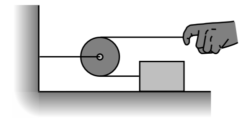


g / Discussion question B.

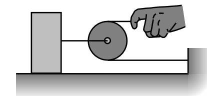
B To stop your car, you must first have time to react, and then it takes some time for the car to slow down. Both of these times contribute to the distance you will travel before you can stop. The figure shows how the average stopping distance increases with speed. Because the stopping distance increases more and more rapidly as you go faster, the rule of one car length per 10 m.p.h. of speed is not conservative enough at high speeds. In terms of work and kinetic energy, what is the reason for the more rapid increase at high speeds?

3.2.5 Simple machines

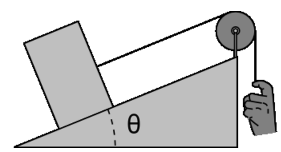
Conservation of energy provided the necessary tools for analyzing some mechanical systems, such as the seesaw on page 46 and the pulley arrangements of the homework problems on page 82, but we could only analyze those machines by computing the total energy of the system. That approach wouldn't work for systems like the biceps/forearm machine on page 46, or the one in figure h, where the energy content of the person's body is impossible to compute directly. Even though the seesaw and the biceps/forearm system were clearly just two different forms of the lever, we had no way to treat them both on the same footing. We can now successfully attack such problems using the work and kinetic energy theorems.



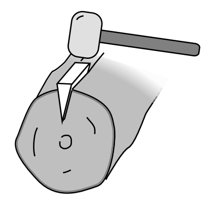
h / The force is transmitted to the block.



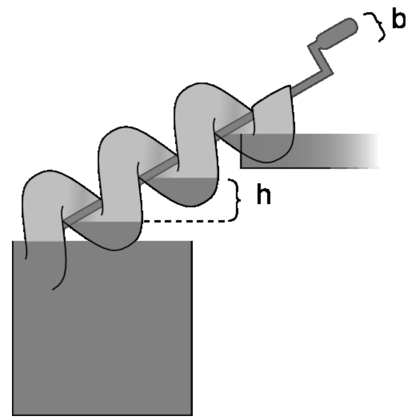
i / A mechanical advantage of 2.



j / An inclined plane.



k / A wedge.



l / Archimedes' screw

Constant tension around a pulley — *example 28*

▷ In figure h, what is the relationship between the force applied by the person's hand and the force exerted on the block?

▷ If we assume the rope and the pulley are ideal, i.e. frictionless and massless, then there is no way for them to absorb or release energy, so the work done by the hand must be the same as the work done on the block. Since the hand and the block move the same distance, the work theorem tells us the two forces are the same.

Similar arguments show that an idealized rope exerts the same force anywhere it's attached to something, and the same amount of force is also exerted by each segment of the rope on the neighboring segments. This amount of force is called the *tension* in the rope. Going around an ideal pulley has no effect on the tension.

This is an example of a simple machine, which is any mechanical system that manipulates forces to do work. This particular machine reverses the direction of the motion, but doesn't change the force or the speed of motion.

A mechanical advantage — *example 29*

The idealized pulley in figure i has negligible mass, so its kinetic energy is zero, and the kinetic energy theorem tells us that the total force on it is zero. We know, as in the preceding example, that the two forces pulling it to the right are equal to each other, so the force on the left must be twice as strong. This simple machine doubles the applied force, and we refer to this ratio as a *mechanical advantage* (M.A.) of 2. There's no such thing as a free lunch, however; the distance traveled by the load is cut in half, and there is no increase in the amount of work done.

Inclined plane and wedge — *example 30*

In figure j, the force applied by the hand is equal to the one applied to the load, but there is a mechanical advantage compared to the force that would have been required to lift the load straight up. The distance traveled up the inclined plane is greater by a factor of $1/\sin\theta$, so by the work theorem, the force is smaller by a factor of $\sin\theta$, and we have M.A.$=1/\sin\theta$. The wedge, k, is similar.

Archimedes' screw — *example 31*

In one revolution, the crank travels a distance $2\pi b$, and the water rises by a height h. The mechanical advantage is $2\pi b/h$.

3.2.6 Force related to interaction energy

In section 2.3, we saw that there were two equivalent ways of looking at gravity, the gravitational field and the gravitational energy. They were related by the equation $\mathrm{d}U = mg\mathrm{d}r$, so if we knew the field, we could find the energy by integration, $U = \int mg\mathrm{d}r$, and if we knew the energy, we could find the field by differentiation, $g = (1/m)\mathrm{d}U/\mathrm{d}r$.

The same approach can be applied to other interactions, for example a mass on a spring. The main difference is that only in gravitational interactions does the strength of the interaction depend on the mass of the object, so in general, it doesn't make sense to separate out the factor of m as in the equation $\mathrm{d}U = mg\mathrm{d}r$. Since $F = mg$ is the gravitational force, we can rewrite the equation in the more suggestive form $\mathrm{d}U = F\mathrm{d}r$. This form no longer refers to gravity specifically, and can be applied much more generally. The only remaining detail is that I've been fairly cavalier about positive and negative signs up until now. That wasn't such a big problem for gravitational interactions, since gravity is always attractive, but it requires more careful treatment for nongravitational forces, where we don't necessarily know the direction of the force in advance, and we need to use positive and negative signs carefully for the direction of the force.

In general, suppose that forces are acting on a particle — we can think of them as coming from other objects that are "off stage" — and that the interaction between the particle and the off-stage objects can be characterized by an interaction energy, U, which depends only on the particle's position, x. Using the kinetic energy theorem, we have $\mathrm{d}K = F\mathrm{d}x$. (It's not necessary to write K_{cm}, since a particle can't have any other kind of kinetic energy.) Conservation of energy tells us $\mathrm{d}K + \mathrm{d}U = 0$, so the relationship between force and interaction energy is $\mathrm{d}U = -F\mathrm{d}x$, or

$$F = -\frac{\mathrm{d}U}{\mathrm{d}x} \qquad \text{[relationship between force and interaction energy]} \quad .$$

Force exerted by a spring — *example 32*

▷ A mass is attached to the end of a spring, and the energy of the spring is $U = (1/2)kx^2$, where x is the position of the mass, and $x = 0$ is defined to be the equilibrium position. What is the force the spring exerts on the mass? Interpret the sign of the result.

▷ Differentiating, we find

$$\begin{aligned} F &= -\frac{\mathrm{d}U}{\mathrm{d}x} \\ &= -kx \quad . \end{aligned}$$

If x is positive, then the force is negative, i.e. it acts so as to bring the mass back to equilibrium, and similarly for $x < 0$ we have $F > 0$.

Most books do the $F = -kx$ form before the $U = (1/2)kx^2$ form, and call it Hooke's law. Neither form is really more fundamental than the other — we can always get from one to the other by integrating or differentiating.

Newton's law of gravity *example 33*

▷ Given the equation $U = -Gm_1m_2/r$ for the energy of gravitational interactions, find the corresponding equation for the gravitational force on mass m_2. Interpret the positive and negative signs.

▷ We have to be a little careful here, because we've been taking r to be positive by definition, whereas the position, x, of mass m_2 could be positive or negative, depending on which side of m_1 it's on.

For positive x, we have $r = x$, and differentiation gives

$$\begin{aligned} F &= -\frac{dU}{dx} \\ &= -Gm_1m_2/x^2 \qquad . \end{aligned}$$

As in the preceding example, we have $F < 0$ when x is positive, because the object is being attracted back toward $x = 0$.

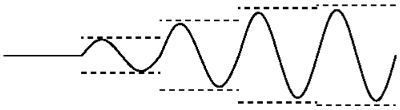
When x is negative, the relationship between r and x becomes $r = -x$, and the result for the force is the same as before, but with a minus sign. We can combine the two equations by writing

$$|F| = Gm_1m_2/r^2 \qquad ,$$

and this is the form traditionally known as Newton's law of gravity. As in the preceding example, the U and F equations contain equivalent information, and neither is more fundamental than the other.

Equilibrium *example 34*

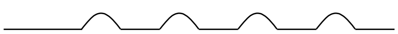
I previously described the condition for equilibrium as a local maximum or minimum of U. A differentiable function has a zero derivative at its extrema, and we can now relate this directly to force: zero force acts on an object when it is at equilibrium.



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

3.3 Resonance

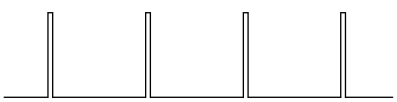
Resonance is a phenomenon in which an oscillator responds most strongly to a driving force that matches its own natural frequency of vibration. For example, suppose a child is on a playground swing with a natural frequency of 1 Hz. That is, if you pull the child away from equilibrium, release her, and 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, figure a, and then give pushes at 1 Hz, the swing's amplitude rapidly builds up, 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.



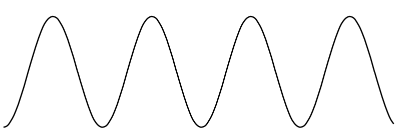
b / A swing pushed at twice its resonant frequency.

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.

Making the math easy



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



d / A sinusoidal driving force.

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. 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.

A second mathematical issue that I glossed over in the swing example was how friction behaves. In section 3.2.3, about forces between solids, the empirical equation for kinetic friction was independent of velocity. If the only type of friction operating on the playground swing was one that behaved in this way, the consequences for the child would be unfortunate: the amplitude, rather than approaching a limiting value as suggested in figure a, would grow without bound, and she would find herself falling out of the swing as it flew up higher than the bar! The main source of friction on the playground swing is air friction, which increases with velocity. In practical machines, moving parts are normally lubricated, and friction at a lubricated surface is not just weaker than dry friction but also, unlike dry friction, displays velocity dependence. How exactly does friction work when liquids and gases are involved? We could imagine that as the child on the swing moves through the air, her body would experience continual collisions with air molecules. In these collisions she would tend to transfer momentum to the air, and force is the rate of momentum transfer. The number of collisions per second would be proportional to her velocity, and we would therefore expect air friction to be proportional to her velocity,

$$F = -bv \qquad ,$$

where the minus sign is because the frictional force opposes her motion. In reality, experiments show that friction involving gases and liquids only behaves according to this equation at extremely low velocities, or for a gas that has a very low density. At more ordinary velocities, the relationship is not a straight proportionality, because turbulent eddies are stirred up. Nevertheless, we'll assume throughout the rest of this section that $F = -bv$ is true, because it ends up giving mathematically simple results!

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. 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 32 for the force exerted by the spring. This becomes a little prettier if we rewrite it in the form

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

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
```

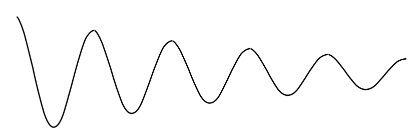
```
turnaround at t= 2.99 , x= 0.728777519477
turnaround at t= 3.99 , x= 0.655817260033
turnaround at t= 4.99 , x= 0.590154191135
turnaround at t= 5.99 , x= 0.531059189965
turnaround at t= 6.99 , x= 0.477875914756
turnaround at t= 7.99 , x= 0.430013546991
turnaround at t= 8.99 , x= 0.386940256644
turnaround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

The spring constant, $k = 4\pi = 39.4784$, 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 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 $b = 0.211$ N·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.

Analytic treatment

Taking advantage of this unexpectedly simple result, let's find an analytic solution for the motion. The numerical output suggests that we assume a solution of the form



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

$$x = Ae^{-ct}\sin(\omega_f t + \delta) \qquad .$$

where the unknown constants ω_f and c will presumably be related to m, b, and k. The constant c indicates how quickly the oscillations die out. The constant ω_f is, as before, defined as 2π times the frequency, with the subscript f to indicate a free (undriven) solution. All our equations will come out much simpler if we use ωs everywhere instead of fs from now on, and, as physicists often do, I'll generally use the word "frequency" to refer to ω when the context makes it clear what I'm talking about. The phase angle δ has no real physical significance, since we can define $t = 0$ to be any moment in time we like. The factor A for the initial amplitude

[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.

can also be omitted without loss of generality, since the equation we're trying to solve, $ma + bv + kx = 0$ is linear. That is, v and a are the first and second derivatives of x, and the derivative of Ax is simply A times the derivative of x. Thus, if $x(t)$ is a solution of the equation, then multiplying it by a constant gives an equally valid solution. For the purpose of determining ω_f and c, the most general form we need to consider is therefore $x = e^{-ct} \sin \omega_f t$, whose first and second derivatives are $v = e^{-ct} \left(-c \sin \omega_f t + \omega \cos \omega_f t\right)$ and $a = e^{-ct} \left(c^2 \sin \omega_f t - 2\omega_f c \cos \omega_f t - \omega_f^2 \sin \omega_f t\right)$. Plugging these into the equation $ma + bv + kx = 0$ and setting the sine and cosine parts equal to zero gives, after some tedious algebra,

$$\omega_f = \sqrt{\frac{k}{m} - \frac{b^2}{4m^2}}$$

and

$$c = \frac{b}{2m} \qquad .$$

The first of these two equations is like the undamped equation $\omega = \sqrt{k/m}$, except for the second term, which is often negligible for small b. The second equation says that c, which indicates how quickly the oscillations damp out, is directly related to b, the strength of the damping.

3.3.2 The quality factor

It's usually impractical to measure b directly and determine c from the equation $c = b/2m$. For a child on a swing, measuring b would require putting the child in a wind tunnel! It's usually much easier to characterize the amount of damping by observing the actual damped oscillations and seeing how many cycles it takes for the mechanical energy to decrease by a certain factor. The unitless *quality factor*, Q, is defined as $Q = \omega_o/2c$, and in the limit of weak damping, where $\omega \approx \omega_o$, this can be interpreted as the number of cycles required for the mechanical energy to fall off by a factor of $e^{2\pi} = 535.49\ldots$. Using this new quantity, we can rewrite the equation for the frequency of damped oscillations in the slightly more elegant form $\omega_f = \omega_o\sqrt{1 - 1/4Q^2}$.

Exponential decay in a trumpet — *example 35*

▷ The vibrations of the air column inside a trumpet have a Q of about 10. This means that even after the trumpet player stops blowing, the note will keep sounding for a short time. If the player suddenly stops blowing, how will the sound intensity 20 cycles later compare with the sound intensity while she was still blowing?

▷ The trumpet's Q is 10, so after 10 cycles the energy will have fallen off by a factor of 535. After another 10 cycles we lose another factor of 535, so the sound intensity is reduced by a factor of $535 \times 535 = 2.9 \times 10^5$.

The decay of a musical sound is part of what gives it its character, and a good musical instrument should have the right Q, but the Q that is considered desirable is different for different instruments. A guitar is meant to keep on sounding for a long time after a string has been plucked, and might have a Q of 1000 or 10000. One of the reasons why a cheap synthesizer sounds so bad is that the sound suddenly cuts off after a key is released.

3.3.3 Driven motion

The driven case is both simpler and more interesting than the undriven one. We have an external driving force $F = F_m \sin \omega t$, where the constant F_m indicates the maximum strength of the force in either direction. The equation of motion is now

[1] $ma + bv + kx = F_m \sin \omega t$

[equation of motion for a driven oscillator] .

After the driving force has been applied for a while, we expect that the amplitude of the oscillations will approach some constant value. This motion is known as the *steady state*, and it's the most interesting thing to find out; as we'll see later, the most general type of motion is only a minor variation on the steady-state motion. For the steady-state motion, we're going to look for a solution of the form

$$x = A \sin(\omega t + \delta) \qquad .$$

In contrast to the undriven case, here it's not possible to sweep A and δ under the rug. The amplitude of the steady-state motion, A, is actually the most interesting thing to know about the steady-state motion, and it's not true that we still have a solution no matter how we fiddle with A; if we have a solution for a certain value of A, then multiplying A by some constant would break the equality between the two sides of the equation of motion. It's also no longer true that we can get rid of δ simply be redefining when we start the clock; here δ represents a *difference* in time between the start of one cycle of the driving force and the start of the corresponding cycle of the motion.

The velocity and acceleration are $v = \omega A \sin(\omega t + \delta)$ and $a = -\omega^2 A \cos(\omega t + \delta)$, and if we plug these into the equation of motion, [1], and simplify a little, we find

[2] $$(k - m\omega^2)\sin(\omega t + \delta) + \omega b \cos(\omega t + \delta) = \frac{F_m}{A} \sin \omega t \qquad .$$

The sum of any two sinusoidal functions with the same frequency is also a sinusoidal, so the whole left side adds up to a sinusoidal. By fiddling with A and δ we can make the amplitudes and phases of the two sides of the equation match up.

Steady state, no damping

A and δ are easy to find in the case where there is no damping at all. There are now no cosines in equation [2] above, only sines, so if we wish we can set δ to zero, and we find $A = F_m/(k - m\omega^2) = F_m/m(\omega_o^2 - \omega^2)$. This, however, makes A negative for $\omega > \omega_o$. The variable δ was designed to represent this kind of phase relationship, so we prefer to keep A positive and set $\delta = \pi$ for $\omega > \omega_o$. Our results are then

$$A = \frac{F_m}{m\,|\omega^2 - \omega_o^2|}$$

and

$$\delta = \begin{cases} 0, & \omega < \omega_o \\ \pi, & \omega > \omega_o \end{cases} \quad .$$

The most important feature of the result is that there is a resonance: the amplitude becomes greater and greater, and approaches infinity, as ω approaches the resonant frequency ω_o. The interpretation of the infinite amplitude is that there really isn't any steady state if we drive the system exactly at resonance — the amplitude will just keep on increasing indefinitely. In real life, there is always some damping, and there will always be some difference, however small, between ω and ω_o.

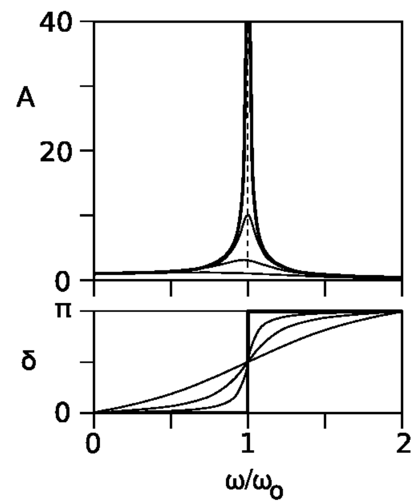
There is a simple interpretation for the surprising behavior of the phase angle δ. The system's mechanical energy can only change due to work done by the driving force, since there is no damping to convert mechanical energy to heat. In the steady state, then, the power transmitted by the driving force over a full cycle of motion must average out to zero. In general, the work theorem $\mathrm{d}E = F\mathrm{d}x$ can always be divided by $\mathrm{d}t$ on both sides to give the useful relation $P = Fv$. If Fv is to average out to zero, then F and v must be out of phase by $\pm\pi/2$, and since v is ahead of x by a phase angle of $\pi/2$, the phase angle between x and F must be zero or π.

A practice mute for a violin *example 36*

The amplitude of the driven vibrations, $A = F_m/(m|\omega^2 - \omega_o^2|)$, contains an inverse proportionality to the mass of the vibrating object. This is simply because a given force will produce less acceleration when applied to a more massive object. An application is shown in figure 36.

In a stringed instrument, the strings themselves don't have enough surface area to excite sound waves very efficiently. In instruments of the violin family, as the strings vibrate from left to right, they cause the bridge (the piece of wood they pass over) to wiggle clockwise and counterclockwise, and this motion is transmitted to the top panel of the instrument, which vibrates and creates sound waves in the air.

A string player who wants to practice at night without bothering the neighbors can add some mass to the bridge. Adding mass to the bridge causes the amplitude of the vibrations to be smaller, and the sound to be much softer. A similar effect is seen when an electric guitar is used without an amp. The body of an electric guitar is so much more massive than the body of an acoustic guitar that the amplitude of its vibrations is very small.

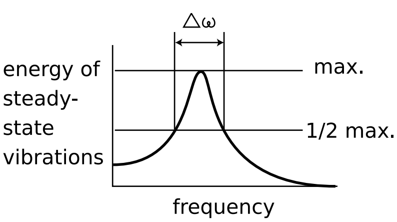


g / Dependence of the amplitude and phase angle on the driving frequency. The undamped case, $Q = \infty$, is shown with heavy lines, and the other curves represent Q=1, 3, and 10. The amplitudes were calculated with F_m, m, and ω_o, all set to 1.

Steady state, with damping

The extension of the analysis to the damped case involves some lengthy algebra, which I've outlined on page 693 in appendix 2. The results are shown in figure g. It's not surprising that the steady state response is weaker when there is more damping, since the steady state is reached when the power extracted by damping matches the

f / Example 36: a viola without a mute (left), and with a mute (right). The mute doesn't touch the strings themselves.



h / The definition of $\Delta\omega$, the full width at half maximum.

power input by the driving force. What is surprising is that the amplitude is strongly affected by damping close to resonance, but only weakly affected far from it. In other words, the shape of the resonance curve is broader with more damping, and even if we were to scale up a high-damping curve so that its maximum was the same as that of a low-damping curve, it would still have a different shape. The standard way of describing the shape numerically is to give the quantity $\Delta\omega$, called the *full width at half-maximum*, or FWHM, which is defined in figure h. Note that the y axis is energy, which is proportional to the square of the amplitude. Our previous observations amount to a statement that $\Delta\omega$ is greater when the damping is stronger, i.e. when the Q is lower. It's not hard to show from the equations on page 693 that for large Q, the FWHM is given approximately by

$$\Delta\omega \approx \omega_o/Q \qquad .$$

(It's clear that this can't be a good approximation for small values of Q, since for very small Q the resonance curve doesn't even have a maximum near $\omega = \omega_o$.)

An opera singer breaking a wineglass — *example 37*

In order to break a wineglass by singing, an opera singer must first tap the glass to find its natural frequency of vibration, and then sing the same note back, so that her driving force will produce a response with the greatest possible amplitude. If she's shopping for the right glass to use for this display of her prowess, she should look for one that has the greatest possible Q, since the resonance curve has a higher maximum for higher values of Q.

Collapse of the Nimitz Freeway *example 38*

Figure i shows a section of the Nimitz Freeway in Oakland, CA, that collapsed during an earthquake in 1989. An earthquake consists of many low-frequency vibrations that occur simultaneously, which is why it sounds like a rumble of indeterminate pitch rather than a low hum. The frequencies that we can hear are not even the strongest ones; most of the energy is in the form of vibrations in the range of frequencies from about 1 Hz to 10 Hz.

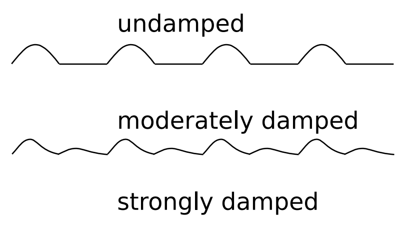
All the structures we build are resting on geological layers of dirt, mud, sand, or rock. When an earthquake wave comes along, the topmost layer acts like a system with a certain natural frequency of vibration, sort of like a cube of jello on a plate being shaken from side to side. The resonant frequency of the layer depends on how stiff it is and also on how deep it is. The ill-fated section of the Nimitz freeway was built on a layer of mud, and analysis by geologist Susan E. Hough of the U.S. Geological Survey shows that the mud layer's resonance was centered on about 2.5 Hz, and had a width covering a range from about 1 Hz to 4 Hz.

When the earthquake wave came along with its mixture of frequencies, the mud responded strongly to those that were close to its own natural 2.5 Hz frequency. Unfortunately, an engineering analysis after the quake showed that the overpass itself had a resonant frequency of 2.5 Hz as well! The mud responded strongly to the earthquake waves with frequencies close to 2.5 Hz, and the bridge responded strongly to the 2.5 Hz vibrations of the mud, causing sections of it to collapse.

i / The collapsed section of the Nimitz Freeway

Physical reason for the relationship between Q and the FWHM

What is the reason for this surprising relationship between the damping and the width of the resonance? Fundamentally, it has to do with the fact that friction causes a system to lose its "memory" of its previous state. If the Pioneer 10 space probe, coasting through the frictionless vacuum of interplanetary space, is detected by aliens a million years from now, they will be able to trace its trajectory backwards and infer that it came from our solar system. On the other hand, imagine that I shove a book along a tabletop, it comes to rest, and then someone else walks into the room. There will be no clue as to which direction the book was moving before it stopped — friction has erased its memory of its motion. Now consider the playground swing driven at twice its natural frequency, figure j, where the undamped case is repeated from figure b on page 122. In the undamped case, the first push starts the swing moving with momentum p, but when the second push comes, if there is no friction at all, it now has a momentum of exactly $-p$, and the momentum transfer from the second push is exactly enough to stop it dead. With moderate damping, however, the momentum on the rebound is not quite $-p$, and the second push's effect isn't quite as disastrous. With very strong damping, the swing comes essentially to rest long before the second push. It has lost all its memory, and the second push puts energy into the system rather than taking it



j / An x-versus-t graph of the steady-state motion of a swing being pushed at twice its resonant frequency by an impulsive force.

out. Although the detailed mathematical results with this kind of impulsive driving force are different,[11] the general results are the same as for sinusoidal driving: the less damping there is, the more of a penalty you pay for driving the system off of resonance.

High-Q speakers — *example 39*

Most good audio speakers have $Q \approx 1$, but the resonance curve for a higher-Q oscillator always lies above the corresponding curve for one with a lower Q, so people who want their car stereos to be able to rattle the windows of the neighboring cars will often choose speakers that have a high Q. Of course they could just use speakers with stronger driving magnets to increase F_m, but the speakers might be more expensive, and a high-Q speaker also has less friction, so it wastes less energy as heat.

One problem with this is that whereas the resonance curve of a low-Q speaker (its "response curve" or "frequency response" in audiophile lingo) is fairly flat, a higher-Q speaker tends to emphasize the frequencies that are close to its natural resonance. In audio, a flat response curve gives more realistic reproduction of sound, so a higher quality factor, Q, really corresponds to a *lower*-quality speaker.

Another problem with high-Q speakers is discussed in example 42 on page 133 .

Changing the pitch of a wind instrument — *example 40*

▷ A saxophone player normally selects which note to play by choosing a certain fingering, which gives the saxophone a certain resonant frequency. The musician can also, however, change the pitch significantly by altering the tightness of her lips. This corresponds to driving the horn slightly off of resonance. If the pitch can be altered by about 5% up or down (about one musical half-step) without too much effort, roughly what is the Q of a saxophone?

▷ Five percent is the width on one side of the resonance, so the full width is about 10%, $\Delta f/f_o \approx 0.1$. The equation $\Delta\omega = \omega_o/Q$ is defined in terms of angular frequency, $\omega = 2\pi f$, and we've been given our data in terms of ordinary frequency, f. The factors of 2π end up canceling out, however:

$$\begin{aligned} Q &= \frac{\omega_o}{\Delta\omega} \\ &= \frac{2\pi f_o}{2\pi\Delta f} \\ &= \frac{f_o}{f} \\ &\approx 10 \end{aligned}$$

In other words, once the musician stops blowing, the horn will continue sounding for about 10 cycles before its energy falls off by a factor of 535. (Blues and jazz saxophone players will typically choose a mouthpiece that gives a low Q, so that they can produce the bluesy pitch-slides

[11] 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 j, however, we can see that impulsive driving at $\omega = 2\omega_o$ produces a steady state with more energy than at $\omega = \omega_o$.

typical of their style. “Legit,” i.e. classically oriented players, use a higher-Q setup because their style only calls for enough pitch variation to produce a vibrato, and the higher Q makes it easier to play in tune.)

Q of a radio receiver *example 41*

▷ A radio receiver used in the FM band needs to be tuned in to within about 0.1 MHz for signals at about 100 MHz. What is its Q?

▷ As in the last example, we're given data in terms of fs, not ωs, but the factors of 2π cancel. The resulting Q is about 1000, which is extremely high compared to the Q values of most mechanical systems.

Transients

What about the motion before the steady state is achieved? When we computed the undriven motion numerically on page 124, the program had to initialize the position and velocity. By changing these two variables, we could have gotten any of an infinite number of simulations.[12] The same is true when we have an equation of motion with a driving term, $ma+bv+kx=F_m\sin\omega t$ (p. 128, equation [1]). The steady-state solutions, however, have no adjustable parameters at all — A and δ are uniquely determined by the parameters of the driving force and the oscillator itself. If the oscillator isn't initially in the steady state, then it will not have the steady-state motion at first. What kind of motion will it have?

The answer comes from realizing that if we start with the solution to the driven equation of motion, and then add to it any solution to the free equation of motion, the result,

$$x=A\sin(\omega t+\delta)+A'e^{-ct}\sin(\omega_f t+\delta') \qquad ,$$

is also a solution of the driven equation. Here, as before, ω_f is the frequency of the free oscillations ($\omega_f\approx\omega_o$ for small Q), ω is the frequency of the driving force, A and δ are related as usual to the parameters of the driving force, and A' and δ' can have any values at all. Given the initial position and velocity, we can always choose A' and δ' to reproduce them, but this is not something one often has to do in real life. What's more important is to realize that the second term dies out exponentially over time, decaying at the same rate at which a free vibration would. For this reason, the A' term is called a transient. A high-Q oscillator's transients take a long time to die out, while a low-Q oscillator always settles down to its steady state very quickly.

Boomy bass *example 42*

In example 39 on page 132, I've already discussed one of the drawbacks of a high-Q speaker, which is an uneven response curve. Another problem is that in a high-Q speaker, transients take a long time to die out. The bleeding-eardrums crowd tend to focus mostly on making

[12] 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.

their bass loud, so it's usually their woofers that have high Qs. The result is that bass notes, "ring" after the onset of the note, a phenomenon referred to as "boomy bass."

Overdamped motion

The treatment of free, damped motion on page 125 skipped over a subtle point: in the equation $\omega_f = \sqrt{k/m - b^2/4m^2} = \omega_o\sqrt{1-1/4Q^2}$, $Q < 1/2$ results in an answer that is the square root of a negative number. For example, suppose we had $k = 0$, which corresponds to a neutral equilibrium. A physical example would be a mass sitting in a tub of syrup. If we set it in motion, it won't oscillate — it will simply slow to a stop. This system has $Q = 0$. The equation of motion in this case is $ma + bv = 0$, or, more suggestively,

$$m\frac{\mathrm{d}v}{\mathrm{d}t} + bv = 0 \qquad .$$

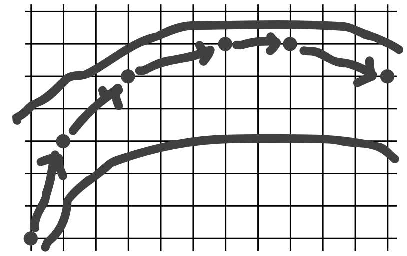
One can easily verify that this has the solution $v = (\text{constant})e^{-bt/m}$, and integrating, we find $x = (\text{constant})e^{-bt/m} + (\text{constant})$. In other words, the reason ω_f comes out to be mathematical nonsense[13] is that we were incorrect in assuming a solution that oscillated at a frequency ω_f. The actual motion is not oscillatory at all.

In general, systems with $Q < 1/2$, called overdamped systems, do not display oscillatory motion. Most cars' shock absorbers are designed with $Q \approx 1/2$, since it's undesirable for the car to undulate up and down for a while after you go over a bump. (Shocks with extremely low values of Q are not good either, because such a system takes a very long time to come back to equilibrium.) It's not particularly important for our purposes, but for completeness I'll note, as you can easily verify, that the general solution to the equation of motion for $0 < Q < 1/2$ is of the form $x = Ae^{-ct} + Be^{-dt}$, while $Q = 1/2$, called the critically damped case, gives $x = (A + Bt)e^{-ct}$.

[13] Actually, if you know about complex numbers and Euler's theorem, it's not quite so nonsensical.

3.4 Motion in Three Dimensions

3.4.1 The Cartesian perspective



a / The car can change its x and y motions by one square every turn.

When my friends and I were bored in high school, we used to play a paper-and-pencil game which, although we never knew it, was Very Educational — in fact, it pretty much embodies the entire world-view of classical physics. To play the game, you draw a racetrack on graph paper, and try to get your car around the track before anyone else. The default is for your car to continue at constant speed in a straight line, so if it moved three squares to the right and one square up last turn, it will do the same this turn. You can also control the car's motion by changing its Δx and Δy by up to one unit. If it moved three squares to the right last turn, you can have it move anywhere from two to four squares to the right this turn.

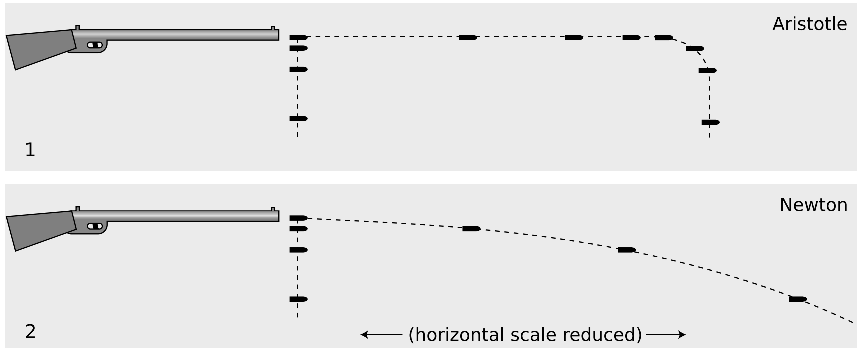
b / French mathematician René Descartes invented analytic geometry; Cartesian (xyz) coordinates are named after him. He did work in philosophy, and was particularly interested in the mind-body problem. He was a skeptic and an antiaristotelian, and, probably for fear of religious persecution, spent his adult life in the Netherlands, where he fathered a daughter with a Protestant peasant whom he could not marry. He kept his daughter's existence secret from his enemies in France to avoid giving them ammunition, but he was crushed when she died of scarlatina at age 5. A pious Catholic, he was widely expected to be sainted. His body was buried in Sweden but then reburied several times in France, and along the way everything but a few fingerbones was stolen by peasants who expected the body parts to become holy relics.

The fundamental way of dealing with the direction of an object's motion in physics is to use conservation of momentum, since momentum depends on direction. Up until now, we've only done momentum in one dimension. How does this relate to the racetrack game? In the game, the motion of a car from one turn to the next is represented by its Δx and Δy. In one dimension, we would only need Δx, which could be related to the velocity, $\Delta x/\Delta t$, and the momentum, $m\Delta x/\Delta t$. In two dimensions, the rules of the game amount to a statement that if there is no momentum transfer, then both $m\Delta x/\Delta t$ and $m\Delta y/\Delta t$ stay the same. In other words, there are two flavors of momentum, and they are *separately* conserved. All of this so far has been done with an artificial division of time into "turns," but we can fix that by redefining everything in terms of derivatives, and for motion in three dimensions rather than two, we augment x and y with z:

$$v_x = \mathrm{d}x/\mathrm{d}t \qquad v_y = \mathrm{d}y/\mathrm{d}t \qquad v_z = \mathrm{d}z/\mathrm{d}t$$

and

$$p_x = mv_x \qquad p_y = mv_y \qquad p_z = mv_z$$



c / Bullets are dropped and shot at the same time.

We call these the x, y, and z components of the velocity and the momentum.

There is both experimental and theoretical evidence that the x, y, and z momentum components are separately conserved, and that a momentum transfer (force) along one axis has no effect on the momentum components along the other two axes. On page 50, for example, I argued that it was impossible for an air hockey puck to make a 180-degree turn spontaneously, because then in the frame moving along with the puck, it would have begun moving after starting from rest. Now that we're working in two dimensions, we might wonder whether the puck could spontaneously make a 90-degree turn, but exactly the same line of reasoning shows that this would be impossible as well, which proves that the puck can't trade x-momentum for y-momentum. A more general proof of separate conservation will be given on page 160, after some of the appropriate mathematical techniques have been introduced.

As an example of the experimental evidence for separate conservation of the momentum components, figure c shows correct and incorrect predictions of what happens if you shoot a rifle and arrange for a second bullet to be dropped from the same height at exactly the same moment when the first one left the barrel. Nearly everyone expects that the dropped bullet will reach the dirt first, and Aristotle would have agreed, since he believed that the bullet had to lose its horizontal motion before it could start moving vertically. In reality, we find that the vertical momentum transfer between the earth and the bullet is completely unrelated to the horizontal momentum. The bullet ends up with $p_y < 0$, while the planet picks up an upward momentum $p_y > 0$, and the total momentum in the y direction remains zero. Both bullets hit the ground at the same time. This is much simpler than the Aristotelian version!

The Pelton waterwheel *example 43*

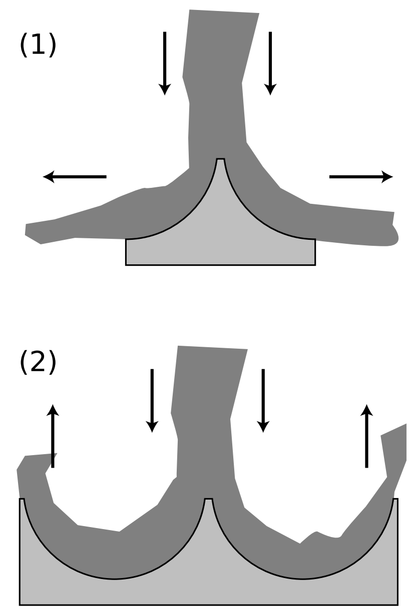
▷ There is a general class of machines that either do work on a gas or liquid, like a boat's propeller, or have work done on them by a gas or liquid, like the turbine in a hydroelectric power plant. Figure d shows two types of surfaces that could be attached to the circumference of an old-fashioned waterwheel. Compare the force exerted by the water in the two cases.

▷ Let the x axis point to the right, and the y axis up. In both cases, the stream of water rushes down onto the surface with momentum $p_{y,i} = -p_o$, where the subscript i stands for "initial," i.e. before the collision.

In the case of surface 1, the streams of water leaving the surface have no momentum in the y direction, and their momenta in the x direction cancel. The final momentum of the water is zero along both axes, so its entire momentum, $-p_o$, has been transferred to the waterwheel.

When the water leaves surface 2, however, its momentum isn't zero. If we assume there is no friction, it's $p_{y,f} = p_o$, with the positive sign indicating upward momentum. The change in the water's momentum is $p_{y,f} - p_{y,i} = 2p_o$, and the momentum transferred to the waterwheel is $-2p_o$.

Force is defined as the rate of transfer of momentum, so surface 2 experiences double the force. A waterwheel constructed in this way is known as a Pelton waterwheel.



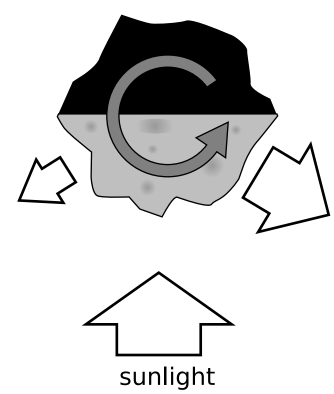
d / Two surfaces that could be used to extract energy from a stream of water.

The Yarkovsky effect *example 44*

We think of the planets and asteroids as inhabiting their orbits permanently, but it is possible for an orbit to change over periods of millions or billions of years, due to a variety of effects. For asteroids with diameters of a few meters or less, an important mechanism is the Yarkovsky effect, which is easiest to understand if we consider an asteroid spinning about an axis that is exactly perpendicular to its orbital plane.

The illuminated side of the asteroid is relatively hot, and radiates more infrared light than the dark (night) side. Light has momentum, and a total force away from the sun is produced by combined effect of the sunlight hitting the asteroid and the imbalance between the momentum radiated away on the two sides. This force, however, doesn't cause the asteroid's orbit to change over time, since it simply cancels a tiny fraction of the sun's gravitational attraction. The result is merely a tiny, undetectable violation of Kepler's law of periods.

Consider the sideways momentum transfers, however. In figure e, the part of the asteroid on the right has been illuminated for half a spin-period (half a "day") by the sun, and is hot. It radiates more light than the morning side on the left. This imbalance produces a total force in the x direction which points to the left. If the asteroid's orbital motion is to the left, then this is a force in the same direction as the motion, which will do positive work, increasing the asteroid's energy and boosting it into an orbit with a greater radius. On the other hand, if the asteroid's spin and orbital motion are in opposite directions, the Yarkovsky push brings the asteroid spiraling in closer to the sun.



e / An asteroid absorbs visible light from the sun, and gets rid of the energy by radiating infrared light.

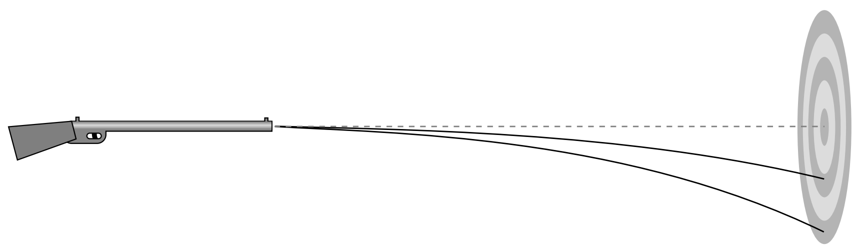
Calculations show that it takes on the order of 10^7 to 10^8 years for the Yarkovsky effect to move an asteroid out of the asteroid belt and into the vicinity of earth's orbit, and this is about the same as the typical age of a meteorite as estimated by its exposure to cosmic rays. The Yarkovsky effect doesn't remove all the asteroids from the asteroid belt, because many of them have orbits that are stabilized by gravitational interactions with Jupiter. However, when collisions occur, the fragments can end up in orbits which are not stabilized in this way, and they may then end up reaching the earth due to the Yarkovsky effect. The cosmic-ray technique is really telling us how long it has been since the fragment was broken out of its parent.

Discussion Questions

A The following is an incorrect explanation of a fact about target shooting:

"Shooting a high-powered rifle with a high muzzle velocity is different from shooting a less powerful gun. With a less powerful gun, you have to aim quite a bit above your target, but with a more powerful one you don't have to aim so high because the bullet doesn't drop as fast."

What is the correct explanation?



f / Discussion question A.

B You have thrown a rock, and it is flying through the air in an arc. If the earth's gravitational force on it is always straight down, why doesn't it just go straight down once it leaves your hand?

C Consider the example of the bullet that is dropped at the same moment another bullet is fired from a gun. What would the motion of the two bullets look like to a jet pilot flying alongside in the same direction as the shot bullet and at the same horizontal speed?

3.4.2 Rotational invariance

The Cartesian approach requires that we choose x, y, and z axes. How do we choose them correctly? The answer is that it had better not matter which directions the axes point (provided they're perpendicular to each other), or where we put the origin, because if it did matter, it would mean that space was asymmetric. If there was a certain point in the universe that was the right place to put the origin, where would it be? The top of Mount Olympus? The United Nations headquarters? We find that experiments come out the same no matter where we do them, and regardless of which way the laboratory is oriented, which indicates that no location in space or direction in space is special in any way.[14]

This is closely related to the idea of Galilean relativity stated on page 22, from which we already know that the absolute *motion* of a frame of reference is irrelevant and undetectable. Observers using frames of reference that are in motion relative to each other will not even agree on the permanent identity of a particular point in space, so it's not possible for the laws of physics to depend on where you are in space. For instance, if gravitational energies were proportional to $m_1 m_2$ in one location but to $(m_1 m_2)^{1.00001}$ in another, then it would be possible to determine when you were in a state of absolute motion, because the behavior of gravitational interactions would change as you moved from one region to the other.

Because of this close relationship, we restate the principle of Galilean relativity in a more general form. This extended principle of Galilean relativity states that the laws of physics are no different in one time and place than in another, and that they also don't depend on your orientation or your motion, provided that your motion is in a straight line and at constant speed.

The irrelevance of time and place could have been stated in chapter 1, but since this section is the first one in which we're dealing with three-dimensional physics in full generality, the irrelevance of orientation is what we really care about right now. This property of the laws of physics is called rotational invariance. The word "invariance" means a lack of change, i.e. the laws of physics don't change when we reorient our frame of reference.

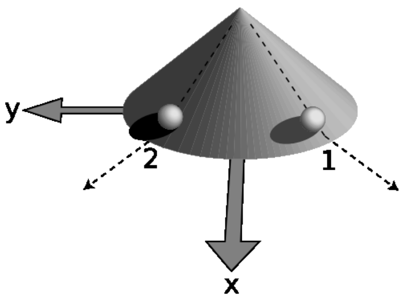
[14]Of course, you could tell in a sealed laboratory which way was down, but that's because there happens to be a big planet nearby, and the planet's gravitational field reaches into the lab, not because space itself has a special down direction. Similarly, if your experiment was sensitive to magnetic 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.

Rotational invariance of gravitational interactions — *example 45*

Gravitational energies depend on the quantity $1/r$, which by the Pythagorean theorem equals

$$\frac{1}{\sqrt{\Delta x^2+\Delta y^2+\Delta z^2}} \quad .$$

Rotating a line segment doesn't change its length, so this expression comes out the same regardless of which way we orient our coordinate axes. Even though Δx, Δy, and Δz are different in differently oriented coordinate systems, r is the same.



g / Two balls roll down a cone and onto a plane.

Kinetic energy — *example 46*

Kinetic energy equals $(1/2)mv^2$, but what does that mean in three dimensions, where we have v_x, v_y, and v_z? If you were tempted to add the components and calculate $K=(1/2)m(v_x+v_y+v_z)^2$, figure g should convince you otherwise. Using that method, we'd have to assign a kinetic energy of zero to ball number 1, since its negative v_y would exactly cancel its positive v_x, whereas ball number 2's kinetic energy wouldn't be zero. This would violate rotational invariance, since the balls would behave differently.

The only possible way to generalize kinetic energy to three dimensions, without violating rotational invariance, is to use an expression that resembles the Pythagorean theorem,

$$v=\sqrt{v_x^2+v_y^2+v_z^2} \quad ,$$

which results in

$$K=\frac{1}{2}m\left(v_x^2+v_y^2+v_z^2\right) \quad .$$

Since the velocity components are squared, the positive and negative signs don't matter, and the two balls in the example behave the same way.

3.4.3 Vectors

Remember the title of this book? It would have been possible to obtain the result of example 46 by applying the Pythagorean theorem to $\mathrm{d}x$, $\mathrm{d}y$, and $\mathrm{d}z$, and then dividing by $\mathrm{d}t$, but the rotational invariance approach is *simpler*, and is useful in a much broader context. Even with a quantity you presently know nothing about, say the magnetic field, you can infer that if the components of the magnetic field are B_x, B_y, and B_z, then the physically useful way to talk about the strength of the magnetic field is to define it as $\sqrt{B_x^2+B_y^2+B_z^2}$. Nature knows your brain cells are precious, and doesn't want you to have to waste them by memorizing mathematical rules that are different for magnetic fields than for velocities.

When mathematicians see that the same set of techniques is useful in many different contexts, that's when they start making definitions that allow them to stop reinventing the wheel. The ancient Greeks, for example, had no general concept of fractions. They couldn't say that a circle's radius divided by its diameter was equal to the number 1/2. They had to say that the radius and the diameter were in the ratio of one to two. With this limited number concept, they couldn't have said that water was dripping out of a tank at a rate of 3/4 of a barrel per day; instead, they would have had to say that over four days, three barrels worth of water would be lost. Once enough of these situations came up, some clever mathematician finally realized that it would make sense to define something called a fraction, and that one could think of these fraction thingies as numbers that lay in the gaps between the traditionally recognized numbers like zero, one, and two. Later generations of mathematicians introduced further subversive generalizations of the number concepts, inventing mathematical creatures like negative numbers, and the square root of two, which can't be expressed as a fraction.

In this spirit, we define a *vector* as any quantity that has both an amount and a direction in space. In contradistinction, a *scalar* has an amount, but no direction. Time and temperature are scalars. Velocity, acceleration, momentum, and force are vectors. In one dimension, there are only two possible directions, and we can use positive and negative numbers to indicate the two directions. In more than one dimension, there are infinitely many possible directions, so we can't use the two symbols + and − to indicate the direction of a vector. Instead, we can specify the three components of the vector, each of which can be either negative or positive. We represent vector quantities in handwriting by writing an arrow above them, so for example the momentum vector looks like this, $\vec{p}$, but the arrow looks ugly in print, so in books vectors are usually shown in bold-face type: $\mathbf{p}$. A straightforward way of thinking about vectors is that a vector equation really represents three different equations. For instance, conservation of momentum could be written in terms of the three components,

$$\begin{aligned}\Delta p_x &= 0\\ \Delta p_y &= 0\\ \Delta p_z &= 0 \qquad ,\end{aligned}$$

or as a single vector equation,[15]

$$\Delta\mathbf{p} = 0 \qquad .$$

The following table summarizes some vector operations.

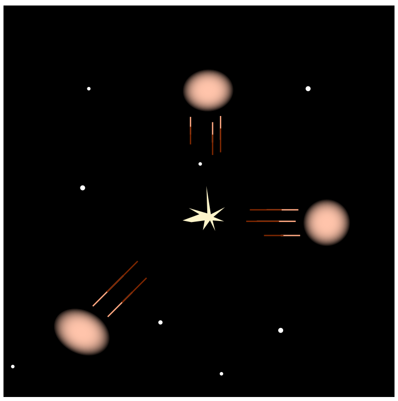
[15]The zero here is really a zero *vector*, i.e. a vector whose components are all zero, so we should really represent it with a boldface 0. There's usually not much danger of confusion, however, so most books, including this one, don't use boldface for the zero vector.

operation	definition
$\lvert\textbf{vector}\rvert$	$\sqrt{vector_x^2 + vector_y^2 + vector_z^2}$
vector + **vector**	Add component by component.
vector − **vector**	Subtract component by component.
vector $\cdot$ *scalar*	Multiply each component by the scalar.
vector / *scalar*	Divide each component by the scalar.

The first of these is called the *magnitude* of the vector; in one dimension, where a vector only has one component, it amounts to taking the absolute value, hence the similar notation.

Self-Check

Translate the equations $F_x=ma_x$, $F_y=ma_y$, and $F_z=ma_z$ into a single equation in vector notation. ▷ Answer, p. 706



h / Example 47.

An explosion — *example 47*

▷ Astronomers observe the planet Mars as the Martians fight a nuclear war. The Martian bombs are so powerful that they rip the planet into three separate pieces of liquefied rock, all having the same mass. If one fragment flies off with velocity components $v_{1x}=0$, $v_{1y}=1.0\times10^4$ km/hr, and the second with $v_{2x}=1.0\times10^4$ km/hr, $v_{2y}=0$, what is the magnitude of the third one's velocity?

▷ We work the problem in the center of mass frame, in which the planet initially had zero momentum. After the explosion, the vector sum of the momenta must still be zero. Vector addition can be done by adding components, so

$$mv_{1x} + mv_{2x} + mv_{3x} = 0$$

and

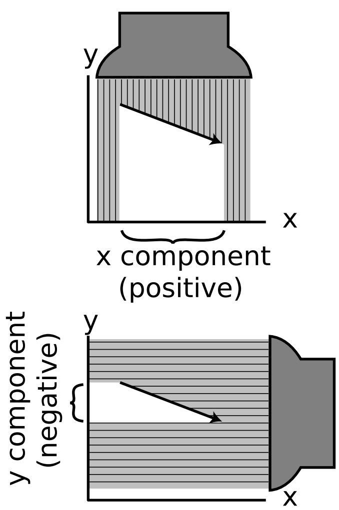
$$mv_{1y} + mv_{2y} + mv_{3y} = 0 \qquad ,$$

where we have used the same symbol m for all the terms, because the fragments all have the same mass. The masses can be eliminated by dividing each equation by m, and we find

$$v_{3x} = 1.0\times10^4 \text{ km/hr}$$
$$v_{3y} = 1.0\times10^4 \text{ km/hr} \qquad ,$$

which gives a magnitude of

$$|\mathbf{v}_3| = \sqrt{v_{3x}^2 + v_{3y}^2}$$
$$= 1.4\times10^4 \text{ km/hr} \qquad .$$



i / The geometric interpretation of a vector's components.

Geometric representation of vectors

A vector in two dimensions can be easily visualized by drawing an arrow whose length represents its magnitude and whose direction represents its direction. The x component of a vector can then be visualized, i, as the length of the shadow it would cast in a beam of light projected onto the x axis, and similarly for the y component.

Shadows with arrowheads pointing back against the direction of the positive axis correspond to negative components.

In this type of diagram, the negative of a vector is the vector with the same magnitude but in the opposite direction. Multiplying a vector by a scalar is represented by lengthening the arrow by that factor, and similarly for division.

Self-Check

Given vector **Q** represented by an arrow below, draw arrows representing the vectors 1.5**Q** and −**Q**.

▷ Answer, p. 706

A useless vector operation *example 48*

The way I've defined the various vector operations above aren't as arbitrary as they seem. There are many different vector operations that we could define, but only some of the possible definitions are mathematically useful. Consider the operation of multiplying two vectors component by component to produce a third vector:

$$R_x = P_x Q_x$$
$$R_y = P_y Q_y$$
$$R_z = P_z Q_z$$

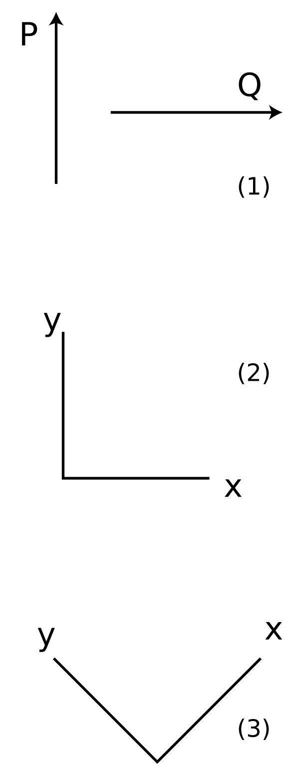
As a simple example, we choose vectors **P** and **Q** to have length 1, and make them perpendicular to each other, as shown in figure j/1. If we compute the result of our new vector operation using the coordinate system shown in j/2, we find:

$$R_x = 0$$
$$R_y = 0$$
$$R_z = 0$$

The x component is zero because $P_x = 0$, the y component is zero because $Q_y = 0$, and the z component is of course zero because both vectors are in the x-y plane. However, if we carry out the same operations in coordinate system j/3, rotated 45 degrees with respect to the previous one, we find

$$R_x = -1/2$$
$$R_y = 1/2$$
$$R_z = 0$$

The operation's result depends on what coordinate system we use, and since the two versions of **R** have different lengths (one being zero and the other nonzero), they don't just represent the same answer expressed in two different coordinate systems. Such an operation will never be useful in physics, because experiments show physics works the same regardless of which way we orient the laboratory building!



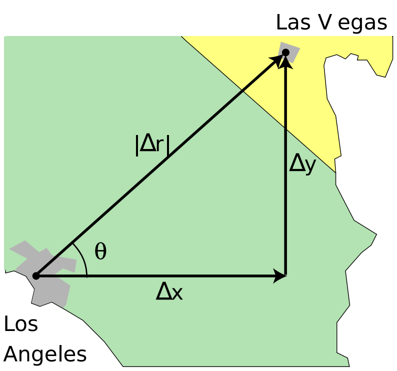
j / Two vectors, 1, to which we apply the same operation in two different frames of reference, 2 and 3.

The useful vector operations, such as addition and scalar multiplication, are rotationally invariant, i.e. come out the same regardless of the orientation of the coordinate system.

All the vector techniques can be applied to any kind of vector, but the graphical representation of vectors as arrows is particularly natural for vectors that represent lengths and distances. We define a vector called **r** whose components are the coordinates of a particular point in space, x, y, and z. The $\Delta\mathbf{r}$ vector, whose components are Δx, Δy, and Δz, can then be used to represent motion that starts at one point and ends at another. Adding two $\Delta\mathbf{r}$ vectors is interpreted as a trip with two legs: by computing the $\Delta\mathbf{r}$ vector going from point A to point B plus the vector from B to C, we find the vector that would have taken us directly from A to C.

Calculations with magnitude and direction

If you ask someone where Las Vegas is compared to Los Angeles, she is unlikely to say that the Δx is 290 km and the Δy is 230 km, in a coordinate system where the positive x axis is east and the y axis points north. She will probably say instead that it's 370 km to the northeast. If she was being precise, she might specify the direction as 38° counterclockwise from east. In two dimensions, we can always specify a vector's direction like this, using a single angle. A magnitude plus an angle suffice to specify everything about the vector. The following two examples show how we use trigonometry and the Pythagorean theorem to go back and forth between the x-y and magnitude-angle descriptions of vectors.



k / Example 49.

Finding the magnitude and angle from the components — *example 49*

▷ Given that the $\Delta\mathbf{r}$ vector from LA to Las Vegas has Δx=290 km and Δy=230 km, how would we find the magnitude and direction of $\Delta\mathbf{r}$?

▷ We find the magnitude of $\Delta\mathbf{r}$ from the Pythagorean theorem:

$$\begin{aligned} |\Delta\mathbf{r}| &= \sqrt{\Delta x^2 + \Delta y^2} \\ &= 370\ \text{km} \end{aligned}$$

We know all three sides of the triangle, so the angle θ can be found using any of the inverse trig functions. For example, we know the opposite and adjacent sides, so

$$\begin{aligned} \theta &= \tan^{-1}\frac{\Delta y}{\Delta x} \\ &= 38^\circ \qquad . \end{aligned}$$

Finding the components from the magnitude and angle — *example 50*

▷ Given that the straight-line distance from Los Angeles to Las Vegas is 370 km, and that the angle θ in the figure is 38°, how can the x and y components of the $\Delta\mathbf{r}$ vector be found?

▷ The sine and cosine of θ relate the given information to the information we wish to find:

$$\cos\theta = \frac{\Delta x}{|\Delta \mathbf{r}|}$$
$$\sin\theta = \frac{\Delta y}{|\Delta \mathbf{r}|}$$

Solving for the unknowns gives

$$\begin{aligned}\Delta x &= |\Delta \mathbf{r}| \cos\theta \\ &= 290\ \text{km} \\ \Delta y &= |\Delta \mathbf{r}| \sin\theta \\ &= 230\ \text{km}\end{aligned}$$

The following example shows the correct handling of the plus and minus signs, which is usually the main cause of mistakes by students.

Negative components *example 51*

▷ San Diego is 120 km east and 150 km south of Los Angeles. An airplane pilot is setting course from San Diego to Los Angeles. At what angle should she set her course, measured counterclockwise from east, as shown in the figure?

▷ If we make the traditional choice of coordinate axes, with x pointing to the right and y pointing up on the map, then her Δx is negative, because her final x value is less than her initial x value. Her Δy is positive, so we have

$$\begin{aligned}\Delta x &= -120\ \text{km} \\ \Delta y &= 150\ \text{km} \quad .\end{aligned}$$

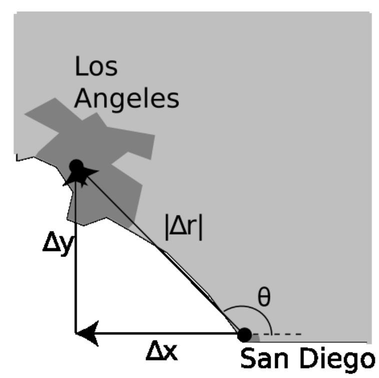
If we work by analogy with the example 49, we get

$$\begin{aligned}\theta &= \tan^{-1}\frac{\Delta y}{\Delta x} \\ &= \tan^{-1}(-1.25) \\ &= -51^\circ \quad .\end{aligned}$$

According to the usual way of defining angles in trigonometry, a negative result means an angle that lies clockwise from the x axis, which would have her heading for the Baja California. What went wrong? The answer is that when you ask your calculator to take the arctangent of a number, there are always two valid possibilities differing by 180°. That is, there are two possible angles whose tangents equal -1.25:

$$\begin{aligned}\tan 129^\circ &= -1.25 \\ \tan(-51^\circ) &= -1.25\end{aligned}$$

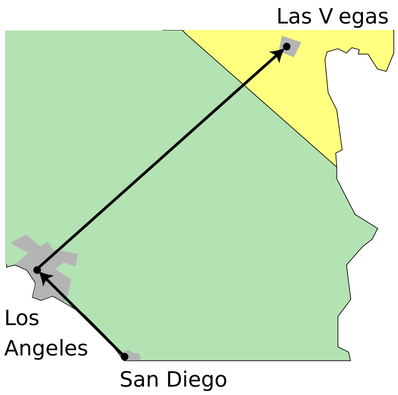
You calculator doesn't know which is the correct one, so it just picks one. In this case, the one it picked was the wrong one, and it was up to you to add 180° to it to find the right answer.



l / Example 51.

Addition of vectors given their components

The easiest type of vector addition is when you are in possession of the components, and want to find the components of their sum.



m / Example 52.

San Diego to Las Vegas — *example 52*

▷ Given the Δx and Δy values from the previous examples, find the Δx and Δy from San Diego to Las Vegas.

▷

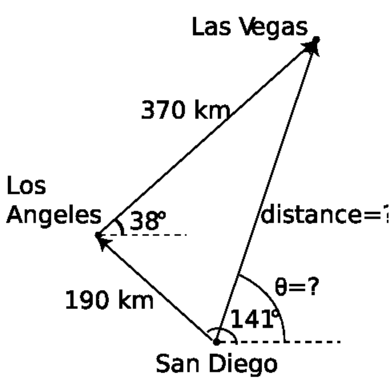
$$\begin{aligned}\Delta x_{total} &= \Delta x_1 + \Delta x_2\\ &= -120\ \text{km} + 290\ \text{km}\\ &= 170\ \text{km}\\ \Delta y_{total} &= \Delta y_1 + \Delta y_2\\ &= 150\ \text{km} + 230\ \text{km}\\ &= 380\end{aligned}$$

Addition of vectors given their magnitudes and directions

In this case, you must first translate the magnitudes and directions into components, and the add the components.

Graphical addition of vectors

Often the easiest way to add vectors is by making a scale drawing on a piece of paper. This is known as graphical addition, as opposed to the analytic techniques discussed previously.



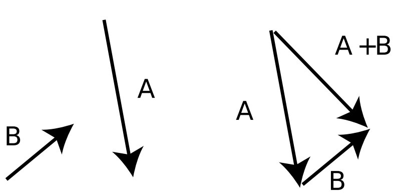
n / Example 53.

From San Diego to Las Vegas, graphically — *example 53*

▷ Given the magnitudes and angles of the $\Delta\mathbf{r}$ vectors from San Diego to Los Angeles and from Los Angeles to Las Vegas, find the magnitude and angle of the $\Delta\mathbf{r}$ vector from San Diego to Las Vegas.

▷ Using a protractor and a ruler, we make a careful scale drawing, as shown in the figure. A scale of 1 cm$\leftrightarrow$10 km was chosen for this solution. With a ruler, we measure the distance from San Diego to Las Vegas to be 3.8 cm, which corresponds to 380 km. With a protractor, we measure the angle θ to be 71 °.

Even when we don't intend to do an actual graphical calculation with a ruler and protractor, it can be convenient to diagram the addition of vectors in this way. With $\Delta\mathbf{r}$ vectors, it intuitively makes sense to lay the vectors tip-to-tail and draw the sum vector from the tail of the first vector to the tip of the second vector. We can do the same when adding other vectors such as force vectors.



o / Adding vectors graphically by placing them tip-to-tail, like a train.

Unit vector notation

When we want to specify a vector by its components, it can be cumbersome to have to write the algebra symbol for each component:

$$\Delta x = 290\ \text{km}, \quad \Delta y = 230\ \text{km}$$

A more compact notation is to write

$$\Delta\mathbf{r} = (290\text{ km})\hat{\mathbf{x}} + (230\text{ km})\hat{\mathbf{y}},$$

where the vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, called the unit vectors, are defined as the vectors that have magnitude equal to 1 and directions lying along the x, y, and z axes. In speech, they are referred to as "x-hat," "y-hat," and "z-hat."

A slightly different, and harder to remember, version of this notation is unfortunately more prevalent. In this version, the unit vectors are called $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$:

$$\Delta\mathbf{r} = (290\text{ km})\hat{\mathbf{i}} + (230\text{ km})\hat{\mathbf{j}} \qquad .$$

Applications to relative motion, momentum, and force

Vector addition is the correct way to generalize the one-dimensional concept of adding velocities in relative motion, as shown in the following example:

Velocity vectors in relative motion *example 54*

▷ You wish to cross a river and arrive at a dock that is directly across from you, but the river's current will tend to carry you downstream. To compensate, you must steer the boat at an angle. Find the angle θ, given the magnitude, $|\mathbf{v}_{WL}|$, of the water's velocity relative to the land, and the maximum speed, $|\mathbf{v}_{BW}|$, of which the boat is capable relative to the water.

▷ The boat's velocity relative to the land equals the vector sum of its velocity with respect to the water and the water's velocity with respect to the land,

$$\mathbf{v}_{BL} = \mathbf{v}_{BW} + \mathbf{v}_W L \qquad .$$

If the boat is to travel straight across the river, i.e., along the y axis, then we need to have $\mathbf{v}_{BL,x} = 0$. This x component equals the sum of the x components of the other two vectors,

$$\mathbf{v}_B L, x = \mathbf{v}_B W, x + \mathbf{v}_W L, x \qquad ,$$

or

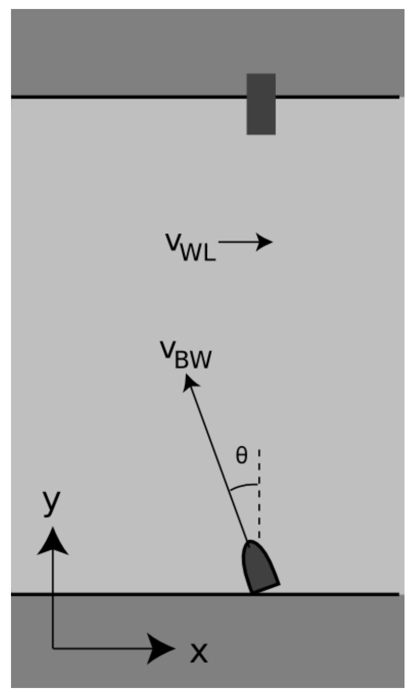
$$0 = -|\mathbf{v}_{BW}|\sin\theta + |\mathbf{v}_{WL}| \qquad .$$

Solving for θ, we find

$$\sin\theta = |\mathbf{v}_{WL}|/|\mathbf{v}_{BW}| \qquad ,$$

so

$$\theta = \sin^{-1}\frac{|\mathbf{v}_{WL}|}{\mathbf{v}_{BW}} \qquad .$$



p / Example 54

How to generalize one-dimensional equations *example 55*

▷ How can the one-dimensional relationships

$$p_{total} = m_{total} v_{cm}$$

and

$$x_{cm} = \frac{\sum_j m_j x_j}{\sum_j m_j}$$

be generalized to three dimensions?

▷ Momentum and velocity are vectors, since they have directions in space. Mass is a scalar. If we rewrite the first equation to show the appropriate quantities notated as vectors,

$$\mathbf{p}_{total} = m_{total} \mathbf{v}_{cm} \quad ,$$

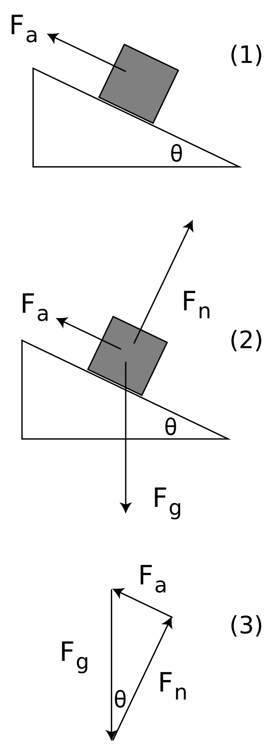
we get a valid mathematical operation, the multiplication of a vector by a scalar. Similarly, the second equation becomes

$$\mathbf{r}_{cm} = \frac{\sum_j m_j \mathbf{r}_j}{\sum_j m_j} \quad ,$$

which is also valid. Each term in the sum on top contains a vector multiplied by a scalar, which gives a vector. Adding up all these vectors gives a vector, and dividing by the scalar sum on the bottom gives another vector.

This kind of wave-the-magic-wand-and-write-it-all-in-bold-face technique will always give the right generalization from one dimension to three, provided that the result makes sense mathematically — if you find yourself doing something nonsensical, such as adding a scalar to a vector, then you haven't found the generalization correctly.

Force is a vector, and we add force vectors when more than one force acts on the same object.



q / Example 56.

Pushing a block up a ramp *example 56*

▷ Figure q/1 shows a block being pushed up a frictionless ramp at constant speed by an applied force F_a. How much force is required, in terms of the block's mass, m, and the angle of the ramp, θ?

▷ We analyzed this simple machine in example 30 on page 119 using the concept of work. Here we'll do it using vector addition of forces. Figure q/2 shows the other two forces acting on the block: a normal force, F_n, created by the ramp, and the gravitational force, F_g. Because the block is being pushed up at constant speed, it has zero acceleration, and the total force on it must be zero. In figure q/3, we position all the force vectors tip-to-tail for addition. Since they have to add up to zero, they must join up without leaving a gap, so they form a triangle. Using trigonometry we find

$$\begin{aligned} F_a &= F_g \sin\theta \\ &= mg \sin\theta \quad . \end{aligned}$$

Buoyancy, again *example 57*

In example 10 on page 45, we found that the energy required to raise a cube immersed in a fluid is as if the cube's mass had been reduced by an amount equal to the mass of the fluid that otherwise would have been in the volume it occupies (Archimedes' principle). From the energy perspective, this effect occurs because raising the cube allows a certain amount of fluid to move downward, and the decreased gravitational energy of the fluid tends to offset the increased gravitational energy of the cube. The proof given there, however, could not easily be extended to other shapes.

Thinking in terms of force rather than energy, it becomes easier to give a proof that works for any shape. A certain upward force is needed to support the object in figure r. If this force was applied, then the object would be in equilibrium: the vector sum of all the forces acting on it would be zero. These forces are $\mathbf{F}_a$, the upward force just mentioned, $\mathbf{F}_g$, the downward force of gravity, and $\mathbf{F}_f$, the total force from the fluid:

$$\mathbf{F}_a + \mathbf{F}_g + \mathbf{F}_f = 0$$

Since the fluid is under more pressure at a greater depth, the part of the fluid underneath the object tends to make more force than the part above, so the fluid tends to help support the object.

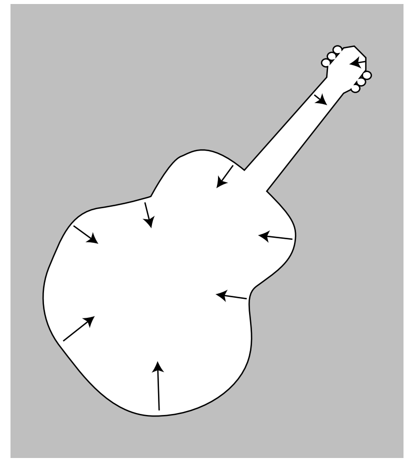
Now suppose the object was removed, and instantly replaced with an equal volume of fluid. The new fluid would be in equilibrium without any force applied to hold it up, so

$$\mathbf{F}_{gf} + \mathbf{F}_f = 0 \qquad ,$$

where $\mathbf{F}_{gf}$, the weight of the fluid, is not the same as $\mathbf{F}_g$, the weight of the object, but $\mathbf{F}_f$ is the same as before, since the pressure of the surrounding fluid is the same as before at any particular depth. We therefore have

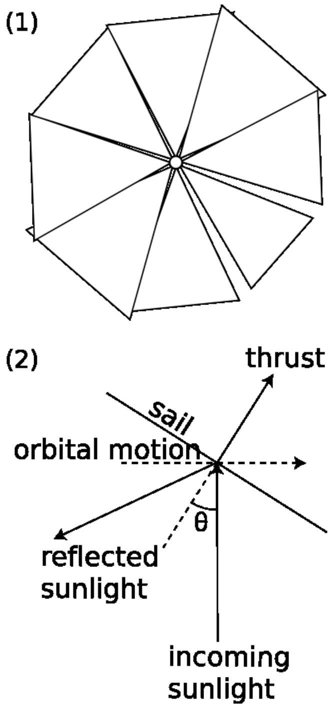
$$\mathbf{F}_a = -\left(\mathbf{F}_g - \mathbf{F}_{gf}\right) \qquad ,$$

which is Archimedes' principle in terms of force: the force required to support the object is lessened by an amount equal to the weight of the fluid that would have occupied its volume.

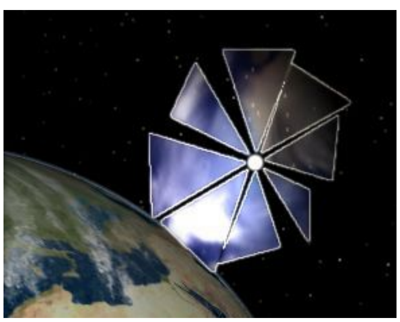


r / Archimedes' principle works regardless of whether the object is a cube. The fluid makes a force on every square millimeter of the object's surface.

By the way, the word "pressure" that I threw around casually in the preceding example has a precise technical definition: force per unit area. The SI units of pressure are N/m^2, which can be abbreviated as pascals, 1 Pa = 1 N/m^2. Atmospheric pressure is about 100 kPa. By applying the equation $\mathbf{F}_g + \mathbf{F}_f = 0$ to the top and bottom surfaces of a cubical volume of fluid, one can easily prove that the difference in pressure between two different depths is $\Delta P = \rho g \Delta y$. (In physics, "fluid" can refer to either a gas or a liquid.) Pressure is discussed in more detail in chapter 5.



s / Example 58.



t / An artist's rendering of what Cosmos 1 would have looked like in orbit.

A solar sail *example 58*

A solar sail, figure s/1, allows a spacecraft to get its thrust without using internal stores of energy or having to carry along mass that it can shove out the back like a rocket. Sunlight strikes the sail and bounces off, transferring momentum to the sail. A working 30-meter-diameter solar sail, Cosmos 1, was built by an American company, and was supposed to be launched into orbit aboard a Russian booster launched from a submarine, but launch attempts in 2001 and 2005 both failed.

In this example, we will calculate the optimal orientation of the sail, assuming that "optimal" means changing the vehicle's energy as rapidly as possible. For simplicity, we model the complicated shape of the sail's surface as a disk, seen edge-on in figure s/2, and we assume that the craft is in a nearly circular orbit around the sun, hence the 90-degree angle between the direction of motion and the incoming sunlight. We assume that the sail is 100% reflective. The orientation of the sail is specified using the angle θ between the incoming rays of sunlight and the perpendicular to the sail. In other words, θ=0 if the sail is catching the sunlight full-on, while $\theta=90°$ means that the sail is edge-on to the sun.

Conservation of momentum gives

$$\mathbf{p}_{light,i} = \mathbf{p}_{light,f} + \Delta\mathbf{p}_{sail} \qquad ,$$

where $\Delta\mathbf{p}_{sail}$ is the change in momentum picked up by the sail. Breaking this down into components, we have

$$\begin{aligned} 0 &= p_{light,f,x} + \Delta p_{sail,x} \qquad \text{and} \\ p_{light,i,y} &= p_{light,f,y} + \Delta p_{sail,y} \qquad . \end{aligned}$$

As in example 44 on page 137, the component of the force that is directly away from the sun (up in figure s/2) doesn't change the energy of the craft, so we only care about $\Delta p_{sail,x}$, which equals $-p_{light,f,x}$. The outgoing light ray forms an angle of 2θ with the negative y axis, or $270° - 2\theta$ measured counterclockwise from the x axis, so the useful thrust depends on $-\cos(270° - 2\theta) = \sin 2\theta$.

However, this is all assuming a given amount of light strikes the sail. During a certain time period, the amount of sunlight striking the sail depends on the cross-sectional area the sail presents to the sun, which is proportional to $\cos\theta$. For $\theta=90°$, $\cos\theta$ equals zero, since the sail is edge-on to the sun.

Putting together these two factors, the useful thrust is proportional to $\sin 2\theta \cos\theta$, and this quantity is maximized for $\theta \approx 35°$. A counterintuitive fact about this maneuver is that as the spacecraft spirals outward, its total energy (kinetic plus gravitational) increases, but its kinetic energy actually decreases!

Discussion Questions

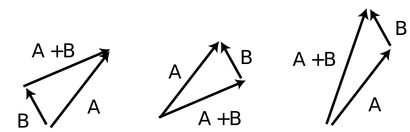
A An object goes from one point in space to another. After it arrives at its destination, how does the magnitude of its $\Delta\mathbf{r}$ vector compare with the distance it traveled?

B In several examples, I've dealt with vectors having negative components. Does it make sense as well to talk about negative and positive vectors?

C If you're doing *graphical* addition of vectors, does it matter which vector you start with and which vector you start from the other vector's tip?

D If you add a vector with magnitude 1 to a vector of magnitude 2, what magnitudes are possible for the vector sum?

E Which of these examples of vector addition are correct, and which are incorrect?

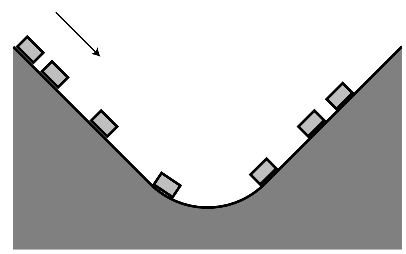


u / Discussion question E.

F Is it possible for an airplane to maintain a constant velocity vector but not a constant $|\mathbf{v}|$? How about the opposite – a constant $|\mathbf{v}|$ but not a constant velocity vector? Explain.

G New York and Rome are at about the same latitude, so the earth's rotation carries them both around nearly the same circle. Do the two cities have the same velocity vector (relative to the center of the earth)? If not, is there any way for two cities to have the same velocity vector?

H The figure shows a roller coaster car rolling down and then up under the influence of gravity. Sketch the car's velocity vectors and acceleration vectors. Pick an interesting point in the motion and sketch a set of force vectors acting on the car whose vector sum could have resulted in the right acceleration vector.



v / Discussion question H.

I The following is a question commonly asked by students:

"Why does the force vector always have to point in the same direction as the acceleration vector? What if you suddenly decide to change your force on an object, so that your force is no longer pointing in the same direction that the object is accelerating?"

What misunderstanding is demonstrated by this question? Suppose, for example, a spacecraft is blasting its rear main engines while moving forward, then suddenly begins firing its sideways maneuvering rocket as well. What does the student think Newton's laws are predicting?

J Debug the following *incorrect* solutions to this vector addition problem.

Problem: Freddi Fish™ swims 5.0 km northeast, and then 12.0 km in the direction 55 degrees west of south. How far does she end up from her starting point, and in what direction is she from her starting point?

Incorrect solution #1:
5.0 km+12.0 km=17.0 km

Incorrect solution #2:
$\sqrt{(5.0\text{ km})^2 + (12.0\text{ km})^2}$=13.0 km

Incorrect solution #3:
Let **A** and **B** be her two $\Delta\mathbf{r}$ vectors, and let $\mathbf{C} = \mathbf{A} + \mathbf{B}$. Then

$$\begin{aligned}
A_x &= (5.0\text{ km})\cos 45^\circ = 3.5\text{ km}\\
B_x &= (12.0\text{ km})\cos 55^\circ = 6.9\text{ km}\\
A_y &= (5.0\text{ km})\sin 45^\circ = 3.5\text{ km}\\
B_y &= (12.0\text{ km})\sin 55^\circ = 9.8\text{ km}\\
C_x &= A_x + B_x\\
&= 10.4\text{ km}\\
C_y &= A_y + B_y\\
&= 13.3\text{ km}\\
|\mathbf{C}| &= \sqrt{C_x^2 + C_y^2}\\
&= 16.9\text{ km}\\
\text{direction} &= \tan^{-1}(13.3/10.4)\\
&= 52^\circ\text{ north of east}
\end{aligned}$$

Incorrect solution #4:
(same notation as above)

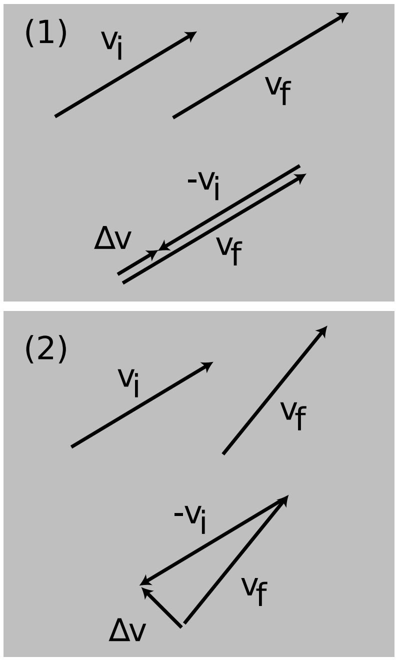
$$\begin{aligned}
A_x &= (5.0\text{ km})\cos 45^\circ = 3.5\text{ km}\\
B_x &= -(12.0\text{ km})\cos 55^\circ = -6.9\text{ km}\\
A_y &= (5.0\text{ km})\sin 45^\circ = 3.5\text{ km}\\
B_y &= -(12.0\text{ km})\sin 55^\circ = -9.8\text{ km}\\
C_x &= A_x + B_x\\
&= -3.4\text{ km}\\
C_y &= A_y + B_y\\
&= -6.3\text{ km}\\
|\mathbf{C}| &= \sqrt{C_x^2 + C_y^2}\\
&= 7.2\text{ km}\\
\text{direction} &= \tan^{-1}(-6.3/-3.4)\\
&= 62^\circ\text{ north of east}
\end{aligned}$$

(continued)

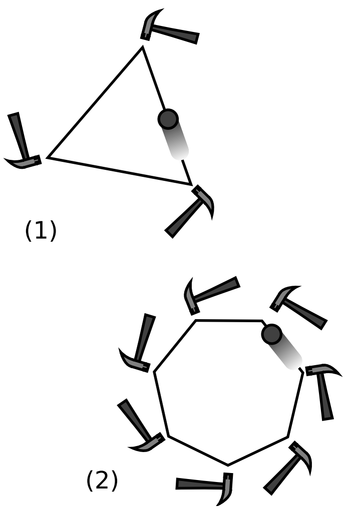
Incorrect solution #5:
(same notation as above)

$$\begin{aligned}
A_x &= (5.0\text{ km})\cos 45^\circ = 3.5\text{ km} \\
B_x &= -(12.0\text{ km})\sin 55^\circ = -9.8\text{ km} \\
A_y &= (5.0\text{ km})\sin 45^\circ = 3.5\text{ km} \\
B_y &= -(12.0\text{ km})\cos 55^\circ = -6.9\text{ km} \\
C_x &= A_x + B_x \\
&= -6.3\text{ km} \\
C_y &= A_y + B_y \\
&= -3.4\text{ km}
\end{aligned}$$

$$\begin{aligned}
|\mathbf{C}| &= \sqrt{C_x^2 + C_y^2} \\
&= 7.2\text{ km} \\
\text{direction} &= \tan^{-1}(-3.4/-6.3) \\
&= 28^\circ\text{ north of east}
\end{aligned}$$



w / Visualizing the acceleration vector.



x / The heptagon, 2, is a better approximation to a circle than the triangle, 1. To make an infinitely good approximation to circular motion, we would need to use an infinitely large number of infinitesimal taps, which would amount to a steady inward force.

3.4.4 Calculus with vectors

Differentiation

In one dimension, we define the velocity as the derivative of the position with respect to time, and we can think of the derivative as what we get when we calculate $\Delta x/\Delta t$ for very short time intervals. The quantity $\Delta x = x_f - x_i$ is calculated by subtraction. In three dimensions, x becomes $\mathbf{r}$, and the $\Delta\mathbf{r}$ vector is calculated by *vector* subtraction, $\Delta\mathbf{r} = \mathbf{r}_f - \mathbf{r}_i$. Vector subtraction is defined component by component, so when we take the derivative of a vector, this means we end up taking the derivative component by component,

$$v_x = \frac{\mathrm{d}x}{\mathrm{d}t}, \quad v_y = \frac{\mathrm{d}y}{\mathrm{d}t}, \quad v_z = \frac{\mathrm{d}z}{\mathrm{d}t}$$

or

$$\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}t}\hat{\mathbf{x}} + \frac{\mathrm{d}y}{\mathrm{d}t}\hat{\mathbf{y}} + \frac{\mathrm{d}z}{\mathrm{d}t}\hat{\mathbf{z}} \qquad .$$

All of this reasoning applies equally well to any derivative of a vector, so for instance we can take the second derivative,

$$a_x = \frac{\mathrm{d}v_x}{\mathrm{d}t}, \quad a_y = \frac{\mathrm{d}v_y}{\mathrm{d}t}, \quad a_z = \frac{\mathrm{d}v_z}{\mathrm{d}t}$$

or

$$\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = \frac{\mathrm{d}v_x}{\mathrm{d}t}\hat{\mathbf{x}} + \frac{\mathrm{d}v_y}{\mathrm{d}t}\hat{\mathbf{y}} + \frac{\mathrm{d}v_z}{\mathrm{d}t}\hat{\mathbf{z}} \qquad .$$

A counterintuitive consequence of this is that the acceleration vector does not need to be in the same direction as the motion. The velocity vector points in the direction of motion, but by Newton's second law, $\mathbf{a} = \mathbf{F}/m$, the acceleration vector points in the same direction as the force, not the motion. This is easiest to understand if we take velocity vectors from two different moments in the motion, and visualize subtracting them graphically to make a $\Delta\mathbf{v}$ vector. The direction of the $\Delta\mathbf{v}$ vector tells us the direction of the acceleration vector as well, since the derivative $\mathrm{d}\mathbf{v}/\mathrm{d}t$ can be approximated as $\Delta\mathbf{v}/\Delta t$. As shown in figure w/1, a change in the magnitude of the velocity vector implies an acceleration that is in the direction of motion. A change in the direction of the velocity vector produces an acceleration perpendicular to the motion, w/2.

Circular motion — *example 59*

▷ An object moving in a circle of radius r in the x-y plane has

$$x = r\cos\omega t \qquad \text{and}$$
$$y = r\sin\omega t \qquad ,$$

where ω is the number of radians traveled per second, and the positive or negative sign indicates whether the motion is clockwise or counterclockwise. What is its acceleration?

▷ The components of the velocity are

$$v_x = -\omega r\sin\omega t \qquad \text{and}$$
$$v_y = \quad \omega r\cos\omega t \qquad ,$$

and for the acceleration we have

$$a_x = -\omega^2 r \cos \omega t \qquad \text{and}$$
$$a_y = -\omega^2 r \sin \omega t \qquad .$$

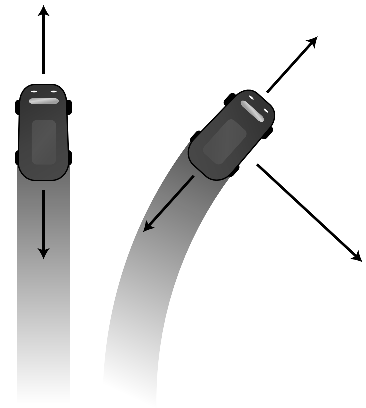
The acceleration vector has cosines and sines in the same places as the **r** vector, but with minus signs in front, so it points in the opposite direction, i.e. toward the center of the circle. By Newton's second law, **a**=**F**/m, this shows that the force must be inward as well; without this force, the object would fly off straight.

The magnitude of the acceleration is

$$\begin{aligned} |\mathbf{a}| &= \sqrt{a_x^2 + a_y^2} \\ &= \omega^2 r \qquad . \end{aligned}$$

It makes sense that ω is squared, since reversing the sign of ω corresponds to reversing the direction of motion, but the acceleration is toward the center of the circle, regardless of whether the motion is clockwise or counterclockwise. This result can also be rewritten in the form

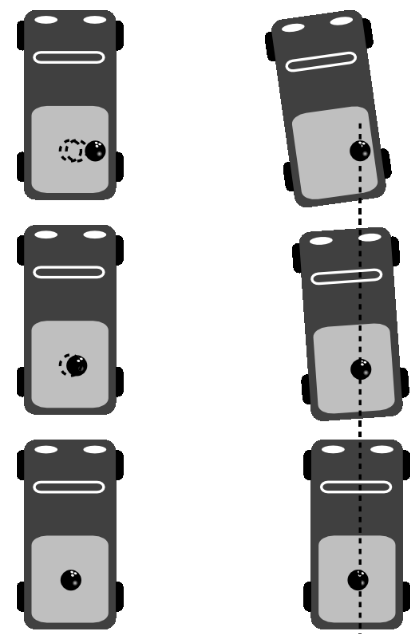
$$|\mathbf{a}| = \frac{|\mathbf{v}|^2}{|\mathbf{r}|} \qquad .$$



y / The total force in the forward-backward direction is zero in both cases.

The results of example 59 are important and useful, but counterintuitive as well. Until Newton, physicists and laypeople alike had assumed that the planets would need a force to push them *forward* in their orbits. Figure x may help to make it more plausible that only an inward force is required. A forward force might be needed in order to cancel out a backward force such as friction, y, but the total force in the forward-backward direction needs to be exactly zero for constant-speed motion.

When you are in a car undergoing circular motion, there is also a strong illusion of an *outward* force. But what object could be making such a force? The car's seat makes an inward force on you, not an outward one. There is no object that could be exerting an outward force on your body. In reality, this force is an illusion that comes from our brain's intuitive efforts to interpret the situation within a noninertial frame of reference. As shown in figure z, we can describe everything perfectly well in an inertial frame of reference, such as the frame attached to the sidewalk. In such a frame, the bowling ball goes straight because there is *no* force on it. The wall of the truck's bed hits the ball, not the other way around.



z / There is no outward force on the bowling ball, but in the noninertial frame it seems like one exists.

Integration

An integral is really just a sum of many infinitesimally small terms. Since vector addition is defined in terms of addition of the components, an integral of a vector quantity is found by doing integrals component by component.

Projectile motion *example 60*

▹ Find the motion of an object whose acceleration vector is constant, for instance a projectile moving under the influence of gravity.

▹ We integrate the acceleration to get the velocity, and then integrate the velocity to get the position as a function of time. Doing this to the x component of the acceleration, we find

$$\begin{aligned} x &= \int \left(\int a_x \, \mathrm{d}t \right) \mathrm{d}t \\ &= \int (a_x t + v_{x\mathrm{o}}) \, \mathrm{d}t \quad , \end{aligned}$$

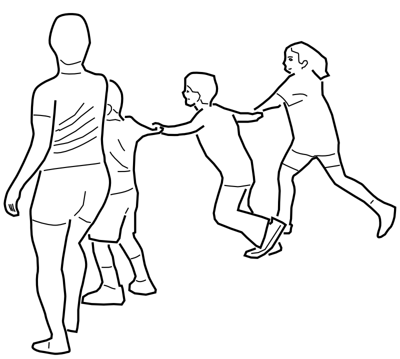
where $v_{x\mathrm{o}}$ is a constant of integration, and

$$x = \frac{1}{2} a_x t^2 + v_{x\mathrm{o}} t + x_\mathrm{o} \quad .$$

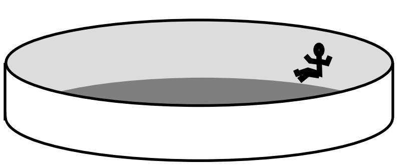
Similarly, y=$(1/2)a_y t^2 + v_{y\mathrm{o}} t + y_\mathrm{o}$ and $z = (1/2)a_z t^2 + v_{z\mathrm{o}} t + z_\mathrm{o}$. Once one has gained a little confidence, it becomes natural to do the whole thing as a single vector integral,

$$\begin{aligned} \mathbf{r} &= \int \left(\int \mathbf{a} \, \mathrm{d}t \right) \mathrm{d}t \\ &= \int (\mathbf{a}t + \mathbf{v}_\mathrm{o}) \, \mathrm{d}t \\ &= \frac{1}{2}\mathbf{a}t^2 + \mathbf{v}_\mathrm{o} t + \mathbf{r}_\mathrm{o} \quad , \end{aligned}$$

where now the constants of integration are vectors.



aa / Discussion question A.



ab / Discussion question E.

Discussion Questions

A In the game of crack the whip, a line of people stand holding hands, and then they start sweeping out a circle. One person is at the center, and rotates without changing location. At the opposite end is the person who is running the fastest, in a wide circle. In this game, someone always ends up losing their grip and flying off. Suppose the person on the end loses her grip. What path does she follow as she goes flying off? (Assume she is going so fast that she is really just trying to put one foot in front of the other fast enough to keep from falling; she is not able to get any significant horizontal force between her feet and the ground.)

B Suppose the person on the outside is still holding on, but feels that she may loose her grip at any moment. What force or forces are acting on her, and in what directions are they? (We are not interested in the vertical forces, which are the earth's gravitational force pulling down, and the ground's normal force pushing up.)

C Suppose the person on the outside is still holding on, but feels that she may loose her grip at any moment. What is wrong with the following analysis of the situation? "The person whose hand she's holding exerts an inward force on her, and because of Newton's third law, there's an equal and opposite force acting outward. That outward force is the one she feels throwing her outward, and the outward force is what might make her go flying off, if it's strong enough."

D If the only force felt by the person on the outside is an inward force, why doesn't she go straight in?

E In the amusement park ride shown in the figure, the cylinder spins faster and faster until the customer can pick her feet up off the floor without falling. In the old Coney Island version of the ride, the floor actually dropped out like a trap door, showing the ocean below. (There is also a version in which the whole thing tilts up diagonally, but we're discussing the version that stays flat.) If there is no outward force acting on her, why does she stick to the wall? Analyze all the forces on her.

F What is an example of circular motion where the inward force is a normal force? What is an example of circular motion where the inward force is friction? What is an example of circular motion where the inward force is the sum of more than one force?

G Does the acceleration vector always change continuously in circular motion? The velocity vector?

H A certain amount of force is needed to provide the acceleration of circular motion. What if were are exerting a force perpendicular to the direction of motion in an attempt to make an object trace a circle of radius r, but the force isn't as big as $m|\mathbf{v}|^2/r$?

I Suppose a rotating space station is built that gives its occupants the illusion of ordinary gravity. What happens when a person in the station lets go of a ball? What happens when she throws a ball straight "up" in the air (i.e. towards the center)?

3.4.5 The dot product

How would we generalize the mechanical work equation $\mathrm{d}E = F\mathrm{d}x$ to three dimensions? Energy is a scalar, but force and distance are vectors, so it might seem at first that the kind of "magic-wand" generalization discussed on page 148 failed here, since we don't know of any way to multiply two vectors together to get a scalar. Actually, this is Nature giving us a hint that there is such a multiplication operation waiting for us to invent it, and since Nature is simple, we can be assured that this operation will work just fine in any situation where a similar generalization is required.

How should this operation be defined? Let's consider what we would get by performing this operation on various combinations of the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$. The conventional notation for the operation is to put a dot, $\cdot$, between the two vectors, and the operation is therefore called the *dot product*. Rotational invariance requires that we handle the three coordinate axes in the same way, without giving special treatment to any of them, so we must have $\hat{\mathbf{x}}\cdot\hat{\mathbf{x}} = \hat{\mathbf{y}}\cdot\hat{\mathbf{y}} = \hat{\mathbf{z}}\cdot\hat{\mathbf{z}}$ and $\hat{\mathbf{x}}\cdot\hat{\mathbf{y}} = \hat{\mathbf{y}}\cdot\hat{\mathbf{z}} = \hat{\mathbf{z}}\cdot\hat{\mathbf{x}}$. This is supposed to be a way of generalizing ordinary multiplication, so for consistency with the property $1\times 1 = 1$ of ordinary numbers, the result of multiplying a magnitude-one vector by itself had better be the scalar 1, so $\hat{\mathbf{x}}\cdot\hat{\mathbf{x}} = \hat{\mathbf{y}}\cdot\hat{\mathbf{y}} = \hat{\mathbf{z}}\cdot\hat{\mathbf{z}} = 1$. Furthermore, there is no way to satisfy rotational invariance unless we define the mixed products to be zero, $\hat{\mathbf{x}}\cdot\hat{\mathbf{y}} = \hat{\mathbf{y}}\cdot\hat{\mathbf{z}} = \hat{\mathbf{z}}\cdot\hat{\mathbf{x}} = 0$; for example, a 90-degree rotation of our frame of reference about the z axis reverses the sign of $\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}$, but rotational invariance requires that $\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}$ produce the same result either way, and zero is the only number that stays the same when we reverse its sign. Establishing these six products of unit vectors suffices to define the operation in general, since any two vectors that we want to multiply can be broken down into components, e.g. $(2\hat{\mathbf{x}}+3\hat{\mathbf{z}})\cdot\hat{\mathbf{z}} = 2\hat{\mathbf{x}}\cdot\hat{\mathbf{z}}+3\hat{\mathbf{z}}\cdot\hat{\mathbf{z}} = 0+3 = 3$. Thus by requiring rotational invariance and consistency with multiplication of ordinary numbers, we find that there is only one possible way to define a multiplication operation on two vectors that gives a scalar as the result.[16] The dot product has all of the properties we normally associate with multiplication, except that there is no "dot division."

Dot product in terms of components *example 61*

If we know the components of any two vectors **b** and **c**, we can find their dot product:

$$\begin{aligned}\mathbf{b}\cdot\mathbf{c} &= \left(b_x\hat{\mathbf{x}}+b_y\hat{\mathbf{y}}+b_z\hat{\mathbf{z}}\right)\cdot\left(c_x\hat{\mathbf{x}}+c_y\hat{\mathbf{y}}+c_z\hat{\mathbf{z}}\right)\\ &= b_xc_x+b_yc_y+b_zc_z \qquad .\end{aligned}$$

[16] There is, however, a different operation, discussed in the next chapter, which multiplies two vectors to give a vector.

Magnitude expressed with a dot product *example 62*

If we take the dot product of any vector **b** with itself, we find

$$\begin{aligned}\mathbf{b}\cdot\mathbf{b} &= \left(b_x\hat{\mathbf{x}}+b_y\hat{\mathbf{y}}+b_z\hat{\mathbf{z}}\right)\cdot\left(b_x\hat{\mathbf{x}}+b_y\hat{\mathbf{y}}+b_z\hat{\mathbf{z}}\right)\\ &= b_x^2+b_y^2+b_z^2 \qquad ,\end{aligned}$$

so its magnitude can be expressed as

$$|\mathbf{b}| = \sqrt{\mathbf{b}\cdot\mathbf{b}}$$

We will often write b^2 to mean $\mathbf{b}\cdot\mathbf{b}$, when the context makes it clear what is intended. For example, we could express kinetic energy as $(1/2)m|\mathbf{v}|^2$, $(1/2)m\mathbf{v}\cdot\mathbf{v}$, or $(1/2)mv^2$. In the third version, nothing but context tells us that v really stands for the magnitude of some vector **v**.

Geometric interpretation *example 63*

In figure ac, vectors **a**, **b**, and **c** represent the sides of a triangle, and $\mathbf{a}=\mathbf{b}+\mathbf{c}$. The law of cosines gives

$$|\mathbf{c}|^2 = |\mathbf{a}|^2+|\mathbf{b}|^2-2|\mathbf{a}||\mathbf{b}|\cos\theta \qquad .$$

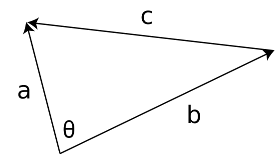
Using the result of example 62, we can also write this as

$$\begin{aligned}|\mathbf{c}|^2 &= \mathbf{c}\cdot\mathbf{c}\\ &= (\mathbf{a}-\mathbf{b})\cdot(\mathbf{a}-\mathbf{b})\\ &= \mathbf{a}\cdot\mathbf{a}+\mathbf{b}\cdot\mathbf{b}-2\mathbf{a}\cdot\mathbf{b} \qquad .\end{aligned}$$

Matching up terms in these two expressions, we find

$$\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta \qquad ,$$

which is a geometric interpretation for the dot product.



ac / The geometric interpretation of the dot product.

The result of the preceding example is very useful. It gives us a way to find the angle between two vectors if we know their components. It can be used to show that the dot product of any two perpendicular vectors is zero. It also leads to a nifty proof that the dot product is rotationally invariant — up until now I've only proved that if a rotationally invariant product exists, the dot product is it — because angles and lengths aren't affected by a rotation, so the right side of the equation is rotationally invariant, and therefore so is the left side.

I introduced the whole discussion of the dot product by way of generalizing the equation $\mathrm{d}E = F\mathrm{d}x$ to three dimensions. In terms of a dot product, we have

$$\mathrm{d}E = \mathbf{F}\cdot\mathrm{d}\mathbf{r} \qquad .$$

If **F** is a constant, integrating both sides gives

$$\Delta E = \mathbf{F}\cdot\Delta\mathbf{r} \qquad .$$

(If that step seemed like black magic, try writing it out in terms of components.) If the force is perpendicular to the motion, as in figure ad, then the work done is zero. The pack horse is doing work within its own body, but is not doing work on the pack.

ad / Breaking trail, by Walter E. Bohl. The pack horse is not doing any work on the pack, because the pack is moving in a horizontal line at constant speed, and therefore there is no kinetic or gravitational energy being transferred into or out of it.

Pushing a lawnmower *example 64*

▷ I push a lawnmower with a force **F**=(110 N)$\hat{\mathbf{x}}$ − (40 N)$\hat{\mathbf{y}}$, and the total distance I travel is (100 m)$\hat{\mathbf{x}}$. How much work do I do?

▷ The dot product is 11000 N·m = 11000 J.

A good application of the dot product is to allow us to write a simple, streamlined proof of separate conservation of the momentum components. (You can skip the proof without losing the continuity of the text.) The argument is a generalization of the one-dimensional proof on page 89, and makes the same assumption about the type of system of particles we're dealing with. The kinetic energy of one of the particles is $(1/2)m\mathbf{v}\cdot\mathbf{v}$, and when we transform into a different frame of reference moving with velocity $\mathbf{u}$ relative to the original frame, the one-dimensional rule $v \rightarrow v+u$ turns into vector addition, $\mathbf{v} \rightarrow \mathbf{v}+\mathbf{u}$. In the new frame of reference, the kinetic energy is $(1/2)m(\mathbf{v}+\mathbf{u})\cdot(\mathbf{v}+\mathbf{u})$. For a system of n particles, we have

$$\begin{aligned} K &= \sum_{j=1}^{n} \frac{1}{2} m_j (\mathbf{v}_j + \mathbf{u})\cdot(\mathbf{v}_j+\mathbf{u}) \\ &= \frac{1}{2}\left[\sum_{j=1}^{n} m_j \mathbf{v}_j\cdot\mathbf{v}_j + 2\sum_{j=1}^{n} m_j\mathbf{v}_j\cdot\mathbf{u} + \sum_{j=1}^{n} m_j \mathbf{u}\cdot\mathbf{u}\right] \qquad . \end{aligned}$$

As in the proof on page 89, the first sum is simply the total kinetic energy in the original frame of reference, and the last sum is a constant, which has no effect on the validity of the conservation law. The middle sum can be rewritten as

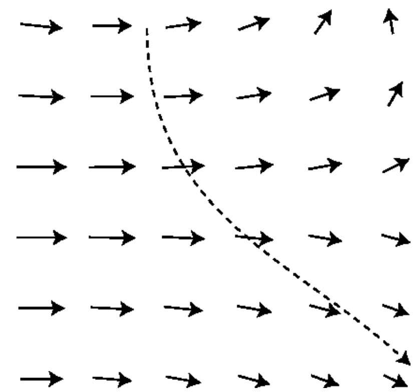
$$\begin{aligned} 2\sum_{j=1}^{n} m_j\mathbf{v}_j\cdot\mathbf{u} &= 2\ \mathbf{u}\cdot\sum_{j=1}^{n} m_j\mathbf{v}_j \\ &= 2\ \mathbf{u}\cdot\sum_{j=1}^{n}\mathbf{p}_j \qquad , \end{aligned}$$

so the only way energy can be conserved for all values of $\mathbf{u}$ is if the vector sum of the momenta is conserved as well.

3.4.6 Gradients and line integrals (optional)

This subsection introduces a little bit of vector calculus. It can be omitted without loss of continuity, but the techniques will be needed in our study of electricity and magnetism, and it may be helpful to be exposed to them in easy-to-visualize mechanical contexts before applying them to invisible electrical and magnetic phenomena.

In physics we often deal with fields of force, meaning situations where the force on an object depends on its position. For instance, figure ae could represent a map of the trade winds affecting a sailing ship, or a chart of the gravitational forces experienced by a space probe entering a double-star system. An object moving under the influence of this force will not necessarily be moving in the same direction as the force at every moment. The sailing ship can tack against the wind, due to the force from the water on the keel. The space probe, if it entered from the top of the diagram at high speed, would start to curve around to the right, but its inertia would carry it forward, and it wouldn't instantly swerve to match the direction of the gravitational force. For convenience, we've defined the gravitational field, $\mathbf{g}$, as the force *per unit mass*, but that trick only leads to a simplification because the gravitational force on an object is proportional to its mass. Since this subsection is meant to apply to any kind of force, we'll discuss everything in terms of the actual force vector, $\mathbf{F}$, in units of newtons.


ae / An object moves through a field of force.

If an object moves through the field of force along some curved path from point $\mathbf{r}_1$ to point $\mathbf{r}_2$, the force will do a certain amount of work on it. To calculate this work, we can break the path up into infinitesimally short segments, find the work done along each segment, and add them all up. For an object traveling along a nice straight x axis, we use the symbol $\mathrm{d}x$ to indicate the length of any infinitesimally short segment. In three dimensions, moving along a curve, each segment is a tiny vector $\mathrm{d}\mathbf{r} = \hat{\mathbf{x}}\mathrm{d}x + \hat{\mathbf{y}}\mathrm{d}y + \hat{\mathbf{z}}\mathrm{d}z$. The work theorem can be expressed as a dot product, so the work done along a segment is $\mathbf{F}\cdot\mathrm{d}\mathbf{r}$. We want to integrate this, but we don't know how to integrate with respect to a variable that's a vector, so let's define a variable s that indicates the distance traveled so far along the curve, and integrate with respect to it instead. The expression $\mathbf{F}\cdot\mathrm{d}\mathbf{r}$ can be rewritten as $|\mathbf{F}|\,|\mathrm{d}\mathbf{r}|\,\cos\theta$, where θ is the angle between $\mathbf{F}$ and $\mathrm{d}\mathbf{r}$. But $|\mathrm{d}\mathbf{r}|$ is simply $\mathrm{d}s$, so the amount of work done becomes

$$\Delta E = \int_{\mathbf{r}_1}^{\mathbf{r}_2} |\mathbf{F}| \cos\theta \; \mathrm{d}s \qquad .$$

Both $\mathbf{F}$ and θ are functions of s. As a matter of notation, it's cumbersome to have to write the integral like this. Vector notation was designed to eliminate this kind of drudgery. We therefore define the line integral

$$\int_C \mathbf{F} \cdot \mathrm{d}\mathbf{r}$$

as a way of notating this type of integral. The 'C' refers to the curve along which the object travels. If we don't know this curve then we typically can't evaluate the line integral just by knowing the initial and final positions $\mathbf{r}_1$ and $\mathbf{r}_2$.

The basic idea of calculus is that integration undoes differentiation, and vice-versa. In one dimension, we could describe an interaction either in terms of a force or in terms of an interaction energy. We could integrate force with respect to position to find minus the energy, or we could find the force by taking minus the derivative of the energy. In the line integral, position is represented by a vector. What would it mean to take a derivative with respect to a vector? The correct way to generalize the derivative $\mathrm{d}U/\mathrm{d}x$ to three dimensions is to replace it with the following vector,

$$\frac{\mathrm{d}U}{\mathrm{d}x}\hat{\mathbf{x}} + \frac{\mathrm{d}U}{\mathrm{d}y}\hat{\mathbf{y}} + \frac{\mathrm{d}U}{\mathrm{d}z}\hat{\mathbf{z}} \qquad ,$$

called the *gradient* of U, and written with an upside-down delta[17] like this, ∇U. Each of these three derivatives is really what's known as a partial derivative. What that means is that when you're differentiating U with respect to x, you're supposed to treat y and z and constants, and similarly when you do the other two derivatives. To emphasize that a derivative is a partial derivative, it's customary to write it using the symbol ∂ in place of the differential d's. Putting all this notation together, we have

$$\nabla U = \frac{\partial U}{\partial x}\hat{\mathbf{x}} + \frac{\partial U}{\partial y}\hat{\mathbf{y}} + \frac{\partial U}{\partial z}\hat{\mathbf{z}} \qquad \text{[definition of the gradient]} \qquad .$$

The gradient looks scary, but it has a very simple physical interpretation. It's a vector that points in the direction in which U is increasing most rapidly, and it tells you how rapidly U is increasing in that direction. For instance, sperm cells in plants and animals find the egg cells by traveling in the direction of the gradient of the concentration of certain hormones. When they reach the location of the strongest hormone concentration, they find their destiny. In terms of the gradient, the force corresponding to a given interaction energy is $\mathbf{F} = -\nabla U$.

Force exerted by a spring — *example 65*

In one dimension, Hooke's law is $U = (1/2)kx^2$. Suppose we tether one end of a spring to a post, but it's free to stretch and swing around in a plane. Let's say its equilibrium length is zero, and let's choose the origin of our coordinate system to be at the post. Rotational invariance

[17] The symbol ∇ is called a "nabla." Cool word!

requires that its energy only depend on the magnitude of the **r** vector, not its direction, so in two dimensions we have $U = (1/2)k|\mathbf{r}|^2 = (1/2)k\left(x^2+y^2\right)$. The force exerted by the spring is then

$$\begin{aligned}\mathbf{F} &= -\nabla U \\ &= -\frac{\partial U}{\partial x}\hat{\mathbf{x}} - \frac{\partial U}{\partial y}\hat{\mathbf{y}} \\ &= -kx\hat{\mathbf{x}} - ky\hat{\mathbf{y}} \qquad .\end{aligned}$$

The magnitude of this force vector is $k|\mathbf{r}|$, and its direction is toward the origin.

This chapter is summarized on page 724. Notation and terminology are tabulated on pages 718-719.

Problems

The symbols √, , etc. are explained on page 172.

1 Derive a formula expressing the kinetic energy of an object in terms of its momentum and mass. √

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. √
(b) Find the bullet's final momentum. √
(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. √

4 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?
▷ Answer, p. 711

5 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.]

6 A car is normally capable of an acceleration of 3 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.]

7 (a) Let T be the maximum tension that the 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, what is the maximum mass that the car can have, including passengers, if the cable is not to break?
(b) Interpret the equation you derived in the special cases of a=0 and of a downward acceleration of magnitude g.

8 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) Plug numbers into your equation from part a, using m=2300 kg, F_{air}=27000 N, and t=4.0 s. √

9 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.)

10 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? ✓

11 If a big truck and a VW bug collide head-on, which will be acted on by the greater force? Which will have the greater acceleration?

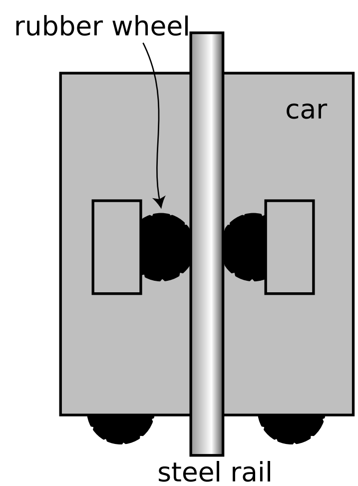
12 The earth is attracted to an object with a force equal and opposite to the force of the earth on the object. If this is true, why is it that when you drop an object, the earth does not have an acceleration equal and opposite to that of the object?

13 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.

14 Today's tallest buildings are really not that much taller than the tallest buildings of the 1940s. The main 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.)

(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?

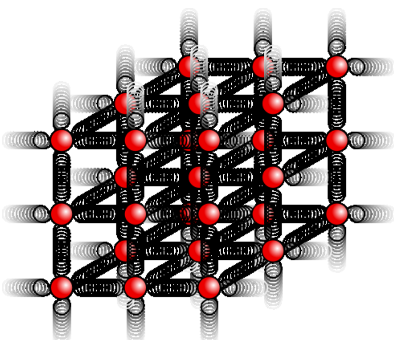


Problem 14.

15 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? ▷ Hint, p. 704
(b) Interpret your answer in the special cases of $M = 0$ and $M = \infty$.

16 Explain why it wouldn't make sense to have kinetic friction be stronger than static friction.

17 (a) Using the solution of problem 37 on page 85, 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, p. 713



Problem 18

18 This problem depends on the results of problems problem 37 on page 85 and problem 17 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 17 and its solution.
(b) Estimate k using the numerical data given in problem 17.
(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?

19 Many fish have an organ known as a swim bladder, an air-filled cavity whose main purpose is to control the fish's buoyancy and allow it to keep from rising or sinking without having to use its muscles. In some fish, however, the swim bladder (or a small extension of it) is linked to the ear and serves the additional purpose of amplifying sound waves. For a typical fish having such an anatomy, the bladder has a resonant frequency of 300 Hz, the bladder's Q is 3, and the maximum amplification is about a factor of 100 in energy. Over what range of frequencies would the amplification be at least a factor of 50?

20 An oscillator with sufficiently strong damping has its maximum response at $\omega = 0$. Using the result derived on page 693, find the value of Q at which this behavior sets in.

▷ Hint, p. 704 ▷ Answer, p. 711

21 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

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

Determine these three quantities numerically using the result derived on page 693, and compare with the approximations.

22 The apparatus in figure d on page 20 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. ▷ Answer, p. 712
(b) Show that even if their estimate of Q was wildly wrong, it wouldn't have affected this result.

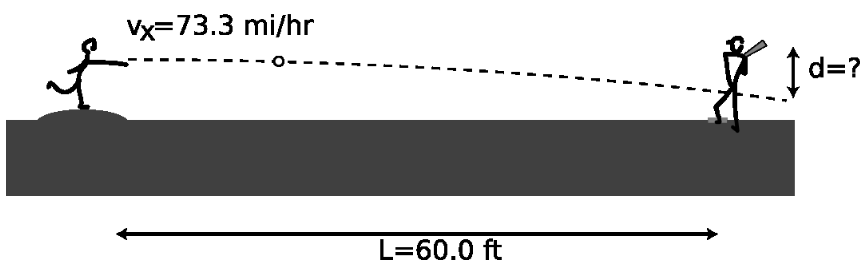
23 (a) A ball is thrown straight up with velocity v. Find an equation for the height to which it rises.
(b) Generalize your equation for a ball thrown at an angle θ above horizontal.

24 At the Salinas Lettuce Festival Parade, Miss Lettuce of 1996 drops her bouquet while riding on a float. Compare the shape of its trajectory as seen by her to the shape seen by one of her admirers standing on the sidewalk.

25 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?

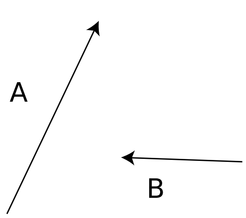
26 A baseball pitcher throws a pitch clocked at v_x=73.3 mi/h. He throws horizontally. By what amount, d, does the ball drop by the time it reaches home plate, L=60.0 ft away?
(a) First find a symbolic answer in terms of L, v_x, and g.
(b) Plug in and find a numerical answer. Express your answer in units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



Problem 26.

27 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.
▷ Answer, p. 712
(b) Interpret your equation in the cases of θ=0 and $\theta = 90\,°$.
(c) Find the angle that gives the maximum range.
▷ Answer, p. 712

28 This problem uses numerical methods to extend the analysis in problem 27 to include air friction. For a game played at sea level, the force due to air friction is approximately $(5 \times 10^{-3}\ \mathrm{N{\cdot}s^2/m^2})v^2$, 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. ▷ Answer, p. 712
(b) How much farther would the ball fly at the Colorado Rockies' stadium, where the thinner air gives 18 percent less air friction?
▷ Answer, p. 712



Problem 30.

29 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. ✓

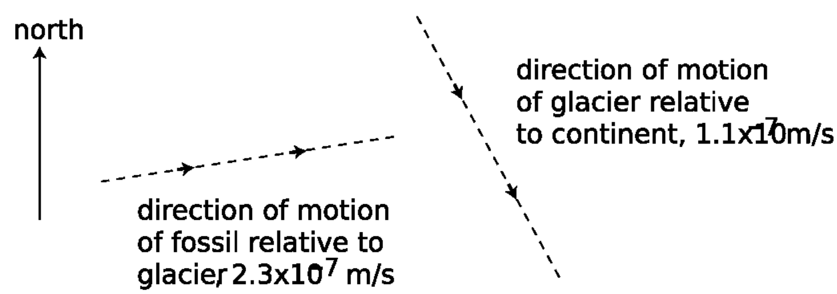
30 Here are two vectors. Graphically calculate $\mathbf{A}+\mathbf{B}$, $\mathbf{A}-\mathbf{B}$, $\mathbf{B}-\mathbf{A}$, $-2\mathbf{B}$, and $\mathbf{A}-2\mathbf{B}$. No numbers are involved.

31 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.

√

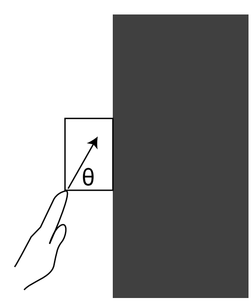
32 Is it possible for a helicopter to have an acceleration due east and a velocity due west? If so, what would be going on? If not, why not?

33 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. √



Problem 33.

34 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.) √



Problem 36

35 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.)

36 Your hand presses a block of mass m against a wall with a force F_h acting at an angle θ. Find the minimum and maximum possible values of $|F_h|$ 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.

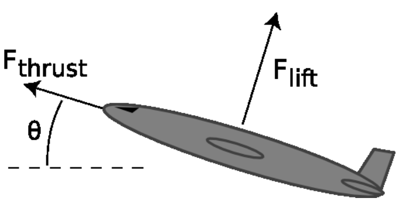
37 A skier of mass m is coasting down a slope inclined at an angle θ compared to horizontal. Assume for simplicity that the standard treatment of kinetic friction given in the text is appropriate here, although a soft and wet surface actually behaves a little differently. The coefficient of kinetic friction 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, θ, μ_k, and b. (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 $\theta < \theta_{min}$.

38 A gun is aimed horizontally to the west, and fired at $t = 0$. The bullet's position vector as a function of time is $\mathbf{r} = b\hat{\mathbf{x}} + ct\hat{\mathbf{x}} + dt^2\hat{\mathbf{x}}$, where b, c, and d are 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.

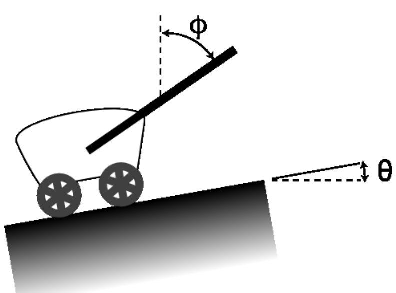
(c) Give physical interpretations of b, c, d, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$.



Problem 40

39 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, p. 713

40 A cargo plane has taken off from a tiny airstrip in the Andes, and is climbing at constant speed, at an angle of $\theta = 17°$ 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, p. 713



Problem 41

41 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 $\phi = 180° - \theta$. ▷ Solution, p. 714

42 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, p. 715

43 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) Suppose you find that the tape doesn't fly off when the motor is on a low speed, but speeding it up does cause it to fly off. Why would the greater speed change things?

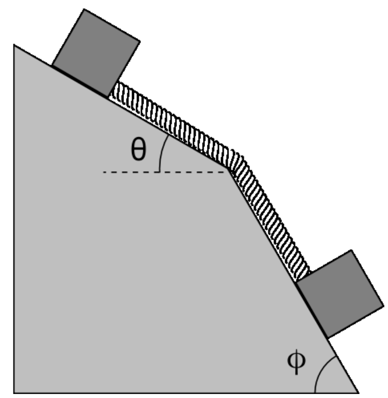
44 Show that the expression $|\mathbf{v}|^2/\mathbf{r}$ has the units of acceleration.

45 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? ✓

46 Find the angle between the following two vectors:

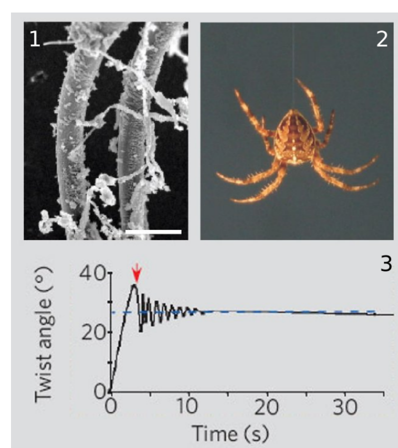
$$\hat{\mathbf{x}} + 2\hat{\mathbf{y}} + 3\hat{\mathbf{z}}$$
$$4\hat{\mathbf{x}} + 5\hat{\mathbf{y}} + 6\hat{\mathbf{z}}$$

▷ Hint, p. 704 ✓



Problem 47.

47 The two blocks shown in the figure have equal mass, m, and the surface is frictionless. What is the tension in the massless rope?



Problem 48.

48 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.

Key to symbols:
easy typical challenging difficult very difficult
$\surd$ An answer check is available at www.lightandmatter.com.

Exercises

Exercise 3A: Force and Motion

Equipment:

- 2-meter pieces of butcher paper
- wood blocks with hooks
- string
- masses to put on top of the blocks to increase friction
- spring scales (preferably calibrated in Newtons)

Suppose a person pushes a crate, sliding it across the floor at a certain speed, and then repeats the same thing but at a higher speed. This is essentially the situation you will act out in this exercise. What do you think is different about her force on the crate in the two situations? Discuss this with your group and write down your hypothesis:

____

1. First you will measure the amount of friction between the wood block and the butcher paper when the wood and paper surfaces are slipping over each other. The idea is to attach a spring scale to the block and then slide the butcher paper under the block while using the scale to keep the block from moving with it. Depending on the amount of force your spring scale was designed to measure, you may need to put an extra mass on top of the block in order to increase the amount of friction. It is a good idea to use long piece of string to attach the block to the spring scale, since otherwise one tends to pull at an angle instead of directly horizontally.

First measure the amount of friction force when sliding the butcher paper as slowly as possible:__________

Now measure the amount of friction force at a significantly higher speed, say 1 meter per second. (If you try to go too fast, the motion is jerky, and it is impossible to get an accurate reading.) __________

Discuss your results. Why are we justified in assuming that the string's force on the block (i.e., the scale reading) is the same amount as the paper's frictional force on the block?

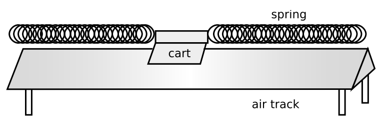
2. Now try the same thing but with the block moving and the paper standing still. Try two different speeds.

Do your results agree with your original hypothesis? If not, discuss what's going on. How does the block "know" how fast to go?

Exercise 3B: Vibrations

Equipment:

- air track and carts of two different masses
- springs
- spring scales



Place the cart on the air track and attach springs so that it can vibrate.

1. Test whether the period of vibration depends on amplitude. Try at least two moderate amplitudes, for which the springs do not go slack, and at least one amplitude that is large enough so that they do go slack.

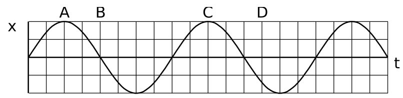
2. Try a cart with a different mass. Does the period change by the expected factor, based on the equation $T=2\pi\sqrt{m/k}$?

3. Use a spring scale to pull the cart away from equilibrium, and make a graph of force versus position. Is it linear? If so, what is its slope?

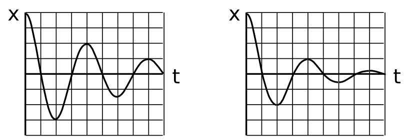
4. Test the equation $T=2\pi\sqrt{m/k}$ numerically.

Exercise 3C: Worksheet on Resonance

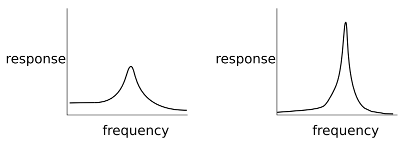
1. Compare the oscillator's energies at A, B, C, and D.



2. Compare the Q values of the two oscillators.



3. Match the x-t graphs in #2 with the amplitude-frequency graphs below.



Exercise 3D: Vectors and Motion

Each diagram on page 177 shows the motion of an object in an $x-y$ plane. Each dot is one location of the object at one moment in time. The time interval from one dot to the next is always the same, so you can think of the vector that connects one dot to the next as a $\mathbf{v}$ vector, and subtract to find $\Delta\mathbf{v}$ vectors.

1. Suppose the object in diagram 1 is moving from the top left to the bottom right. Deduce whatever you can about the force acting on it. Does the force always have the same magnitude? The same direction?

Invent a physical situation that this diagram could represent.

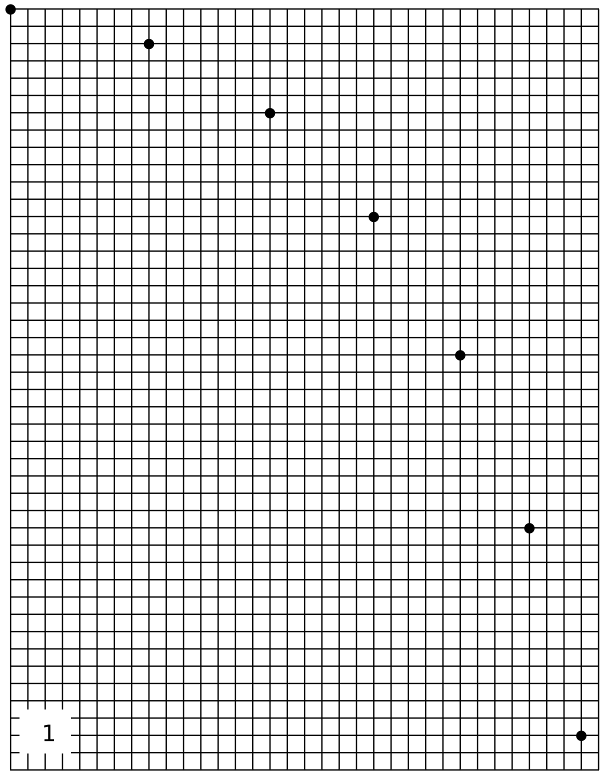
What if you reinterpret the diagram, and reverse the object's direction of motion?

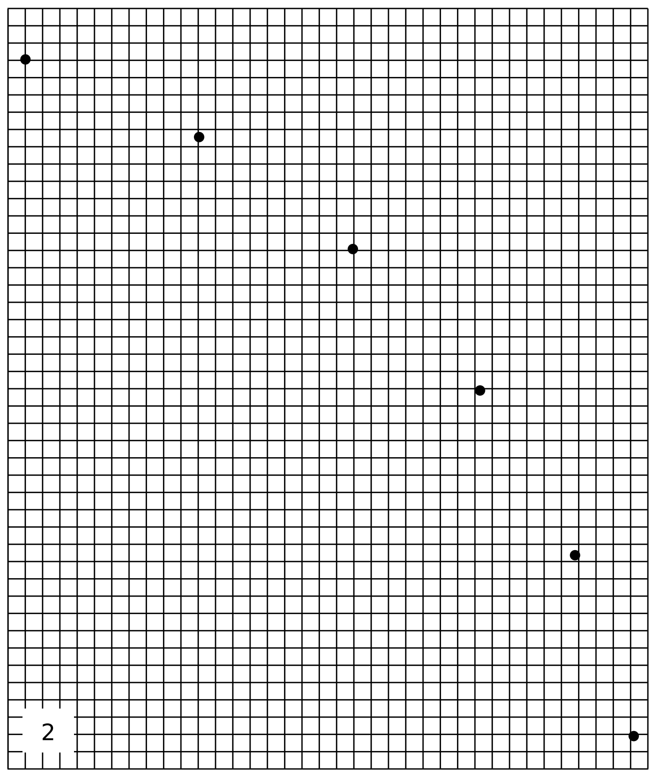
2. What can you deduce about the force that is acting in diagram 2?

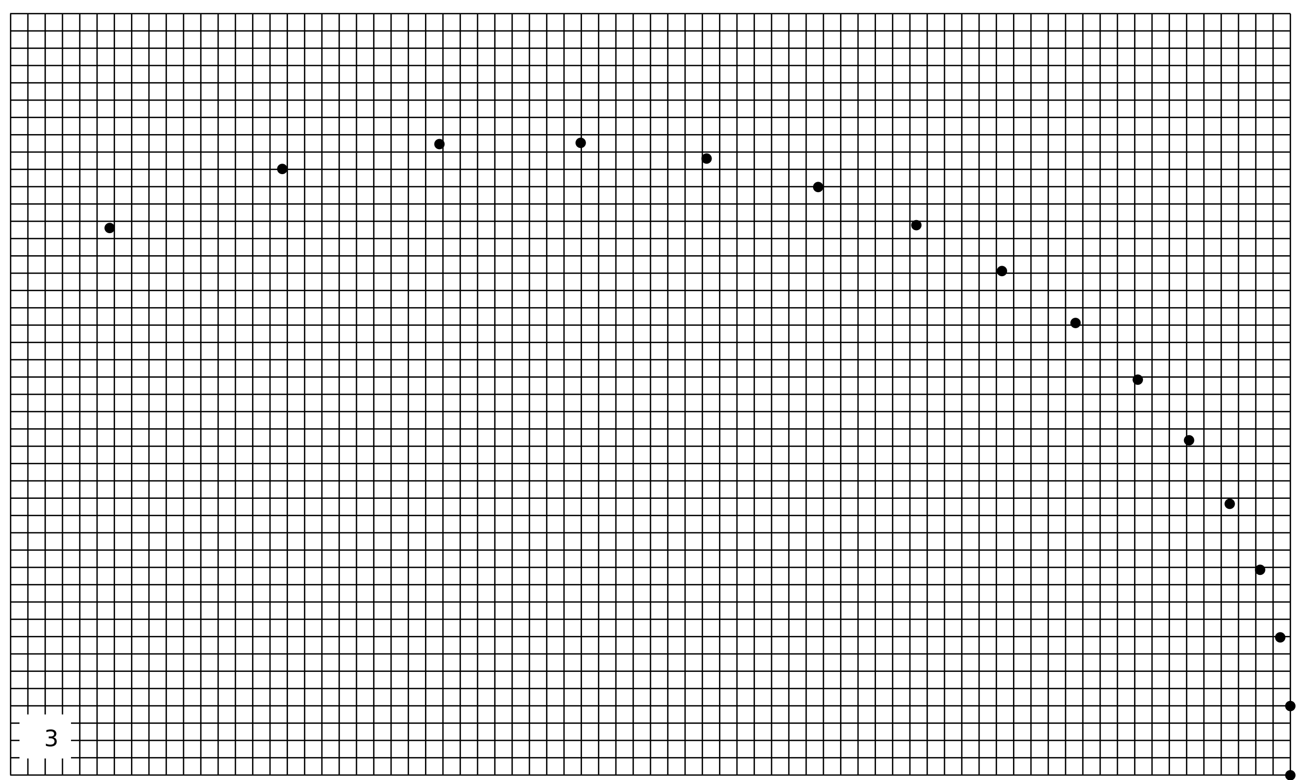
Invent a physical situation that diagram 2 could represent.

3. What can you deduce about the force that is acting in diagram 3?

Invent a physical situation.


1


2


3

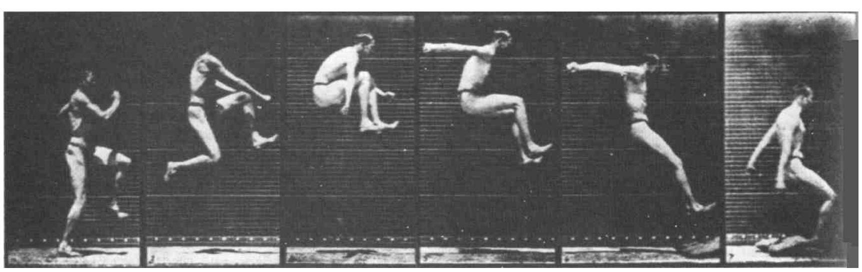
Chapter 4

Conservation of Angular Momentum

4.1 Angular Momentum in Two Dimensions

4.1.1 Angular momentum

"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:

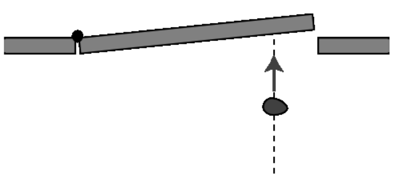
A closed system is involved. Nothing is making an effort to twist the earth, the hydrogen atom, or the high-diver. They are isolated from rotation-changing influences, i.e. they are closed systems.

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.

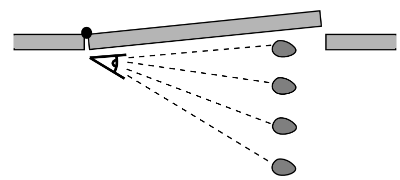


b / An overhead view of a piece of putty being thrown at a door. Even though the putty is neither spinning nor traveling along a curve, we must define it has having some kind of "rotation" because it is able to make the door rotate.

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.

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, 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, 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.

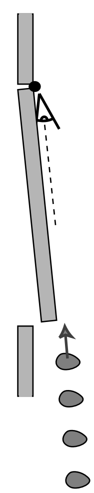


c / As seen by someone standing at the axis, the putty changes its angular position. We therefore define it as having angular momentum.

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,

$$L \propto m \qquad .$$



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.

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,

$$L \propto mv \qquad .$$

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. 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,

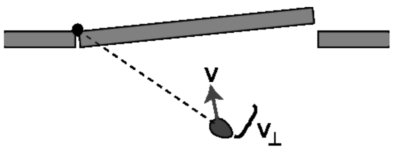
$$L \propto mvr \qquad .$$

We are almost done, but there is one missing ingredient. We know on grounds of symmetry that a putty ball thrown directly

[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.

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,

$$L = mv_\perp r \qquad .$$



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.

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.

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 (kg·m/s)·m, i.e. kg·m^2/s. Summarizing, we have

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

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.

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

A figure skater pulls her arms in. — *example 1*

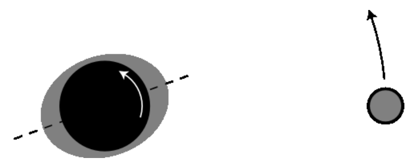
When a figure skater is twirling, there is very little friction between her and the ice, so she is essentially a closed system, and her angular momentum is conserved. If she pulls her arms in, she is decreasing *r* for all the atoms in her arms. It would violate conservation of angular momentum if she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

Earth's slowing rotation and the receding moon *example 2*

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 2 millimeters 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.

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.



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

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.

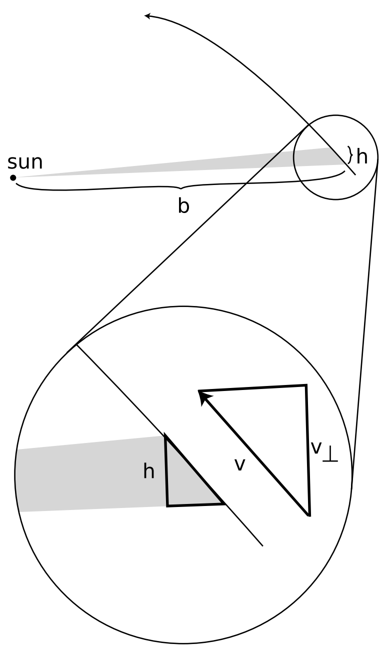
Discussion Questions

A 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.



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:

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

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

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

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

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

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

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

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.

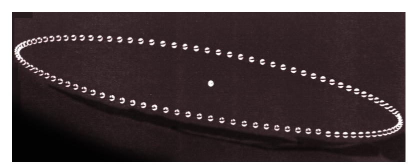
We thus find that Kepler's equal-area law is equivalent to a statement that the planet's angular momentum remains constant. But wait, why should it remain constant? — the planet is not a closed system, since it is being acted on by the sun's gravitational force. There are two valid answers. The first 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 187) 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

A 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?

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?

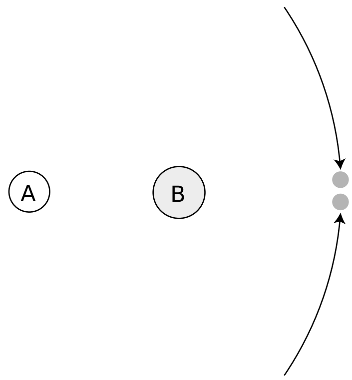


i / Discussion question B.

4.1.3 Two Theorems About Angular Momentum

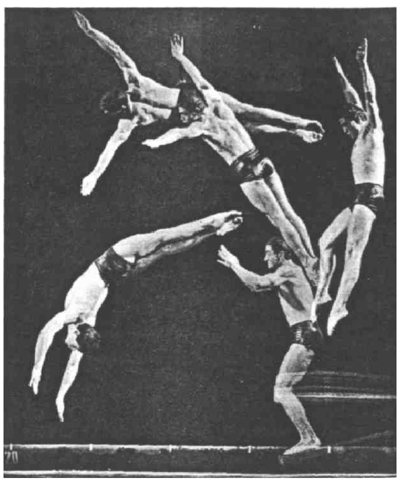
With plain old momentum, $\mathbf{p}$, we had the freedom to work in any inertial frame of reference we liked. The same object could have different values of momentum in two different frames, if the frames were not at rest with respect to each other. Conservation of momentum, however, would be true in 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 695, use the vector cross-product technique introduced in section 4.3, which greatly simplifies them.



j / Two asteroids collide.

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.



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.

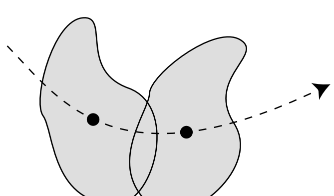
Colliding asteroids described with different axes *example 3*

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.

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:



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

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.

A system with its center of mass at rest *example 4*
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.)

Angular momentum of a rigid object *example 5*
▷ 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?

▷ 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

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

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

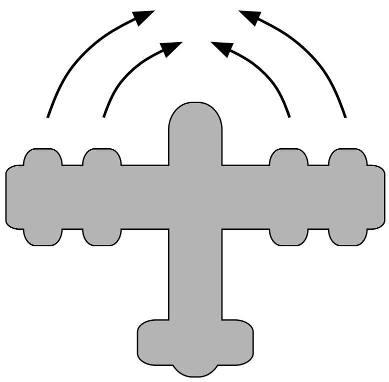
Force is the rate of transfer of momentum. The corresponding quantity in the case of angular momentum is called torque (rhymes with "fork"). Where force tells us how hard we are pushing or pulling on something, torque indicates how hard we are twisting on it. Torque is represented by the Greek letter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

$$\tau_{total} = \mathrm{d}L/\mathrm{d}t \qquad .$$

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

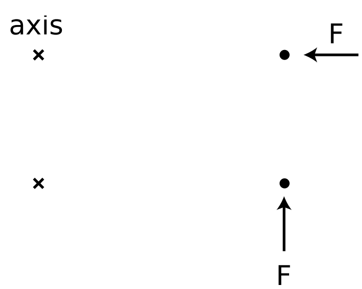
Of course a force is necessary in order to create a torque — you can't twist a screw without pushing on the wrench — but force and torque are two different things. One distinction between them is direction. We use positive and negative signs to represent forces in the two possible directions along a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



m / The plane's four engines produce zero total torque but not zero total force.

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.



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.

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 216, 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

$$a = F/m \qquad ,$$

and using $a = \mathrm{d}v/\mathrm{d}t$ we find the velocity the object acquires after a time $\mathrm{d}t$,

$$\mathrm{d}v = F\mathrm{d}t/m \qquad .$$

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

$$\begin{aligned} m\,\mathrm{d}v\,r &= F\,\mathrm{d}t\,r \\ \mathrm{d}L &= F\,\mathrm{d}t\,r \qquad . \end{aligned}$$

Dividing by $\mathrm{d}t$ gives the torque:

$$\begin{aligned} \frac{\mathrm{d}L}{\mathrm{d}t} &= Fr \\ \tau &= Fr \qquad . \end{aligned}$$

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,

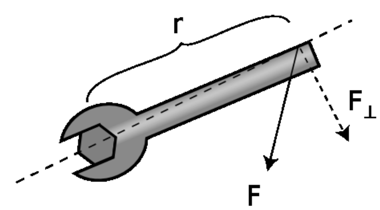
$$\tau = F_\perp r \qquad .$$

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

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

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

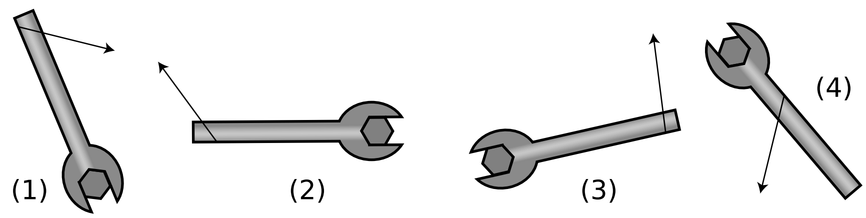
where r is the distance from the axis to the point of application of the force, and $F_\perp$ is the component of the force that is perpendicular to the line joining the axis to the point of application.



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

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 N·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.



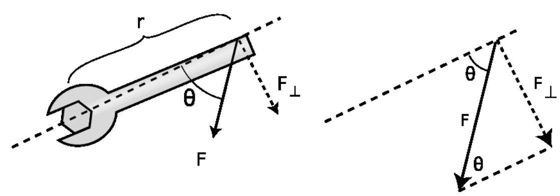
p / Self-check.

Self-Check
Compare the magnitudes and signs of the four torques shown in figure p. ▷ Answer, p. 706

How torque depends on the direction of the force *example 6*

▷ 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,

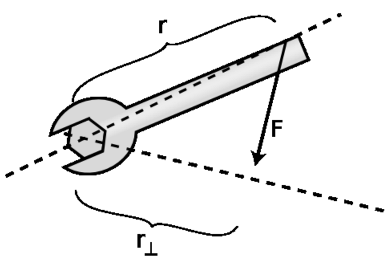


and the interior angle opposite to $F_\perp$ equals θ. The absolute value of $F_\perp$ can thus be expressed as

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

leading to

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



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

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:

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

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

[2] A proof is given in example 28 on page 216

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.

Gravitational torque on an outstretched arm *example 7*

▷ 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?

▷ The total gravitational force acting on your arm is

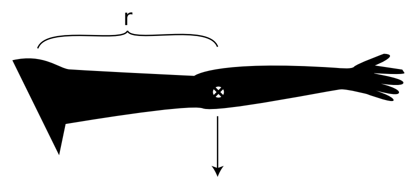
$$|\mathbf{F}| = (3.0\ \text{kg})(9.8\ \text{m/s}^2) = 29\ \text{N} \qquad .$$

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

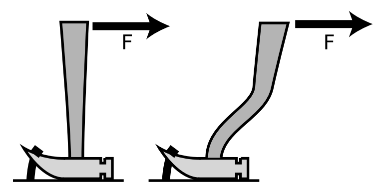
$$F_\perp = |\mathbf{F}| = 29\ \text{N} \qquad .$$

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

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



r / Example 7.



s / Discussion question D.

t / Discussion question E.

Discussion Questions

A 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.

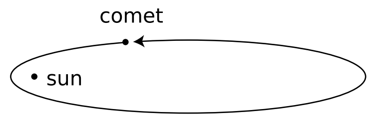
(2) Is the comet's angular momentum increasing, decreasing, or staying the same? Explain.

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.

B 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?

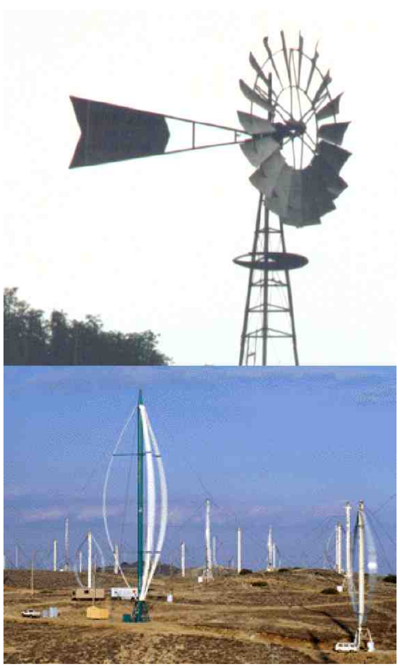
C A helicopter has, in addition to the huge fan blades on top, a smaller propeller mounted on the tail that rotates in a vertical plane. Why?

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?

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

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.



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.

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:

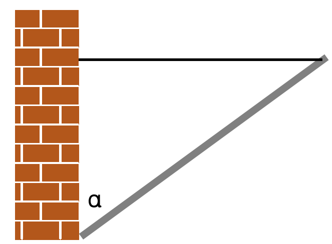
- You know the behaviors of the various types of forces, for example that a frictional force is always parallel to the surface of contact.

- 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.



w / Example 8.

A flagpole *example 8*

▷ 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?

▷ ▷ 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 $|\tau| = r|F|\sin\theta$, we have

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

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

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

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

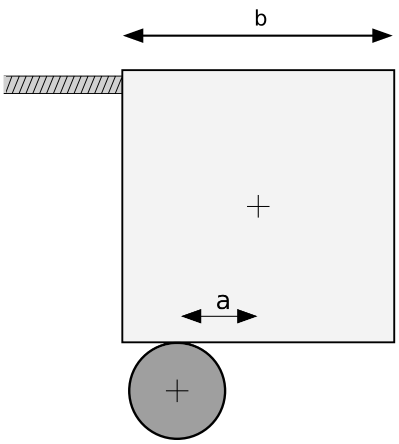
$$2|F_{cable}|\sin(90° - \alpha) - |F_{grav}|\sin\alpha = 0 \quad .$$

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 $\sin(90° - x) = \cos x$,

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

which allows us to find

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



x / Example 9.

Art! *example 9*

▷ 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.

▷ 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.

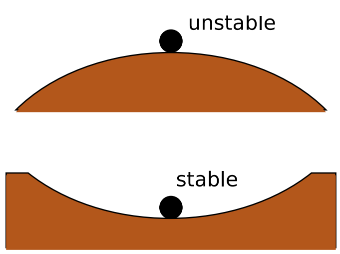
The total force on the cube in the vertical direction is zero:

$$F_N - mg = 0 \qquad .$$

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:

$$bF_T - F_N a = 0 \qquad .$$

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 $F_T = (a/b)mg$.



y / Stable and unstable equilibria.

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.

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

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.

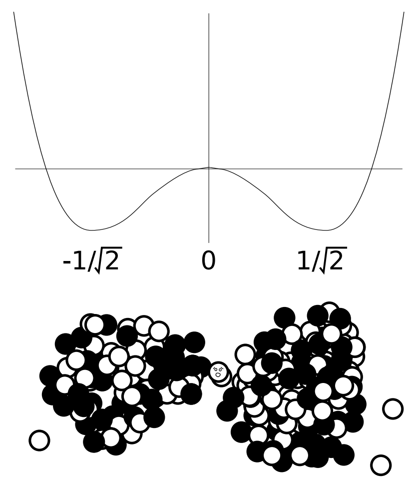
This idea can be rephrased in terms of energy. The difference between the stable and unstable equilibria shown in figure y is that in the stable equilibrium, the potential energy is at a minimum, and moving to either side of equilibrium will increase it, whereas the unstable equilibrium represents a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

An application of calculus *example 10*

▷ Nancy Neutron is living in a uranium nucleus that is undergoing fission. Nancy's potential energy as a function of position can be approximated by $PE = x^4 - x^2$, 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.

▷ The equilibrium points occur where the PE 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 $\mathrm{d}PE/\mathrm{d}x = 4x^3 - 2x$, 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 the PE 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.



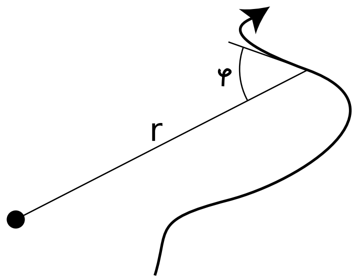
aa / Example 10.

4.1.6 Proof of Kepler's elliptical orbit law

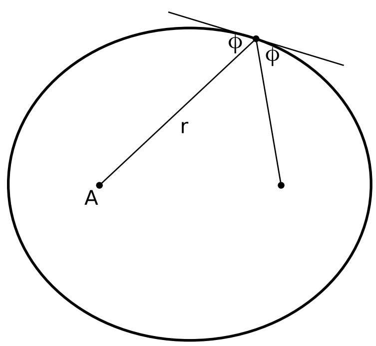
Kepler determined purely empirically that the planets' orbits were ellipses, without understanding the underlying reason in terms of physical law. Newton's proof of this fact based on his laws of motion and law of gravity was considered his crowning achievement both by him and by his contemporaries, because it showed that the same physical laws could be used to analyze both the heavens and the earth. Newton's proof was very lengthy, but by applying the more recent concepts of conservation of energy and angular momentum we can carry out the proof quite simply and succinctly. This subsection can be skipped without losing the continuity of the text.

The basic idea of the proof is that we want to describe the shape of the planet's orbit with an equation, and then show that this equation is exactly the one that represents an ellipse. Newton's original proof had to be very complicated because it was based directly on his laws of motion, which include time as a variable. To make any statement about the shape of the orbit, he had to eliminate time from his equations, leaving only space variables. But conservation laws tell us that certain things don't change over time, so they have already had time eliminated from them.

There are many ways of representing a curve by an equation, of which the most familiar is $y = ax + b$ for a line in two dimensions. It would be perfectly possible to describe a planet's orbit using an x-y equation like this, but remember that we are applying conservation of angular momentum, and the space variables that occur in the equation for angular momentum are the distance from the axis, r, and the angle between the velocity vector and the r vector, which we will call ϕ. The planet will have $\phi = 90°$ when it is moving perpendicular to the r vector, i.e. at the moments when it is at its smallest or greatest distances from the sun. When ϕ is less than $90°$ the planet is approaching the sun, and when it is greater than $90°$ it is receding from it. Describing a curve with an r-ϕ equation is like telling a driver in a parking lot a certain rule for what direction to steer based on the distance from a certain streetlight in the middle of the lot.



ab / Describing a curve by giving ϕ as a function of r.



ac / Proof that the two angles labeled ϕ are in fact equal: The definition of an ellipse is that the sum of the distances from the two foci stays constant. If we move a small distance ℓ along the ellipse, then one distance shrinks by an amount $\ell \cos \phi_1$, while the other grows by $\ell \cos \phi_2$. These are equal, so $\phi_1 = \phi_2$.

The proof is broken into the three parts for easier digestion. The first part is a simple and intuitively reasonable geometrical fact about ellipses, whose proof we relegate to the caption of figure ac; you will not be missing much if you merely absorb the result without reading the proof.

(1) If we use one of the two foci of an ellipse as an axis for defining the variables r and ϕ, then the angle between the tangent line and the line drawn to the other focus is the same as ϕ, i.e. the two angles labeled ϕ in the figure are in fact equal.

The other two parts form the meat of our proof. We state the results first and then prove them.

(2) A planet, moving under the influence of the sun's gravity with less then the energy required to escape, obeys an equation of the form

$$\sin\,\phi = \frac{1}{\sqrt{-pr^2+qr}} \qquad ,$$

where p and q are positive constants that depend on the planet's energy and angular momentum and p is greater than zero.

(3) A curve is an ellipse if and only if its r-ϕ equation is of the form

$$\sin\,\phi = \frac{1}{\sqrt{-pr^2+qr}} \qquad ,$$

where p and q are positive constants that depend on the size and shape of the ellipse.

Proof of part (2)

The component of the planet's velocity vector that is perpendicular to the **r** vector is $v_\perp = v \sin\phi$, so conservation of angular momentum tells us that $L = mrv\sin\phi$ is a constant. Since the planet's mass is a constant, this is the same as the condition

$$rv\sin\phi = \text{constant} \qquad .$$

Conservation of energy gives

$$\frac{1}{2}mv^2 - G\frac{Mm}{r} = \text{constant} \qquad .$$

We solve the first equation for v and plug into the second equation to eliminate v. Straightforward algebra then leads to the equation claimed above, with the constant p being positive because of our assumption that the planet's energy is insufficient to escape from the sun, i.e. its total energy is negative.

Proof of part (3)

We define the quantities α, d, and s as shown in figure ad. The law of cosines gives

$$d^2 = r^2 + s^2 - 2rs\ \cos\ \alpha \qquad .$$

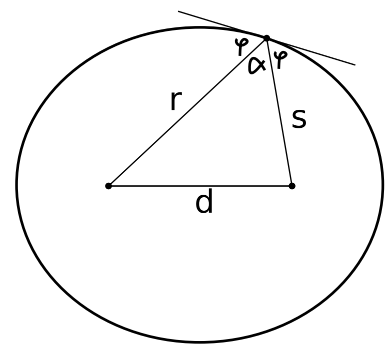
Using $\alpha = 180\,^\circ - 2\phi$ and the trigonometric identities $\cos(180\,^\circ - x) = -\cos x$ and $\cos 2x = 1 - 2\ \sin^2\ x$, we can rewrite this as

$$d^2 = r^2 + s^2 - 2rs\left(2\sin^2\phi - 1\right) \qquad .$$

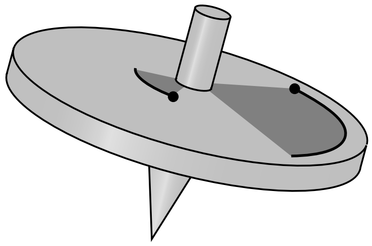
Straightforward algebra transforms this into

$$\sin\ \phi = \sqrt{\frac{(r+s)^2 - d^2}{4rs}} \qquad .$$

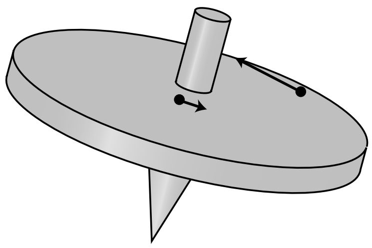
Since $r + s$ is constant, the top of the fraction is constant, and the denominator can be rewritten as $4rs = 4r(\text{constant} - r)$, which is equivalent to the desired form.



ad / Quantities referred to in the proof of part (3).



a / The two atoms cover the same angle in a given time interval.



b / Their velocity vectors, however, differ in both magnitude and direction.

4.2 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 $\mathrm{d}\theta$, and all parts of the top have the same value of $\mathrm{d}\theta$ over a certain time interval $\mathrm{d}t$. We define the angular velocity, ω (Greek omega),

$$\omega = \frac{\mathrm{d}\theta}{\mathrm{d}t} \qquad ,$$

[definition of angular velocity; θ in units of radians]

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.

Self-Check

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, p. 707

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}.

A 78-rpm record *example 11*

▷ 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

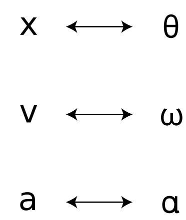
$$\frac{\text{78 revolutions}}{\text{1 minute}} \times \frac{2\pi\text{ radians}}{\text{1 revolution}} \times \frac{\text{1 minute}}{\text{60 seconds}} = 8.2\ \mathrm{s}^{-1} \qquad .$$

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,

$$\alpha = \frac{\mathrm{d}\omega}{\mathrm{d}t} \qquad , \qquad \text{[definition of angular acceleration]}$$

The symbol is the Greek letter alpha. The units of this quantity are rad/s^2, or simply 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.



c / Analogies between rotational and linear quantities.

The synodic period — *example 12*

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.

A neutron star — *example 13*

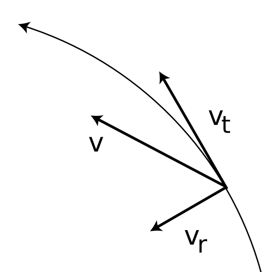
▷ A neutron star is initially observed to be rotating with an angular velocity of 2.0 s^{-1}, determined via the radio pulses it emits. If its angular acceleration is a constant -1.0×10^{-8} s^{-2}, how many rotations will it complete before it stops? (In reality, the angular acceleration is not always 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

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

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.



d / 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 dθ by the definition of radian measure, $\mathrm{d}\theta = \mathrm{d}s/r$, where positive and negative values of ds represent the two possible directions of motion along the tangential axis. We then have $v_t = \mathrm{d}s/\mathrm{d}t = r\mathrm{d}\theta/\mathrm{d}t = \omega r$,

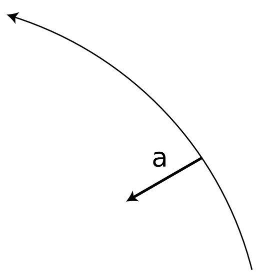
or

$$v_t = \omega r \qquad . \qquad \text{[tangential velocity of a point at a distance } r \text{ from the axis of rotation]}$$

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

$$v_r = 0 \qquad . \qquad \text{[radial velocity of a point at a distance } r \text{ from the axis of rotation]}$$

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 $\mathrm{d}\theta_{degrees}$ and $\mathrm{d}s$ is $\mathrm{d}\theta_{degrees} = (360/2\pi)s/r$, 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.



e / 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 154, the magnitude of this acceleration is

$$a_r = \omega^2 r \qquad . \qquad \text{[radial acceleration of a point at a distance } r \text{ from the axis]}$$

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

$$a_t = \alpha r \qquad . \qquad \text{[radial acceleration of a point at a distance } r \text{ from the axis]}$$

Self-Check

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, p. 707

Radial acceleration at the surface of the Earth *example 14*

▷ 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

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

which gives an acceleration of

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

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 $L = rp = (r)(mv) = (r)(m\omega r) = mr^2\omega$. 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

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

[definition of the moment of inertia; for rigid-body rotation in a plane; r is the distance from the axis, measured perpendicular to the axis]

The angular momentum of a rigidly rotating body is then

$$L = I\omega \qquad .$$

[angular momentum of rigid-body rotation in a plane]

$x \longleftrightarrow \theta$

$v \longleftrightarrow \omega$

$a \longleftrightarrow \alpha$

$m \longleftrightarrow I$

$p \longleftrightarrow L$

$F \longleftrightarrow \tau$

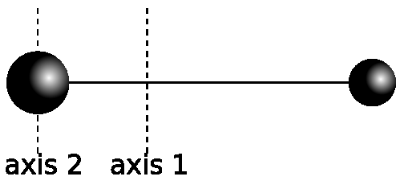
f / Analogies between rotational and linear quantities.

Since torque is defined as $\mathrm{d}L/\mathrm{d}t$, and a rigid body has a constant moment of inertia, we have $\tau = \mathrm{d}L/\mathrm{d}t = I\mathrm{d}\omega/\mathrm{d}t = I\alpha$,

$$\tau = I\alpha \quad , \qquad \text{[relationship between torque and angular acceleration for rigid-body rotation in a plane]}$$

which is analogous to $F = ma$.

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



g / Example 15

A barbell *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

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

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:

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

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.

The parallel axis theorem *example 16*

▷ 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 m)+h, and the small one's is (2.0 m)-h. The new moment of inertia is

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

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, 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.

Scaling of the moment of inertia *example 17*

▷ (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. (We'll also use them in subsection 11.2.3.) 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

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

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

$$\sum_{r=1}^{8} 8 \qquad .$$

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 $\mathrm{d}x$ and height $\mathrm{d}y$, and therefore the infinitesimal area $\mathrm{d}A = \mathrm{d}x\mathrm{d}y$. For convenience, we'll imagine that the rectangle's lower left corner is at the origin. Then the area is given by this integral:

$$\begin{aligned}
\text{area} &= \int_{y=0}^{b}\int_{x=0}^{a} \mathrm{d}A \\
&= \int_{y=0}^{b}\int_{x=0}^{a} \mathrm{d}x\,\mathrm{d}y
\end{aligned}$$

Notice how the leftmost integral sign, over y, and the rightmost differential, $\mathrm{d}y$, 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 $\mathrm{d}x$ 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

$$\begin{aligned}
\text{area} &= \int_{y=0}^{b}\int_{x=0}^{a} \mathrm{d}A \\
&= \int_{y=0}^{b}\int_{x=0}^{a} \mathrm{d}x\,\mathrm{d}y \\
&= \int_{y=0}^{b}\left(\int_{x=0}^{a} \mathrm{d}x\right)\mathrm{d}y \\
&= \int_{y=0}^{b} a\,\mathrm{d}y \\
&= a\int_{y=0}^{b} \mathrm{d}y \\
&= ab \qquad .
\end{aligned}$$

Area of a triangle — *example 18*

▷ 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

$$
\begin{aligned}
\text{area} &= \int_{y=0}^{a} \int_{x=0}^{y} \mathrm{d}A \\
&= \int_{y=0}^{a} \int_{x=0}^{y} \mathrm{d}x\, \mathrm{d}y \\
&= \int_{y=0}^{a} \left(\int_{x=0}^{y} \mathrm{d}x \right) \mathrm{d}y \\
&= \int_{y=0}^{a} y\, \mathrm{d}y \\
&= \frac{1}{2} a^2
\end{aligned}
$$

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.

Volume of a cube *example 19*

▷ 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 $\mathrm{d}V = \mathrm{d}x\mathrm{d}y\mathrm{d}z$.

$$
\begin{aligned}
\text{volume} &= \int_{z=0}^{a} \int_{y=0}^{a} \int_{x=0}^{a} \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z \\
&= \int_{z=0}^{a} \int_{y=0}^{a} a\, \mathrm{d}y\, \mathrm{d}z \\
&= a \int_{z=0}^{a} \int_{y=0}^{a} \mathrm{d}y\, \mathrm{d}z \\
&= a \int_{z=0}^{a} a\, \mathrm{d}z \\
&= a^3
\end{aligned}
$$

Area of a circle *example 20*

▷ 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

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

Substituting $u = x/r$,

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

The definite integal 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,

$$\begin{aligned} I &= \int r^2 \mathrm{d}m \\ &= \int_{-L/2}^{L/2} x^2\, \frac{M}{L}\, \mathrm{d}x \qquad [r = |x|, \text{ so } r^2 = x^2] \\ &= \frac{1}{12} M L^2 \end{aligned}$$

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 $r + \mathrm{d}r$), the slice's contribution to the total moment of inertia is simply $r^2\mathrm{d}m$. We then have

$$\begin{aligned} I &= \int r^2 \mathrm{d}m \\ &= \int r^2 \rho \, \mathrm{d}V \qquad , \end{aligned}$$

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, $\mathrm{d}V = \pi(r + \mathrm{d}r)^2 t - \pi r^2 t = 2\pi t r \mathrm{d}r$.

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

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.

$$\begin{aligned} I &= \int r^2 \mathrm{d}m \\ &= \rho \int r^2 \mathrm{d}V \\ &= \rho \int_{b/2}^{b/2} \int_{b/2}^{b/2} \int_{b/2}^{b/2} \left(y^2 + z^2\right) \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \\ &= \rho b \int_{b/2}^{b/2} \int_{b/2}^{b/2} \left(y^2 + z^2\right) \mathrm{d}y \, \mathrm{d}z \end{aligned}$$

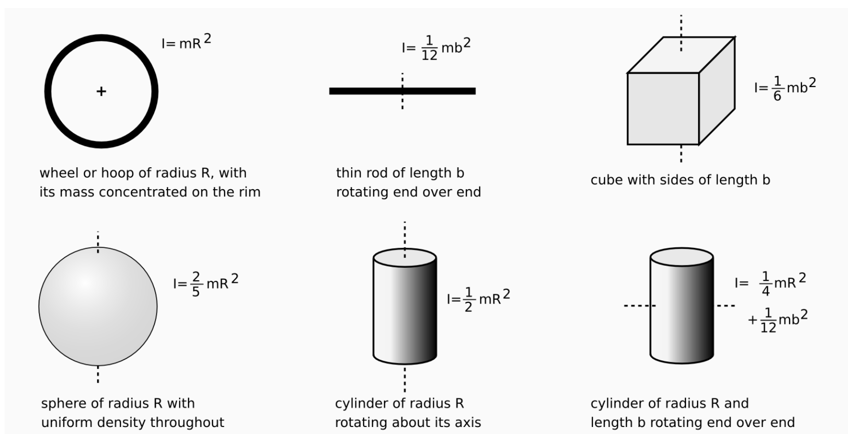
The fact that the last step is a trivial integral results from the symmetry of the problem. The integrand of the remaining double integral breaks down into two terms, each of which depends on only one of the variables, so we break it into two integrals,

$$I = \rho b \int_{b/2}^{b/2} \int_{b/2}^{b/2} y^2 \mathrm{d}y\, \mathrm{d}z + \rho b \int_{b/2}^{b/2} \int_{b/2}^{b/2} z^2 \mathrm{d}y\, \mathrm{d}z$$

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

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

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



h / Momenta of inertia of some geometric shapes.

The hammer throw *example 21*

▷ 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.

Toppling a rod *example 22*

▷ 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?

▷ 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

$$p = Am^j b^k g^l \qquad ,$$

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

$$j = 1 \qquad ,$$

meters to the first power requires

$$k + l = 1 \qquad ,$$

and seconds to the power -1 implies

$$l = 1/2 \qquad .$$

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

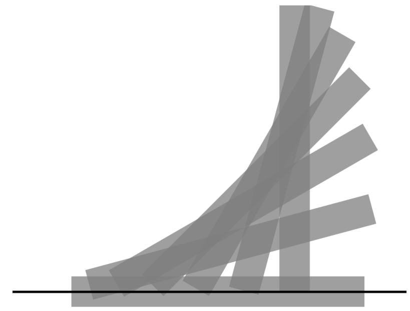
$$p = Am\sqrt{bg} \qquad .$$

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 225.



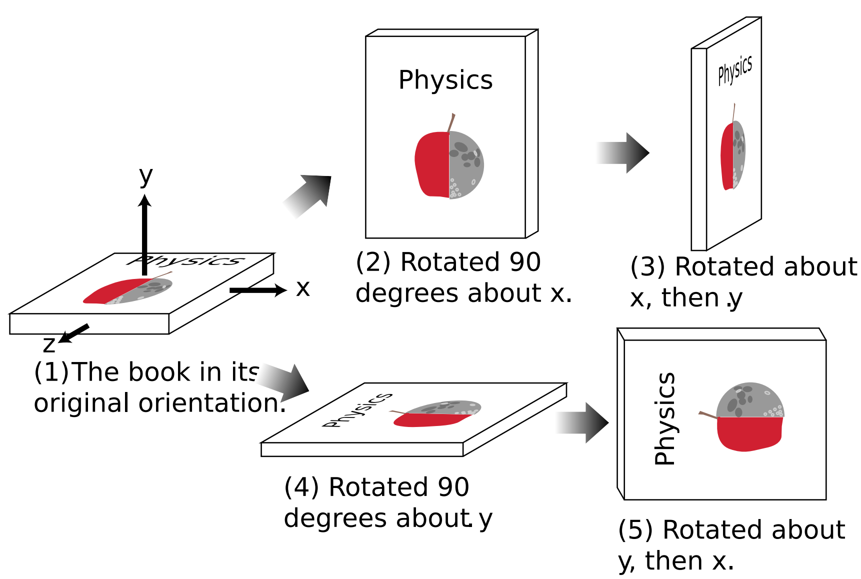
i / Example 22.

4.3 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.

4.3.1 Rigid-body kinematics in three dimensions

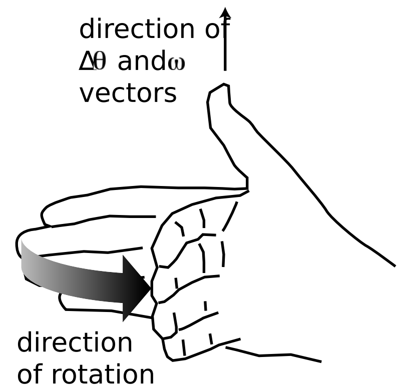
How do we generalize rigid-body kinematics to three dimensions? When we wanted to generalize the kinematics of a moving particle to three dimensions, we made the numbers r, v, and a into vectors $\mathbf{r}$, $\mathbf{v}$, and $\mathbf{a}$. This worked because these quantities all obeyed the same laws of vector addition. For instance, one of the laws of vector addition is that, just like addition of numbers, vector addition gives the same result regardless of the order of the two quantities being added. Thus you can step sideways 1 m to the right and then step forward 1 m, and the end result is the same as if you stepped forward first and then to the side. In order words, it didn't matter whether you took $\Delta\mathbf{r}_1 + \Delta\mathbf{r}_2$ or $\Delta\mathbf{r}_2 + \Delta\mathbf{r}_1$. In math this is called the commutative property of addition.



a / Performing the rotations in one order gives one result, 3, while reversing the order gives a different result, 5.

Angular motion, unfortunately doesn't have this property, as shown in figure a. Doing a rotation about the x axis and then about y gives one result, while doing them in the opposite order gives a different result. These operations don't "commute," i.e. it makes a difference what order you do them in.

This means that there is in general no possible way to construct a $\Delta\boldsymbol{\theta}$ vector. However, if you try doing the operations shown in figure a using small rotation, say about 10 degrees instead of 90, you'll find that the result is nearly the same regardless of what order you use; small rotations are very nearly commutative. Not only that, but the result of the two 10-degree rotations is about the same as a single, somewhat larger, rotation about an axis that lies symmetrically at between the x and y axes at 45 degree angles to each one. This is exactly what we would expect if the two small rotations did act like vectors whose directions were along the axis of rotation. We therefore define a $\mathrm{d}\boldsymbol{\theta}$ vector whose magnitude is the amount of rotation in units of radians, and whose direction is along the axis of rotation. Actually this definition is ambiguous, because there it could point in either direction along the axis. We therefore use a right-hand rule as shown in figure b to define the direction of the $\mathrm{d}\boldsymbol{\theta}$ vector, and the $\boldsymbol{\omega}$ vector, $\boldsymbol{\omega}=\mathrm{d}\boldsymbol{\theta}/\mathrm{d}t$, based on it. Aliens on planet Tammyfaye may decide to define it using their left hands rather than their right, but as long as they keep their scientific literature separate from ours, there is no problem. When entering a physics exam, always be sure to write a large warning note on your left hand in magic marker so that you won't be tempted to use it for the right-hand rule while keeping your pen in your right.



b / The right-hand rule for associating a vector with a direction of rotation.

Self-Check

Use the right-hand rule to determine the directions of the ω vectors in each rotation shown in figures a/1 through a/5. ▷ Answer, p. 707

Because the vector relationships among $\mathrm{d}\boldsymbol{\theta}$, $\boldsymbol{\omega}$, and $\boldsymbol{\alpha}$ are strictly analogous to the ones involving $\mathrm{d}\mathbf{r}$, $\mathbf{v}$, and $\mathbf{a}$ (with the proviso that we avoid describing large rotations using $\Delta\boldsymbol{\theta}$ vectors), any operation in $\mathbf{r}$-$\mathbf{v}$-$\mathbf{a}$ vector kinematics has an exact analog in $\boldsymbol{\theta}$-$\boldsymbol{\omega}$-$\boldsymbol{\alpha}$ kinematics.

Result of successive 10-degree rotations *example 23*

▷ What is the result of two successive (positive) 10-degree rotations about the x and y axes? That is, what single rotation about a single axis would be equivalent to executing these in succession?

▷ The result is only going to be approximate, since 10 degrees is not an infinitesimally small angle, and we are not told in what order the rotations occur. To some approximation, however, we can add the $\Delta\theta$ vectors in exactly the same way we would add $\Delta\mathbf{r}$ vectors, so we have

$$\begin{aligned}\Delta\theta &\approx \Delta\theta_1+\Delta\theta_2\\ &\approx (10\text{ degrees})\hat{\mathbf{x}}+(10\text{ degrees})\hat{\mathbf{y}}\quad.\end{aligned}$$

This is a vector with a magnitude of $\sqrt{(10\text{ deg})^2+(10\text{ deg})^2}=14$ deg, and it points along an axis midway between the x and y axes.

4.3.2 Angular momentum in three dimensions

The vector cross product

In order to expand our system of three-dimensional kinematics to include dynamics, we will have to generalize equations like $v_t = \omega r$, $\tau = rF\sin\theta_{rF}$, and $L = rp\sin\theta_{rp}$, each of which involves three quantities that we have either already defined as vectors or that we want to redefine as vectors. Although the first one appears to differ from the others in its form, it could just as well be rewritten as $v_t = \omega r\sin\theta_{\omega r}$, since $\theta_{\omega r} = 90\,°$, and $\sin\theta_{\omega r} = 1$.

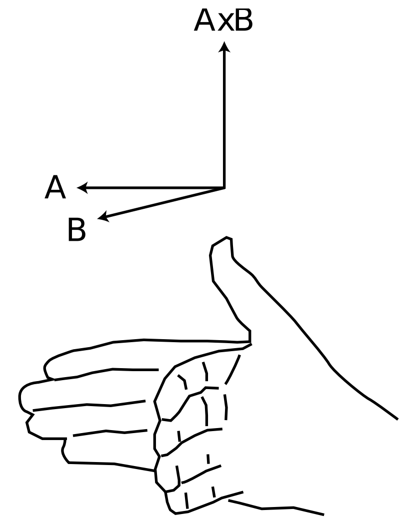
It thus appears that we have discovered something general about the physically useful way to relate three vectors in a multiplicative way: the magnitude of the result always seems to be proportional to the product of the magnitudes of the two vectors being "multiplied," and also to the sine of the angle between them.

Is this pattern just an accident? Actually the sine factor has a very important physical property: it goes to zero when the two vectors are parallel. This is a Good Thing. The generalization of angular momentum into a three-dimensional vector, for example, is presumably going to describe not just the clockwise or counterclockwise nature of the motion but also from which direction we would have to view the motion so that it was clockwise or counterclockwise. (A clock's hands go counterclockwise as seen from behind the clock, and don't rotate at all as seen from above or to the side.) Now suppose a particle is moving directly away from the origin, so that its $\mathbf{r}$ and $\mathbf{p}$ vectors are parallel. It is not going around the origin from any point of view, so its angular momentum vector had better be zero.

Thinking in a slightly more abstract way, we would expect the angular momentum vector to point perpendicular to the plane of motion, just as the angular velocity vector points perpendicular to the plane of motion. The plane of motion is the plane containing both $\mathbf{r}$ and $\mathbf{p}$, if we place the two vectors tail-to-tail. But if $\mathbf{r}$ and $\mathbf{p}$ are parallel and are placed tail-to-tail, then there are infinitely many planes containing them both. To pick one of these planes in preference to the others would violate the symmetry of space, since they should all be equally good. Thus the zero-if-parallel property is a necessary consequence of the underlying symmetry of the laws of physics.

The following definition of a kind of vector multiplication is consistent with everything we've seen so far, and later we'll prove that the definition is unique, i.e. if we believe in the symmetry of space, it is essentially the only way of defining the multiplication of two vectors to produce a third vector:

Definition of the vector cross product:
The cross product $\mathbf{A}\times\mathbf{B}$ of two vectors $\mathbf{A}$ and $\mathbf{B}$ is defined as follows:
(1) Its magnitude is defined by $|\mathbf{A}\times\mathbf{B}| = |\mathbf{A}||\mathbf{B}|\sin\theta_{AB}$, where θ_{AB} is the angle between $\mathbf{A}$ and $\mathbf{B}$ when they are placed tail-to-tail.
(2) Its direction is along the line perpendicular to both $\mathbf{A}$ and $\mathbf{B}$. Of the two such directions, it is the one that obeys the right-hand rule shown in figure c.



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

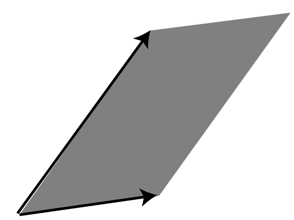
The name "cross product" refers to the symbol, and distinguishes it from the dot product, which acts on two vectors but produces a scalar.

Although the vector cross-product has nearly all the properties of numerical multiplication, e.g. $\mathbf{A}\times(\mathbf{B}+\mathbf{C}) = \mathbf{A}\times\mathbf{B}+\mathbf{A}\times\mathbf{C}$, it lacks the usual property of commutativity. Try applying the right-hand rule to find the direction of the vector cross product $\mathbf{B}\times\mathbf{A}$ using the two vectors shown in the figure. This requires starting with a flattened hand with the four fingers pointing along $\mathbf{B}$, and then curling the hand so that the fingers point along $\mathbf{A}$. The only possible way to do this is to point your thumb toward the floor, in the opposite direction. Thus for the vector cross product we have

$$\mathbf{A}\times\mathbf{B} = -\mathbf{B}\times\mathbf{A} \qquad ,$$

a property known as anticommutativity. The vector cross product is also not associative, i.e. $\mathbf{A}\times(\mathbf{B}\times\mathbf{C})$ is usually not the same as $(\mathbf{A}\times\mathbf{B})\times\mathbf{C}$.

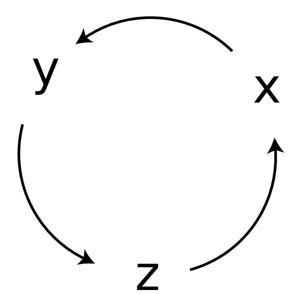
A geometric interpretation of the cross product, d, is that if both $\mathbf{A}$ and $\mathbf{B}$ are vectors with units of distance, then the magnitude of their cross product can be interpreted as the area of the parallelogram they form when placed tail-to-tail.


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

A useful expression for the components of the vector cross product in terms of the components of the two vectors being multiplied is as follows:

$$\begin{aligned}(\mathbf{A}\times\mathbf{B})_x &= A_yB_z - B_yA_z\\(\mathbf{A}\times\mathbf{B})_y &= A_zB_x - B_zA_x\\(\mathbf{A}\times\mathbf{B})_z &= A_xB_y - B_xA_y\end{aligned}$$

I'll prove later that these expressions are equivalent to the previous definition of the cross product. Although they may appear formidable, they have a simple structure: the subscripts on the right are the other two besides the one on the left, and each equation is related to the preceding one by a cyclic change in the subscripts, e. If the subscripts were not treated in some completely symmetric manner like this, then the definition would provide some way to distinguish one axis from another, which would violate the symmetry of space.



e / A cyclic change in the *x*, *y*, and *z* subscripts.

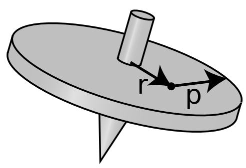
Self-Check
Show that the component equations are consistent with the rule $\mathbf{A}\times\mathbf{B} = -\mathbf{B}\times\mathbf{A}$. ▷ Answer, p. 707

Angular momentum in three dimensions

In terms of the vector cross product, we have:

$$\mathbf{v} = \boldsymbol{\omega}\times\mathbf{r}$$
$$\mathbf{L} = \mathbf{r}\times\mathbf{p}$$
$$\boldsymbol{\tau} = \mathbf{r}\times\mathbf{F}$$

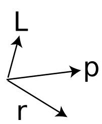
But wait, how do we know these equations are even correct? For instance, how do we know that the quantity defined by $\mathbf{r}\times\mathbf{p}$ is in fact conserved? Well, just as we saw on page 158 that the dot product is unique (i.e., can only be defined in one way while observing rotational invariance), the cross product is also unique, as proved on page 694. If $\mathbf{r}\times\mathbf{p}$ was not conserved, then there could not be any generally conserved quantity that would reduce to our old definition of angular momentum in the special case of plane rotation. This doesn't prove conservation of angular momentum — only experiments can prove that — but it does prove that if angular momentum is conserved in three dimensions, there is only one possible way to generalize from two dimensions to three.



f / The position and momentum vectors of an atom in the spinning top.

Angular momentum of a spinning top — *example 24*

As an illustration, we consider the angular momentum of a spinning top. Figures f and g show the use of the vector cross product to determine the contribution of a representative atom to the total angular momentum. Since every other atom's angular momentum vector will be in the same direction, this will also be the direction of the total angular momentum of the top. This happens to be rigid-body rotation, and perhaps not surprisingly, the angular momentum vector is along the same direction as the angular velocity vector.



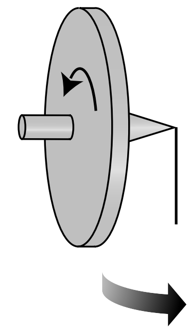
g / The right-hand rule for the atom's contribution to the angular momentum.

Three important points are illustrated by this example: (1) When we do the full three-dimensional treatment of angular momentum, the "axis" from which we measure the position vectors is just an arbitrarily chosen point. If this had not been rigid-body rotation, we would not even have been able to identify a single line about which every atom circled. (2) Starting from figure f, we had to rearrange the vectors to get them tail-to-tail before applying the right-hand rule. If we had attempted to apply the right-hand rule to figure f, the direction of the result would have been exactly the opposite of the correct answer. (3) The equation $\mathbf{L} = \mathbf{r}\times\mathbf{p}$ cannot be applied all at once to an entire system of particles. The total momentum of the top is zero, which would give an erroneous result of zero angular momentum (never mind the fact that $\mathbf{r}$ is not well defined for the top as a whole).

Doing the right-hand rule like this requires some practice. I urge you to make models like g out of rolled up pieces of paper and to practice with the model in various orientations until it becomes natural.

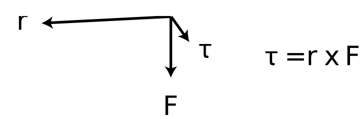
Precession *example 25*

Figure h shows a counterintuitive example of the concepts we've been discussing. One expects the torque due to gravity to cause the top to flop down. Instead, the top remains spinning in the horizontal plane, but its axis of rotation starts moving in the direction shown by the shaded arrow. This phenomenon is called precession. Figure i shows that the torque due to gravity is out of the page. (Actually we should add up all the torques on all the atoms in the top, but the qualitative result is the same.) Since torque is the rate of change of angular momentum, $\tau = d\mathbf{L}/dt$, the $\Delta\mathbf{L}$ vector must be in the same direction as the torque (division by a positive scalar doesn't change the direction of the vector). As shown in j, this causes the angular momentum vector to twist in space without changing its magnitude.

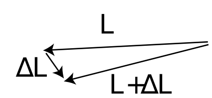


h / A top is supported at its tip by a pinhead. (More practical devices to demonstrate this would use a double bearing.)

For similar reasons, the Earth's axis precesses once every 26,000 years (although not through a great circle, since the angle between the axis and the force isn't 90 degrees as in figure h). This precession is due to a torque exerted by the moon. If the Earth was a perfect sphere, there could be no precession effect due to symmetry. However, the Earth's own rotation causes it to be slightly flattened (oblate) relative to a perfect sphere, giving it "love handles" on which the moon's gravity can act. The moon's gravity on the nearer side of the equatorial bulge is stronger, so the torques do not cancel out perfectly. Presently the earth's axis very nearly lines up with the star Polaris, but in 12,000 years, the pole star will be Vega instead.



i / The torque made by gravity is in the horizontal plane.



j / The $\Delta\mathbf{L}$ vector is in the same direction as the torque, out of the page.

The frisbee *example 26*

The flow of the air over a flying frisbee generates lift, and the lift at the front and back of the frisbee isn't necessarily balanced. If you throw a frisbee without rotating it, as if you were shooting a basketball with two hands, you'll find that it pitches, i.e., its nose goes either up or down. When I do this with my frisbee, it goes nose down, which apparently means that the lift at the back of the disc is greater than the lift at the front. The two torques are unbalanced, resulting in a total torque that points to the left.

The way you actually throw a frisbee is with one hand, putting a lot of spin on it. If you throw backhand, which is how most people first learn to do it, the angular momentum vector points down (assuming you're right-handed). On my frisbee, the aerodynamic torque to the left would therefore tend to make the angular momentum vector precess in the clockwise direction as seen by the thrower. This would cause the disc to roll to the right, and therefore follow a curved trajectory. Some specialized discs, used in the sport of disc golf, are actually designed intentionally to show this behavior; they're known as "understable" discs. However, the typical frisbee that most people play with is designed to be stable: as the disc rolls to one side, the airflow around it is altered in way that tends to bring the disc back into level flight. Such a disc will therefore tend to fly in a straight line, provided that it is thrown with enough angular momentum.

r	4	5	0
F	1	2	3

k / Example 27.

Finding a cross product by components *example 27*

▷ What is the torque produced by a force given by $\hat{\mathbf{x}}+2\hat{\mathbf{y}}+3\hat{\mathbf{z}}$ (in units of Newtons) acting on a point whose radius vector is $4\hat{\mathbf{x}}+2\hat{\mathbf{y}}$ (in meters)?

▷ It's helpful to make a table of the components as shown in the figure. The results are

$$\begin{aligned}\tau_x &= r_y F_z - F_y r_z = 15\ \text{N}\cdot\text{m}\\ \tau_y &= r_z F_x - F_z r_x = -12\ \text{N}\cdot\text{m}\\ \tau_z &= r_x F_y - F_x r_y = 3\ \text{N}\cdot\text{m}\end{aligned}$$

Torque and angular momentum *example 28*

In this example, we prove explicitly the consistency of the equations involving torque and angular momentum that we proved above based purely on symmetry. The proof uses calculus. Starting from the definition of torque, we have

$$\begin{aligned}\tau &= \frac{\mathrm{d}\mathbf{L}}{\mathrm{d}t}\\ &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_i \mathbf{r}_i \times \mathbf{p}_i\\ &= \sum_i \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{r}_i \times \mathbf{p}_i) \qquad .\end{aligned}$$

The derivative of a cross product can be evaluated in the same was as the derivative of an ordinary scalar product:

$$\tau = \sum_i \left[\left(\frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} \times \mathbf{p}_i\right) + \left(\mathbf{r}_i \times \frac{\mathrm{d}\mathbf{p}_i}{\mathrm{d}t}\right)\right]$$

The first term is zero for each particle, since the velocity vector is parallel to the momentum vector. The derivative appearing in the second term is the force acting on the particle, so

$$\tau = \sum_i \mathbf{r}_i \times \mathbf{F}_i \qquad ,$$

which is the relationship we set out to prove.

4.3.3 Rigid-body dynamics in three dimensions

The student who is not madly in love with mathematics may wish to skip the rest of this section after absorbing the statement that, for a typical, asymmetric object, the angular momentum vector and the angular velocity vector need not be parallel. That is, only for a body that possesses symmetry about the rotation axis is it true that $\mathbf{L} = I\boldsymbol{\omega}$ (the rotational equivalent of $\mathbf{p} = m\mathbf{v}$) for some scalar I.

Let's evaluate the angular momentum of a rigidly rotating system of particles:

$$\begin{aligned}\mathbf{L} &= \sum_i \mathbf{r}_i \times \mathbf{p}_i \\ &= \sum_i m_i \mathbf{r}_i \times \mathbf{v}_i \\ &= \sum_i m_i \mathbf{r}_i \times (\boldsymbol{\omega} \times \mathbf{r}_i)\end{aligned}$$

An important mathematical skill is to know when to give up and back off. This is a complicated expression, and there is no reason to expect it to simplify and, for example, take the form of a scalar multiplied by ω. Instead we examine its general characteristics. If we expanded it using the equation that gives the components of a vector cross product, every term would have one of the ω components raised to the first power, multiplied by a bunch of other stuff. The most general possible form for the result is

$$\begin{aligned}L_x &= I_{xx}\omega_x + I_{xy}\omega_y + I_{xz}\omega_z \\ L_y &= I_{yx}\omega_x + I_{yy}\omega_y + I_{yz}\omega_z \\ L_z &= I_{zx}\omega_x + I_{zy}\omega_y + I_{zz}\omega_z \qquad ,\end{aligned}$$

which you may recognize as a case of matrix multiplication. The moment of inertia is not a scalar, and not a three-component vector. It is a matrix specified by nine numbers, called its matrix elements.

The elements of the moment of inertia matrix will depend on our choice of a coordinate system. In general, there will be some special coordinate system, in which the matrix has a simple diagonal form:

$$\begin{aligned}L_x &= I_{xx}\omega_x \\ L_y &= \qquad I_{yy}\omega_y \\ L_z &= \qquad\qquad I_{zz}\omega_z \qquad .\end{aligned}$$

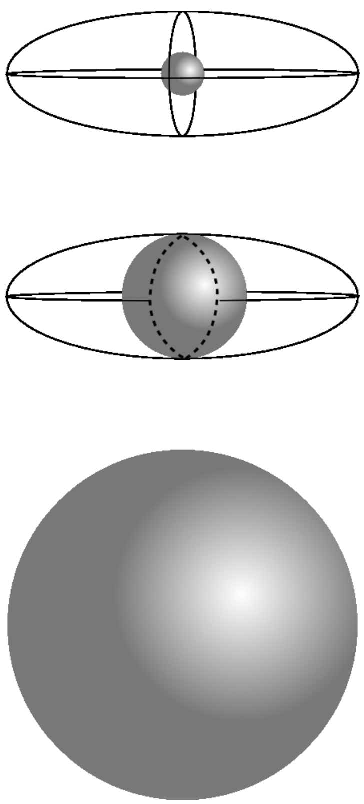
The three special axes that cause this simplification are called the principal axes of the object, and the corresponding coordinate system is the principal axis system. For symmetric shapes such as a rectangular box or an ellipsoid, the principal axes lie along the intersections of the three symmetry planes, but even an asymmetric body has principal axes.

We can also generalize the plane-rotation equation $K = (1/2)I\omega^2$ to three dimensions as follows:

$$\begin{aligned} K &= \sum_i \frac{1}{2} m_i v_i^2 \\ &= \frac{1}{2} \sum_i m_i (\boldsymbol{\omega} \times \mathbf{r}_i) \cdot (\boldsymbol{\omega} \times \mathbf{r}_i) \end{aligned}$$

We want an equation involving the moment of inertia, and this has some evident similarities to the sum we originally wrote down for the moment of inertia. To massage it into the right shape, we need the vector identity $(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} = (\mathbf{B} \times \mathbf{C}) \cdot \mathbf{A}$, which we state without proof. We then write

$$\begin{aligned} K &= \frac{1}{2} \sum_i m_i \left[\mathbf{r}_i \times (\boldsymbol{\omega} \times \mathbf{r}_i) \right] \cdot \boldsymbol{\omega} \\ &= \frac{1}{2} \boldsymbol{\omega} \cdot \sum_i m_i \mathbf{r}_i \times (\boldsymbol{\omega} \times \mathbf{r}_i) \\ &= \frac{1}{2} \mathbf{L} \cdot \boldsymbol{\omega} \end{aligned}$$



l / Visualizing surfaces of constant energy and angular momentum in L_x-L_y-L_z space.

As a reward for all this hard work, let's analyze the problem of the spinning shoe that I posed at the beginning of the chapter. The three rotation axes referred to there are approximately the principal axes of the shoe. While the shoe is in the air, no external torques are acting on it, so its angular momentum vector must be constant in magnitude and direction. Its kinetic energy is also constant. That's in the room's frame of reference, however. The principal axis frame is attached to the shoe, and tumbles madly along with it. In the principal axis frame, the kinetic energy and the magnitude of the angular momentum stay constant, but the actual direction of the angular momentum need not stay fixed (as you saw in the case of rotation that was initially about the intermediate-length axis). Constant $|\mathbf{L}|$ gives

$$L_x^2 + L_y^2 + L_z^2 = \text{constant} \qquad .$$

In the principal axis frame, it's easy to solve for the components of $\boldsymbol{\omega}$ in terms of the components of $\mathbf{L}$, so we eliminate ω from the expression $2K = \mathbf{L} \cdot \boldsymbol{\omega}$, giving

$$\frac{1}{I_{xx}} L_x^2 + \frac{1}{I_{yy}} L_y^2 + \frac{1}{I_{zz}} L_z^2 = \text{constant \#2.}$$

The first equation is the equation of a sphere in the three dimensional space occupied by the angular momentum vector, while the second one is the equation of an ellipsoid. The top figure corresponds to the case of rotation about the shortest axis, which has the greatest moment of inertia element. The intersection of the two surfaces consists only of the two points at the front and back of the sphere. The angular momentum is confined to one of these points, and can't change its direction, i.e. its orientation with respect to the principal axis system, which is another way of saying that the shoe can't change its orientation with respect to the angular momentum vector. In the bottom figure, the shoe is rotating about the longest axis. Now the angular momentum vector is trapped at one of the two points on the right or left. In the case of rotation about the axis with the intermediate moment of inertia element, however, the intersection of the sphere and the ellipsoid is not just a pair of isolated points but the curve shown with the dashed line. The relative orientation of the shoe and the angular momentum vector can and will change.

As an example of this, you can try tossing a box-shaped object up in the air, spinning about one of its principal axes. A book works, if you tape it shut, but I've also used a shoe. If you spin it about its intermediate-moment-of-inertia axis, it tumbles irregularly.

One more exotic example has to do with nuclear physics. Although you have probably visualized atomic nuclei as nothing more than featureless points, or perhaps tiny spheres, they are often ellipsoids with one long axis and two shorter, equal ones. Although a spinning nucleus normally gets rid of its angular momentum via gamma ray emission within a period of time on the order of picoseconds, it may happen that a deformed nucleus gets into a state in which has a large angular momentum is along its long axis, which is a very stable mode of rotation. Such states can live for seconds or even years! (There is more to the story — this is the topic on which I wrote my Ph.D. thesis — but the basic insight applies even though the full treatment requires fancy quantum mechanics.)

m / The Explorer I satellite.

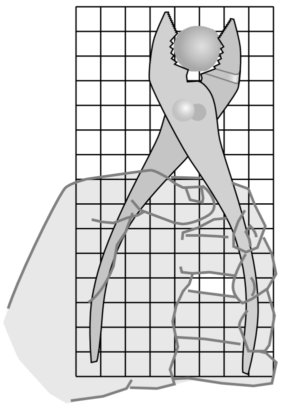
Our analysis has so far assumed that the kinetic energy of rotation energy can't be converted into other forms of energy such as heat, sound, or vibration. When this assumption fails, then rotation about the axis of least moment of inertia becomes unstable, and will eventually convert itself into rotation about the axis whose moment of inertia is greatest. This happened to the U.S.'s first artificial satellite, Explorer I, launched in 1958. It had long, floppy antennas, Note the long, floppy antennas, which tended to dissipare kinetic energy into vibration. It had been designed to spin about its minimimum-moment-of-inertia axis, but almost immediately, as soon as it was in space, it began spinning end over end. It was nevertheless able to carry out its science mission, which didn't depend on being able to maintain a stable orientation, and it discovered the Van Allen radiation belts.

This chapter is summarized on page 726. Notation and terminology are tabulated on pages 718-719.

Problems

The symbols √, , etc. are explained on page 225.

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, p. 715



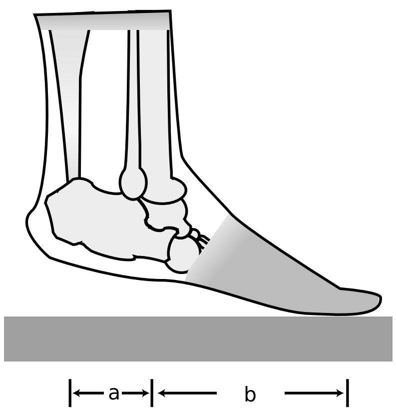
Problem 1.

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? √

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.) √

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.

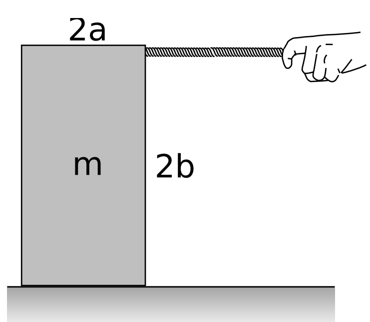
5 An object is observed to have constant angular momentum. Can you conclude that no torques are acting on it? Explain. [Based on a problem by Serway and Faughn.]



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.

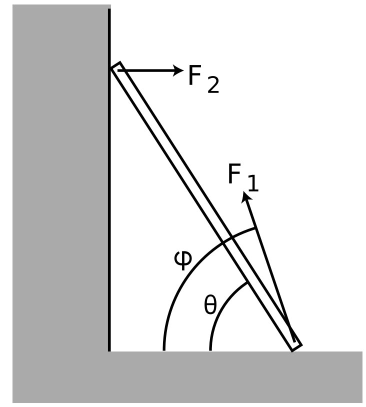
7 Two objects have the same momentum vector. Can you conclude that their angular momenta are the same? Explain. [Based on a problem by Serway and Faughn.]



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?
(b) Repeat part a, but now let the coefficient of kinetic friction be μ_k.
(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°.

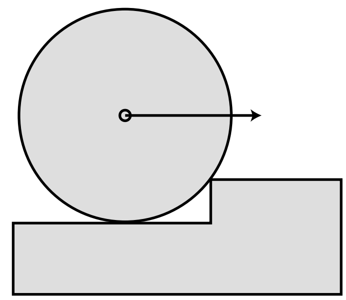


Problems 9 and 10.

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.)

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?

▷ Hint, p. 704



Problem 11.

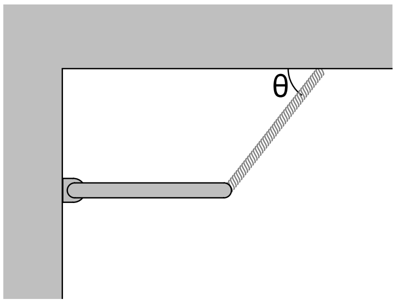
12 A ball is connected by a string to a vertical post. The ball is set in horizontal motion so that it starts winding the string around the post. Assume that the motion is confined to a horizontal plane, i.e. ignore gravity. Michelle and Astrid are trying to predict the final velocity of the ball when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

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

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

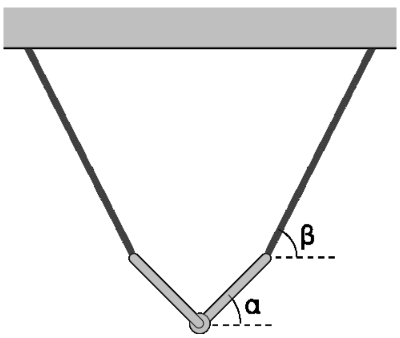
[Note that both factors are negative, giving a positive result.]

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.
(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.



Problem 14.

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.
(b) Using your answer to part a, sketch the configurations for $\beta \to 0$, $\beta = 45\,°$, and $\beta = 90\,°$. Do your results make sense intuitively?

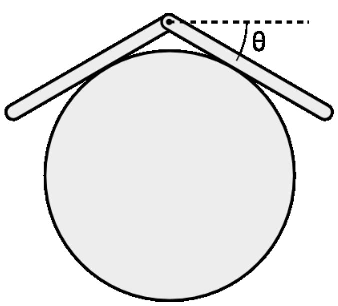


Problem 15.

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

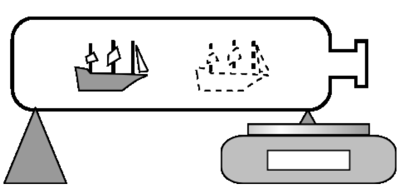
$$\frac{r}{\ell} = \frac{\cos^2\theta}{2\,\tan\theta} \qquad .$$

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



Problem 16.

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 two different locations.


Problem 17.

18 Suppose that we lived in a universe in which Newton's law of gravity gave forces proportional to r^{-7} rather than r^{-2}. 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:

$$K = p^2/2m$$
$$\Delta x = v_o \Delta t + (1/2)a\Delta t^2$$
$$W = F\Delta x$$

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

20 For a one-dimensional harmonic oscillator, the solution to the energy conservation equation,

$$U + K = \frac{1}{2}kx^2 + \frac{1}{2}mv^2 = \text{constant} \qquad ,$$

is an oscillation with frequency $f = (1/2\pi)\sqrt{k/m}$.

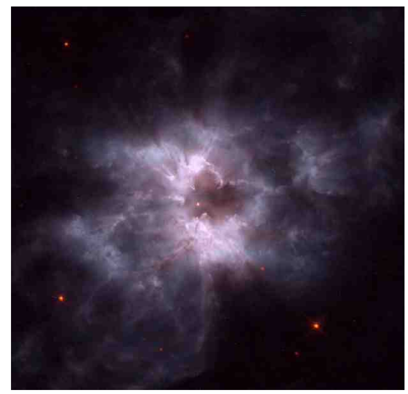
Now consider an analogous system consisting of a bar magnet hung from a thread, which acts like a magnetic compass. A normal compass is full of water, so its oscillations are strongly damped, but the magnet-on-a-thread compass has very little friction, and will oscillate repeatedly around its equilibrium direction. The magnetic energy of the bar magnet is

$$U = -Bm\cos\theta \qquad ,$$

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.

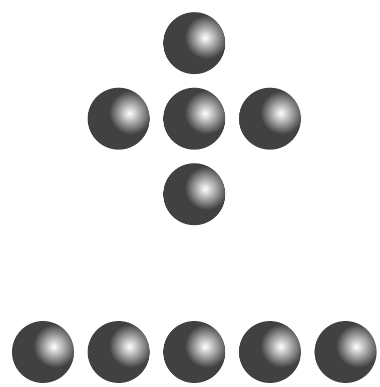
(a) For small θ, the magnetic energy can be approximated by $U \approx (1/2)\kappa\theta^2$. Relate κ to the other quantities. (Assume the thread is so thin that it does not have any significant effect compared to earth's magnetic field.)
(b) Problem 19 gave some examples of how to construct analogies between rotational and linear motion. Use a similar technique to solve for the frequency of the compass's vibrations, stating your result in terms of the variables that will be relevant.

21 (a) Find the angular velocities of the earth's rotation and of the earth's motion around the sun.
(b) Which motion involves the greater acceleration?



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? √
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 every 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? √



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).

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

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

29 Prove property (3) of the vector cross product from the theorem in on page 694.

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 two 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) $(\mathbf{A}\cdot\mathbf{B})\times\mathbf{C}$

33 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.

34 Find the moment of inertia of a rectangular box of mass M whose sides are of length a, b, and c, for rotation about an axis through its center parallel to the edges of length a.

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} kg $\cdot$ m^2/s. The table shows some of the observed angular momenta and energies of ^{168}Er, in SI units (kg $\cdot$ m^2/s and joules).

$L \times 10^{34}$	$E \times 10^{17}$
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.
(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 fm $= 10^{-15}$ m), and assume they are closely packed in the nucleus. ▷ Hint, p. 704

36 Find 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. Rather than starting directly from the results given in the text, you'll have an easier time if you us the following reformulation of the moment of inertia matrix: $I_{xx} = \int x^2 \mathrm{d}m$, etc.

37 In example 22 on page 209, prove that if the rod is sufficiently thin, it can be toppled without scraping on the floor.
▷ Solution, p. 715

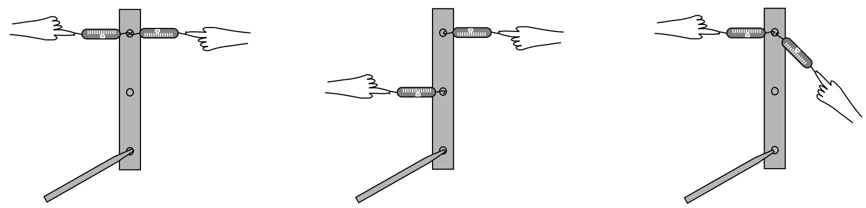
Key to symbols:
easy   typical   challenging   difficult   very difficult
√ An answer check is available at www.lightandmatter.com.

Exercises

Exercise 4A: Torque

Equipment:

- rulers with holes in them
- spring scales (two per group)



While one person holds the pencil which forms the axle for the ruler, the other members of the group pull on the scale and take readings. In each case, calculate the total torque on the ruler, and find out whether it equals zero to roughly within the accuracy of the experiment.

Chapter 5
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.

5.1 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.

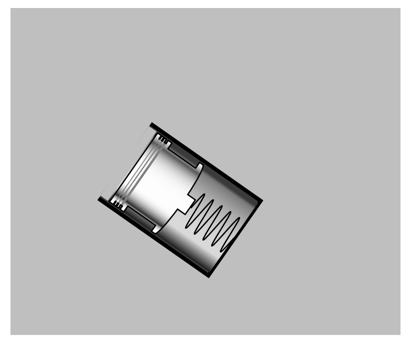
5.1.1 Pressure

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 different in pressure between one part of it and another part. The SI units of pressure are evidently N/m^2, and this combination can be abbreviated as the pascal, 1 Pa=1 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.

Pressure in U.S. units — *example 1*

In U.S. units, the unit of force is the pound, and the unit of distance is the inch. The unit of pressure is therefore pounds per square inch, or p.s.i. (Note that the pound is not a unit of mass.)

Atmospheric pressure in U.S. and metric units — *example 2*

▷ A figure that many people in the U.S. remember is that atmospheric pressure is about 15 pounds per square inch. What is this in metric units? Solution:

$$\begin{aligned}(15\ \text{lb})/(1\ \text{in}^2) &= 1.0\times10^5\ \text{N/m}^2\\ &= 100\ \text{kPa}\end{aligned}$$

Only pressure differences are normally significant.

If you spend enough time on an airplane, the pain in your ears subsides. This is because your body has gradually been able to admit more air into the cavity behind the eardrum. Once the pressure inside is equalized with the pressure outside, the inward and outward forces on your eardrums cancel out, and there is no physical sensation to tell you that anything unusual is going on. For this reason, it is normally only pressure differences that have any physical significance. Thus deep-sea fish are perfectly healthy in their habitat because their bodies have enough internal pressure to cancel the pressure from the water in which they live; if they are caught in a net and brought to the surface rapidly, they explode because their internal pressure is so much greater than the low pressure outside.

Getting killed by a pool pump *example 3*

▷ My house has a pool, which I maintain myself. A pool always needs to have its water circulated through a filter for several hours a day in order to keep it clean. The filter is a large barrel with a strong clamp that holds the top and bottom halves together. My filter has a prominent warning label that warns me not to try to open the clamps while the pump is on, and it shows a cartoon of a person being struck by the top half of the pump. The cross-sectional area of the filter barrel is 0.25 m^2. Like most pressure gauges, the one on my pool pump actually reads the difference in pressure between the pressure inside the pump and atmospheric pressure. The gauge reads 90 kPa. What is the force that is trying to pop open the filter?

▷ If the gauge told us the absolute pressure of the water inside, we'd have to find the force of the water pushing outward and the force of the air pushing inward, and subtract in order to find the total force. Since air surrounds us all the time, we would have to do such a subtraction every time we wanted to calculate anything useful based on the gauge's reading. The manufacturers of the gauge decided to save us from all this work by making it read the difference in pressure between inside and outside, so all we have to do is multiply the gauge reading by the cross-sectional area of the filter:

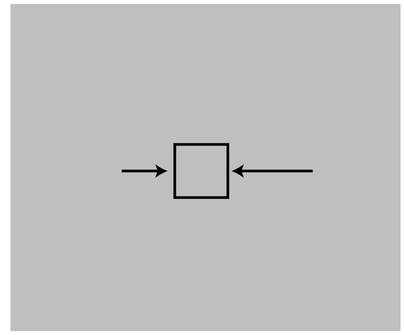
$$\begin{aligned} F &= PA \\ &= (90 \times 10^3\ \text{N}/\text{m}^2)(0.25\ \text{m}^2) \\ &= 22000\ \text{N} \end{aligned}$$

That's a lot of force!

The word "suction" and other related words contain a hidden misunderstanding related to this point about pressure differences. When you suck water up through a straw, there is nothing in your mouth that is attracting the water upward. The force that lifts the water is from the pressure of the water in the cup. By creating a partial vacuum in your mouth, you decreased the air's downward force on the water so that it no longer exactly canceled the upward force.

Variation of pressure with depth

The pressure within a fluid in equilibrium can only depend on depth, due to gravity. If the pressure could vary from side to side, then a piece of the fluid in between, b, would be subject to unequal forces from the parts of the fluid on its two sides. Since fluids do not exhibit shear forces, there would be no other force that could keep this piece of fluid from accelerating. This contradicts the assumption that the fluid was in equilibrium.



b / This doesn't happen. If pressure could vary horizontally in equilibrium, the cube of water would accelerate horizontally. This is a contradiction, since we assumed the fluid was in equilibrium.

Self-Check

How does this proof fail for solids? ▷ Answer, p. 707

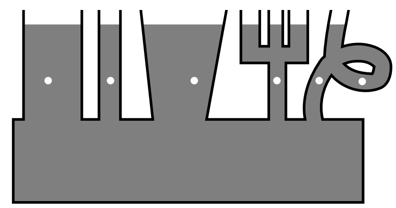
To find the variation with depth, we consider the vertical forces acting on a tiny, imaginary cube of the fluid having infinitesimal height $\mathrm{d}y$ and areas $\mathrm{d}A$ on the top and bottom. Using positive numbers for upward forces, we have

$$P_{bottom}\mathrm{d}A - P_{top}\mathrm{d}A - F_g = 0 \qquad .$$

The weight of the fluid is $F_g = mg = \rho V g = \rho\, \mathrm{d}A\mathrm{d}y\, g$, where ρ is the density of the fluid, so the difference in pressure is

$$\mathrm{d}P = -\rho g \mathrm{d}y \qquad .$$

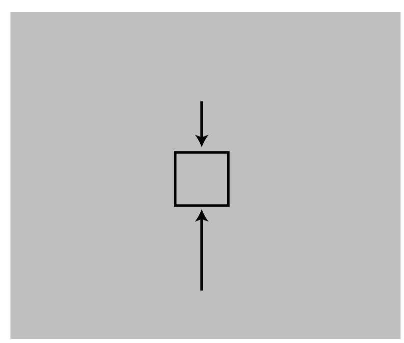
[variation in pressure with depth for a fluid of density ρ in equilibrium; positive y is up.]



c / The pressure is the same at all the points marked with dots.

A more elegant way of writing this is in terms of a dot product, $\mathrm{d}P = \rho\mathbf{g}\cdot\mathrm{d}\mathbf{y}$, which automatically takes care of the plus or minus sign, depending on the relative directions of the $\mathbf{g}$ and $\mathrm{d}\mathbf{y}$ vectors, and avoids any requirements about the coordinate system.

The factor of ρ explains why we notice the difference in pressure when diving 3 m down in a pool, but not when going down 3 m of stairs. The equation only tells us the difference in pressure, not the absolute pressure. The pressure at the surface of a swimming pool equals the atmospheric pressure, not zero, even though the depth is zero at the surface. The blood in your body does not even have an upper surface.



d / This does happen. The sum of the forces from the surrounding parts of the fluid is upward, canceling the downward force of gravity.

In cases where g and ρ are independent of depth, we can integrate both sides of the equation to get everything in terms of finite differences rather than differentials: $\Delta P = -\rho g \Delta y$.

Self-Check

In which of the following situations is the equation $\Delta P = -\rho g \Delta y$ valid? Why?

(1) difference in pressure between a tabletop and the feet (i.e. predicting the pressure of the feet on the floor)

(2) difference in air pressure between the top and bottom of a tall building

(3) difference in air pressure between the top and bottom of Mt. Everest

(4) difference in pressure between the top of the earth's mantle and the

center of the earth
(5) difference in pressure between the top and bottom of an airplane's wing ▷ Answer, p. 707

Pressure of lava underneath a volcano *example 4*
▷ A volcano has just finished erupting, and a pool of molten lava is lying at rest in the crater. The lava has come up through an opening inside the volcano that connects to the earth's molten mantle. The density of the lava is 4.1 g/cm^3. What is the pressure in the lava underneath the base of the volcano, 3000 m below the surface of the pool?

▷

$$\begin{aligned}\Delta P &= \rho g \Delta y \\ &= (4.1\ \text{g/cm}^3)(9.8\ \text{m/s}^2)(3000\ \text{m}) \\ &= (4.1 \times 10^6\ \text{g/m}^3)(9.8\ \text{m/s}^2)(3000\ \text{m}) \\ &= (4.1 \times 10^3\ \text{kg/m}^3)(9.8\ \text{m/s}^2)(3000\ \text{m}) \\ &= 1.2 \times 10^8\ \text{N/m}^2 \\ &= 1.2 \times 10^8\ \text{Pa}\end{aligned}$$

This is the difference between the pressure we want to find and atmospheric pressure at the surface. The latter, however, is tiny compared to the ΔP we just calculated, so what we've found is essentially the pressure, P.

Atmospheric pressure *example 5*
Gases, unlike liquids, are quite compressible, and at a given temperature, the density of a gas is approximately proportional to the pressure. The proportionality constant is discussed on page 237, but for now let's just call it k, $\rho = kP$. Using this fact, we can find the variation of atmospheric pressure with altitude, assuming constant temperature:

$$\begin{aligned}dP &= -\rho g\, dy \\ dP &= -kPg\, dy \\ \frac{dP}{P} &= -kg\, dy \\ \ln P &= -kgy + \text{constant} && \text{[integrating both sides]} \\ P &= (\text{constant})e^{-kgy} && \text{[exponentiating both sides]}\end{aligned}$$

Pressure falls off exponentially with height. There is no sharp cutoff to the atmosphere, but the exponential gets extremely small by the time you're ten or a hundred miles up.

e / We have to wait for the thermometer to equilibrate its temperature with the temperature of Irene's armpit.

5.1.2 Temperature

Thermal equilibrium

We use the term temperature casually, but what is it exactly? Roughly speaking, temperature is a measure of how concentrated the heat energy is in an object. A large, massive object with very little heat energy in it has a low temperature.

But physics deals with operational definitions, i.e. definitions of how to measure the thing in question. How do we measure temperature? One common feature of all temperature-measuring devices is that they must be left for a while in contact with the thing whose temperature is being measured. When you take your temperature with a fever thermometer, you are waiting for the mercury inside to come up to the same temperature as your body. The thermometer actually tells you the temperature of its own working fluid (in this case the mercury). In general, the idea of temperature depends on the concept of thermal equilibrium. When you mix cold eggs from the refrigerator with flour that has been at room temperature, they rapidly reach a compromise temperature. What determines this compromise temperature is conservation of energy, and the amount of energy required to heat or cool each substance by one degree. But without even having constructed a temperature scale, we can see that the important point is the phenomenon of thermal equilibrium itself: two objects left in contact will approach the same temperature. We also assume that if object A is at the same temperature as object B, and B is at the same temperature as C, then A is at the same temperature as C. This statement is sometimes known as the zeroth law of thermodynamics, so called because after the first, second, and third laws had been developed, it was realized that there was another law that was even more fundamental.

f / Thermal equilibrium can be prevented. Otters have a coat of fur that traps air bubbles for insulation. If a swimming otter was in thermal equilibrium with cold water, it would be dead. Heat is still conducted from the otter's body to the water, but much more slowly than it would be in a warm-blooded animal that didn't have this special adaptation.

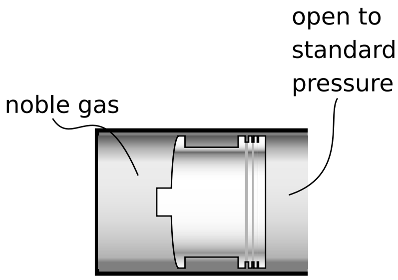
Thermal expansion

The familiar mercury thermometer operates on the principle that the mercury, its working fluid, expands when heated and contracts when cooled. In general, all substances expand and contract with changes in temperature. The zeroth law of thermodynamics guarantees that we can construct a comparative scale of temperatures that is independent of what type of thermometer we use. If a thermometer gives a certain reading when it's in thermal equilibrium with object A, and also gives the same reading for object B, then A and B must be the same temperature, regardless of the details of how the thermometers works.

g / A hot air balloon is inflated. Because of thermal expansion, the hot air is less dense than the surrounding cold air, and therefore floats as the cold air drops underneath it and pushes it up out of the way.

What about constructing a temperature scale in which every degree represents an equal step in temperature? The Celsius scale has 0 as the freezing point of water and 100 as its boiling point. The hidden assumption behind all this is that since two points define a line, any two thermometers that agree at two points must agree at all other points. In reality if we calibrate a mercury thermometer and an alcohol thermometer in this way, we will find that a graph of one thermometer's reading versus the other is not a perfectly straight $y = x$ line. The subtle inconsistency becomes a drastic one when we try to extend the temperature scale through the points where mercury and alcohol boil or freeze. Gases, however, are much more consistent among themselves in their thermal expansion than solids or liquids, and the noble gases like helium and neon are more consistent with each other than gases in general. Continuing to search for consistency, we find that noble gases are more consistent with each other when their pressure is very low.

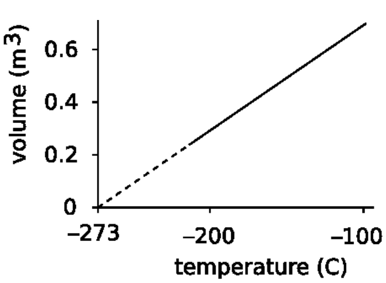
As an idealization, we imagine a gas in which the atoms interact only with the sides of the container, not with each other. Such a gas is perfectly nonreactive (as the noble gases very nearly are), and never condenses to a liquid (as the noble gases do only at extremely low temperatures). Its atoms take up a negligible fraction of the available volume. Any gas can be made to behave very much like this if the pressure is extremely low, so that the atoms hardly ever encounter each other. Such a gas is called an ideal gas, and we define the Celsius scale in terms of the volume of the gas in a thermometer whose working substance is an ideal gas maintained at a fixed (very low) pressure, and which is calibrated at 0 and 100 degrees according to the melting and boiling points of water. The Celsius scale is not just a comparative scale but an additive one as well: every step in temperature is equal, and it makes sense to say that the difference in temperature between 18 and 28 °C is the same as the difference between 48 and 58.



h / A simplified version of an ideal gas thermometer. The whole instrument is allowed to come into thermal equilibrium with the substance whose temperature is to be measured, and the mouth of the cylinder is left open to standard pressure. The volume of the noble gas gives an indication of temperature.

Absolute zero and the Kelvin scale

We find that if we extrapolate a graph of volume versus temperature, the volume becomes zero at nearly the same temperature for all gases: -273 °C. Real gases will all condense into liquids at some temperature above this, but an ideal gas would achieve zero volume at this temperature, known as absolute zero. The most useful temperature scale in scientific work is one whose zero is defined by absolute zero, rather than by some arbitrary standard like the melting point of water. The temperature scale used universally in scientific work, called the Kelvin scale, is the same as the Celsius scale, but shifted by 273 degrees to make its zero coincide with absolute zero. Scientists use the Celsius scale only for comparisons or when a change in temperature is all that is required for a calculation. Only on the Kelvin scale does it make sense to discuss ratios of temperatures, e.g. to say that one temperature is twice as hot as another.



i / The volume of 1 kg of neon gas as a function of temperature (at standard pressure). Although neon would actually condense into a liquid at some point, extrapolating the graph gives to zero volume gives the same temperature as for any other gas: absolute zero.

Which temperature scale to use *example 6*

▷ You open an astronomy book and encounter the equation

$$\text{(light emitted)} = \text{(constant)} \times T^4$$

for the light emitted by a star as a function of its surface temperature. What temperature scale is implied?

▷ The equation tells us that doubling the temperature results in the emission of 16 times as much light. Such a ratio only makes sense if the Kelvin scale is used.

Although we can achieve as good an approximation to an ideal gas as we wish by making the pressure very low, it seems nevertheless that there should be some more fundamental way to define temperature. We will construct a more fundamental scale of temperature in section 5.4.

5.2 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

$$PV \propto nT, \qquad \text{[ideal gas]}$$

which really only holds exactly for an ideal gas. Here n is the number of molecules in the sample of gas.

Volume related to temperature *example 7*
The proportionality of volume to temperature at fixed pressure was the basis for our definition of temperature.

Pressure related to temperature *example 8*
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

$$F = \sum \frac{\Delta p_{x,i}}{\Delta t_i} \qquad \text{[monoatomic ideal gas]} \qquad ,$$

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

$$F = \sum \frac{mv_{x,i}^2}{L} \qquad \text{[monoatomic ideal gas]} \qquad .$$

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. By symmetry, it must therefore equal $2/3$ of the total kinetic energy, so

$$F = \frac{2K_{total}}{3L} \qquad \text{[monoatomic ideal gas]} \qquad .$$

Dividing by A and using $AL = V$, we have

$$P = \frac{2K_{total}}{3V} \qquad \text{[monoatomic ideal gas]} \qquad .$$

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:

$$PV = \frac{2}{3}n\bar{K} \qquad \text{[monoatomic ideal gas]} \qquad .$$

For the first time we have an interpretation for the 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

$$PV = nkT \qquad , \qquad \text{[ideal gas]}$$

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.)

Pressure in a car tire *example 9*

▷ 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,

$$\begin{aligned} P_2/P_1 &= T_2/T_1 \\ &= (308\text{ K})/(283\text{ K}) \\ &= 1.09 \qquad , \end{aligned}$$

or a 9% increase.

Discussion Questions

A 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.)

B Repeat discussion question A if the comparison is 1 kg of helium versus 1 kg of xenon (equal masses, rather than equal volumes).

C 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.)

5.3 Entropy as a Macroscopic Quantity

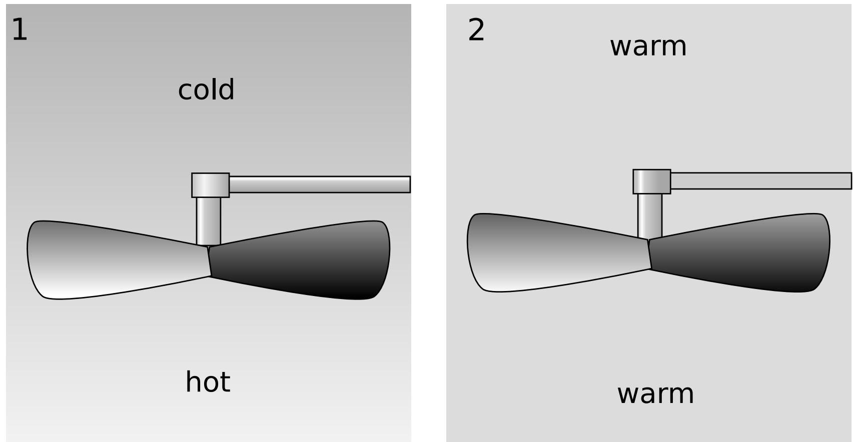
5.3.1 Efficiency and grades of energy

Some forms of energy are more convenient than others in certain situations. You can't run a spring-powered mechanical clock on a battery, and you can't run a battery-powered clock with mechanical energy. However, there is no fundamental physical principle that prevents you from converting 100% of the electrical energy in a battery into mechanical energy or vice-versa. More efficient motors and generators are being designed every year. In general, the laws of physics permit perfectly efficient conversion within a broad class of forms of energy.

Heat is different. Friction tends to convert other forms of energy into heat even in the best lubricated machines. When we slide a book on a table, friction brings it to a stop and converts all its kinetic energy into heat, but we never observe the opposite process, in which a book spontaneously converts heat energy into mechanical energy and starts moving! Roughly speaking, heat is different because it is disorganized. Scrambling an egg is easy. Unscrambling it is harder.

We summarize these observations by saying that heat is a lower grade of energy than other forms such as mechanical energy.

Of course it is possible to convert heat into other forms of energy such as mechanical energy, and that is what a car engine does with the heat created by exploding the air-gasoline mixture. But a car engine is a tremendously inefficient device, and a great deal of the heat is simply wasted through the radiator and the exhaust. Engineers have never succeeded in creating a perfectly efficient device for converting heat energy into mechanical energy, and we now know that this is because of a deeper physical principle that is far more basic than the design of an engine.



a / 1. The temperature difference between the hot and cold parts of the air can be used to extract mechanical energy, for example with a fan blade that spins because of the rising hot air currents. 2. If the temperature of the air is first allowed to become uniform, then no mechanical energy can be extracted. The same amount of heat energy is present, but it is no longer accessible for doing mechanical work.

5.3.2 Heat engines

Heat may be more useful in some forms than in other, i.e., there are different grades of heat energy. In figure a/1, the difference in temperature can be used to extract mechanical work with a fan blade. This principle is used in power plants, where steam is heated by burning oil or by nuclear reactions, and then allowed to expand through a turbine which has cooler steam on the other side. On a smaller scale, there is a Christmas toy, b, that consists of a small propeller spun by the hot air rising from a set of candles, very much like the setup shown in the figure.

b / A heat engine. Hot air from the candles rises through the fan blades, and makes the angels spin.

In figure a/2, however, no mechanical work can be extracted because there is no difference in temperature. Although the air in a/2 has the same total amount of energy as the air in a/1, the heat in a/2 is a lower grade of energy, since none of it is accessible for doing mechanical work.

In general, we define a heat engine as any device that takes heat from a reservoir of hot matter, extracts some of the heat energy to do mechanical work, and expels a lesser amount of heat into a reservoir of cold matter. The efficiency of a heat engine equals the amount of useful work extracted, W, divided by the amount of energy we had to pay for in order to heat the hot reservoir. This latter amount of heat is the same as the amount of heat the engine extracts from the high-temperature reservoir, Q_H. (The letter Q is the standard notation for a transfer of heat.) By conservation of energy, we have $Q_H = W + Q_L$, where Q_L is the amount of heat expelled into the low-temperature reservoir, so the efficiency of a heat engine, W/Q_H, can be rewritten as

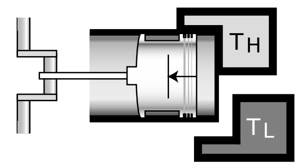
$$\text{efficiency} = 1 - \frac{Q_L}{Q_H} \qquad . \qquad \text{[efficiency of any heat engine]}$$

c / Sadi Carnot (1796-1832)

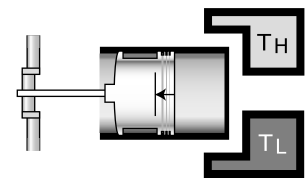
It turns out that there is a particular type of heat engine, the Carnot engine, which, although not 100% efficient, is more efficient than any other. The grade of heat energy in a system can thus be unambiguously defined in terms of the amount of heat energy in it that cannot be extracted even by a Carnot engine.

How can we build the most efficient possible engine? Let's start with an unnecessarily inefficient engine like a car engine and see how it could be improved. The radiator and exhaust expel hot gases, which is a waste of heat energy. These gases are cooler than the exploded air-gas mixture inside the cylinder, but hotter than the air that surrounds the car. We could thus improve the engine's efficiency by adding an auxiliary heat engine to it, which would operate with the first engine's exhaust as its hot reservoir and the air as its cold reservoir. In general, any heat engine that expels heat at an intermediate temperature can be made more efficient by changing it so that it expels heat only at the temperature of the cold reservoir.

Similarly, any heat engine that absorbs some energy at an intermediate temperature can be made more efficient by adding an auxiliary heat engine to it which will operate between the hot reservoir and this intermediate temperature.



d / The beginning of the first expansion stroke, in which the working gas is kept in thermal equilibrium with the hot reservoir.

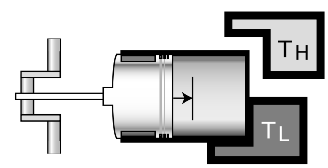


e / The beginning of the second expansion stroke, in which the working gas is thermally insulated. The working gas cools because it is doing work on the piston and thus losing energy.

Based on these arguments, we define a Carnot engine as a heat engine that absorbs heat only from the hot reservoir and expels it only into the cold reservoir. Figures d-g show a realization of a Carnot engine using a piston in a cylinder filled with a monoatomic ideal gas. This gas, known as the working fluid, is separate from, but exchanges energy with, the hot and cold reservoirs. As proved on page 697, this particular Carnot engine has an efficiency given by

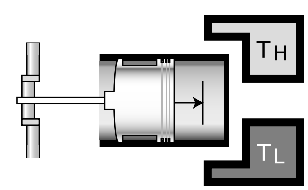
$$\text{efficiency} = 1 - \frac{T_L}{T_H} \qquad , \qquad \text{[efficiency of a Carnot engine]}$$

where T_L is the temperature of the cold reservoir and T_H is the temperature of the hot reservoir.



f / The beginning of the first compression stroke. The working gas begins the stroke at the same temperature as the cold reservoir, and remains in thermal contact with it the whole time. The engine does negative work.

Even if you do not wish to dig into the details of the proof, the basic reason for the temperature dependence is not so hard to understand. Useful mechanical work is done on strokes d and e, in which the gas expands. The motion of the piston is in the same direction as the gas's force on the piston, so positive work is done on the piston. In strokes f and g, however, the gas does negative work on the piston. We would like to avoid this negative work, but we must design the engine to perform a complete cycle. Luckily the pressures during the compression strokes are lower than the ones during the expansion strokes, so the engine doesn't undo all its work with every cycle. The ratios of the pressures are in proportion to the ratios of the temperatures, so if T_L is 20% of T_H, the engine is 80% efficient.



g / The beginning of the second compression stroke, in which mechanical work is absorbed, heating the working gas back up to T_H.

We have already proved that any engine that is not a Carnot engine is less than optimally efficient, and it is also true that all

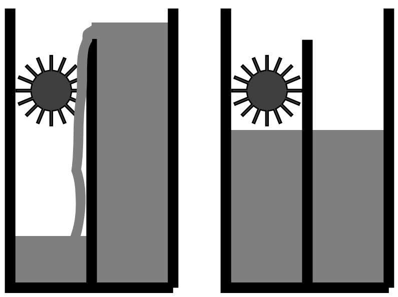
Carnot engines operating between a given pair of temperatures T_H and T_L have the same efficiency. (This can be proved by the methods of section 5.4.) Thus a Carnot engine is the most efficient possible heat engine.

5.3.3 Entropy

We would like to have some numerical way of measuring the grade of energy in a system. We want this quantity, called entropy, to have the following two properties:

(1) Entropy is additive. When we combine two systems and consider them as one, the entropy of the combined system equals the sum of the entropies of the two original systems. (Quantities like mass and energy also have this property.)

(2) The entropy of a system is not changed by operating a Carnot engine within it.



h / Entropy can be understood using the metaphor of a water wheel. Letting the water levels equalize is like letting the entropy maximize. Taking water from the high side and putting it into the low side increases the entropy. Water levels in this metaphor correspond to temperatures in the actual definition of entropy.

It turns out to be simpler and more useful to define changes in entropy than absolute entropies. Suppose as an example that a system contains some hot matter and some cold matter. It has a relatively high grade of energy because a heat engine could be used to extract mechanical work from it. But if we allow the hot and cold parts to equilibrate at some lukewarm temperature, the grade of energy has gotten worse. Thus putting heat into a hotter area is more useful than putting it into a cold area. Motivated by these considerations, we define a change in entropy as follows:

$$\Delta S = \frac{Q}{T} \quad \text{[change in entropy when adding heat } Q \text{ to matter at temperature } T\text{; } \Delta S \text{ is negative if heat is taken out]}$$

A system with a higher grade of energy has a lower entropy.

Entropy is additive. *example 10*

Since changes in entropy are defined by an additive quantity (heat) divided by a non-additive one (temperature), entropy is additive.

Entropy isn't changed by a Carnot engine. *example 11*

The efficiency of a heat engine is defined by

$$\text{efficiency} = 1 - Q_L/Q_H \quad ,$$

and the efficiency of a Carnot engine is

$$\text{efficiency} = 1 - T_L/T_H \quad ,$$

so for a Carnot engine we have $Q_L/Q_H = T_L/T_H$, which can be rewritten as $Q_L/T_L = Q_H/T_H$. The entropy lost by the hot reservoir is therefore the same as the entropy gained by the cold one.

Entropy increases in heat conduction. *example 12*

When a hot object gives up energy to a cold one, conservation of energy tells us that the amount of heat lost by the hot object is the same as the amount of heat gained by the cold one. The change in entropy is $-Q/T_H + Q/T_L$, which is positive because $T_L < T_H$.

Entropy is increased by a non-Carnot engine. *example 13*

The efficiency of a non-Carnot engine is less than 1 - T_L/T_H, so $Q_L/Q_H > T_L/T_H$ and $Q_L/T_L > Q_H/T_H$. This means that the entropy increase in the cold reservoir is greater than the entropy decrease in the hot reservoir.

A book sliding to a stop *example 14*

A book slides across a table and comes to a stop. Once it stops, all its kinetic energy has been transformed into heat. As the book and table heat up, their entropies both increase, so the total entropy increases as well.

All of these examples involved closed systems, and in all of them the total entropy either increased or stayed the same. It never decreased. Here are two examples of schemes for decreasing the entropy of a closed system, with explanations of why they don't work.

Using a refrigerator to decrease entropy? *example 15*

▷ A refrigerator takes heat from a cold area and dumps it into a hot area. (1) Does this lead to a net decrease in the entropy of a closed system? (2) Could you make a Carnot engine more efficient by running a refrigerator to cool its low-temperature reservoir and eject heat into its high-temperature reservoir?

▷ (1) No. The heat that comes off of the radiator coils is a great deal more than the heat the fridge removes from inside; the difference is what it costs to run your fridge. The heat radiated from the coils is so much more than the heat removed from the inside that the increase in the entropy of the air in the room is greater than the decrease of the entropy inside the fridge. The most efficient refrigerator is actually a Carnot engine running in reverse, which leads to neither an increase nor a decrease in entropy.

(2) No. The most efficient refrigerator is a reversed Carnot engine. You will not achieve anything by running one Carnot engine in reverse and another forward. They will just cancel each other out.

Maxwell's demon *example 16*

▷ Maxwell imagined a pair of rooms, their air being initially in thermal equilibrium, having a partition across the middle with a tiny door. A miniscule demon is posted at the door with a little ping-pong paddle, and his duty is to try to build up faster-moving air molecules in room B and slower moving ones in room A. For instance, when a fast molecule is headed through the door, going from A to B, he lets it by, but when a slower than average molecule tries the same thing, he hits it back into room A. Would this decrease the total entropy of the pair of rooms?

▷ No. The demon needs to eat, and we can think of his body as a little heat engine, and his metabolism is less efficient than a Carnot engine, so he ends up increasing the entropy rather than decreasing it.

Observations such as these lead to the following hypothesis, known as the second law of thermodynamics:

The entropy of a closed system always increases, or at best stays the same: $\Delta S \geq 0$.

At present our arguments to support this statement may seem less than convincing, since they have so much to do with obscure facts about heat engines. In the following section we will find a more satisfying and fundamental explanation for the continual increase in entropy. To emphasize the fundamental and universal nature of the second law, here are a few exotic examples.

Entropy and evolution *example 17*

A favorite argument of many creationists who don't believe in evolution is that evolution would violate the second law of thermodynamics: the death and decay of a living thing releases heat (as when a compost heap gets hot) and lessens the amount of energy available for doing useful work, while the reverse process, the emergence of life from nonliving matter, would require a decrease in entropy. Their argument is faulty, since the second law only applies to closed systems, and the earth is not a closed system. The earth is continuously receiving energy from the sun.

The heat death of the universe *example 18*

Living things have low entropy: to demonstrate this fact, observe how a compost pile releases heat, which then equilibrates with the cooler environment. We never observe dead things to leap back to life after sucking some heat energy out of their environments! The only reason life was able to evolve on earth was that the earth was not a closed system: it got energy from the sun, which presumably gained more entropy than the earth lost.

Victorian philosophers spent a lot of time worrying about the heat death of the universe: eventually the universe would have to become a high-entropy, lukewarm soup, with no life or organized motion of any kind. Fortunately (?), we now know a great many other things that will make the universe inhospitable to life long before its entropy is maximized. Life on earth, for instance, will end when the sun evolves into a giant star and vaporizes our planet.

Hawking radiation *example 19*

Any process that could destroy heat (or convert it into nothing but mechanical work) would lead to a reduction in entropy. Black holes are supermassive stars whose gravity is so strong that nothing, not even light, can escape from them once it gets within a boundary known as the event horizon. Black holes are commonly observed to suck hot gas into them. Does this lead to a reduction in the entropy of the universe? Of course one could argue that the entropy is still there inside the black hole, but being able to "hide" entropy there amounts to the same thing as being able to destroy entropy.

The physicist Steven Hawking was bothered by this question, and finally realized that although the actual stuff that enters a black hole is lost forever, the black hole will gradually lose energy in the form of light emitted from just outside the event horizon. This light ends up reintroducing the original entropy back into the universe at large.

Discussion Question

A In this discussion question, you'll think about a car engine in terms of thermodynamics. Note that an internal combustion engine doesn't fit very well into the theoretical straightjacket of a heat engine. For instance, a heat engine has a high-temperature heat reservoir at a single well-defined temperature, T_H. In a typical car engine, however, there are several very different temperatures you could imagine using for T_H: the temperature of the engine block ($\sim 100\,^\circ\text{C}$), the walls of the cylinder ($\sim 250\,^\circ\text{C}$), or the temperature of the exploding air-gas mixture ($\sim 1000\,^\circ\text{C}$, with significant changes over a four-stroke cycle). Let's use $T_H \sim 1000\,^\circ\text{C}$.

Burning gas supplies heat energy Q_H to your car's engine. The engine does mechanical work W, but also expels heat Q_L into the environment through the radiator and the exhaust. Conservation of energy gives

$$Q_H = Q_L + W \qquad ,$$

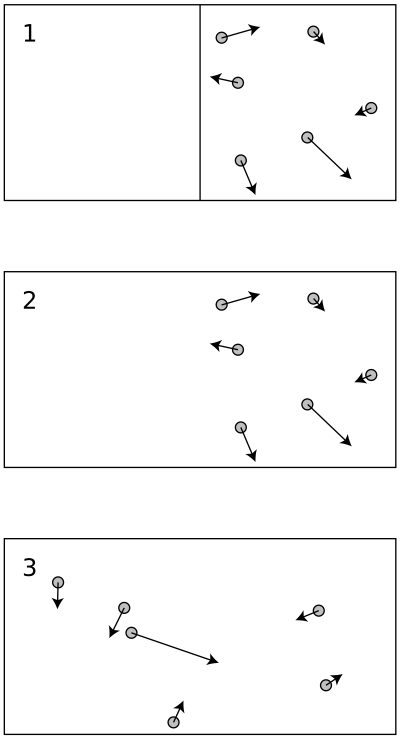
and the relative proportions of Q_L and W are usually about 90% to 10%. (Actually it depends quite a bit on the type of car, the driving conditions, etc.)

(1) Q_L is obviously undesirable: you pay for it, but all it does is heat the neighborhood. Suppose that engineers do a really good job of getting rid of the effects that create Q_L, such as friction. Could Q_L ever be reduced to zero, at least theoretically?

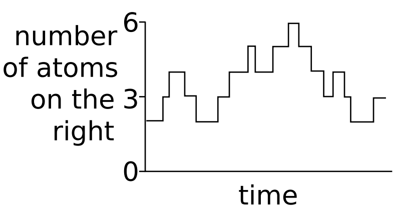
(2) A gallon of gas releases about 140 MJ of heat Q_H when burned. Estimate the change in entropy of the universe due to running a typical car engine and burning one gallon of gas. (You'll have to estimate how hot the environment is. For the sake of argument, assume that the work done by the engine, W, remains in the form of mechanical energy, although in reality it probably ends up being changed into heat when you step on the brakes.) Is your result consistent with the second law of thermodynamics?

(3) What would happen if you redid the calculation in #2, but assumed $Q_L = 0$? Is this consistent with your answer to #1?

B Example 11 on page 242 showed that entropy isn't changed by a Carnot engine, by using the expression for the efficiency of a Carnot engine. Prove the same thing by applying the definition of entropy to each of the four strokes of the Carnot engine.



a / A gas expands freely, doubling its volume.



b / An unusual fluctuation in the distribution of the atoms between the two sides of the box. There has been no external manipulation as in figure a/1.

5.4 Entropy as a Microscopic Quantity (Optional)

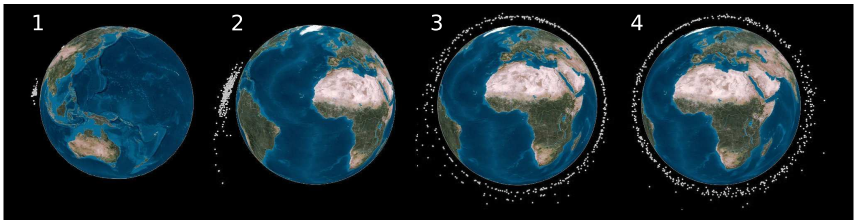
5.4.1 A microscopic view of entropy

To understand why the second law of thermodynamics is always true, we need to see what entropy really means at the microscopic level. An example that is easy to visualize is the free expansion of a monoatomic gas. Figure a/1 shows a box in which all the atoms of the gas are confined on one side. We very quickly remove the barrier between the two sides, a/2, and some time later, the system has reached an equilibrium, a/3. Each snapshot shows both the positions and the momenta of the atoms, which is enough information to allow us in theory to extrapolate the behavior of the system into the future, or the past. However, with a realistic number of atoms, rather than just six, this would be beyond the computational power of any computer.[1]

But suppose we show figure a/2 to a friend without any further information, and ask her what she can say about the system's behavior in the future. She doesn't know how the system was prepared. Perhaps, she thinks, it was just a strange coincidence that all the atoms happened to be in the right half of the box at this particular moment. In any case, she knows that this unusual situation won't last for long. She can predict that after the passage of any significant amount of time, a surprise inspection is likely to show roughly half the atoms on each side. The same is true if you ask her to say what happened in the past. She doesn't know about the barrier, so as far as she's concerned, extrapolation into the past is exactly the same kind of problem as extrapolation into the future. We just have to imagine reversing all the momentum vectors, and then all our reasoning works equally well for backwards extrapolation. She would conclude, then, that the gas in the box underwent an unusual fluctuation, b, and she knows that the fluctuation is very unlikely to exist very far into the future, or to have existed very far into the past.

What does this have to do with entropy? Well, state a/3 has a greater entropy than state a/2. It would be easy to extract mechanical work from a/2, for instance by letting the gas expand while pressing on a piston rather than simply releasing it suddenly into the void. There is no way to extract mechanical work from state a/3. Roughly speaking, our microscopic description of entropy relates to the *number of possible states*. There are a lot more states like a/3 than there are states like a/2. Over long enough periods of time — long enough for equilibration to occur — the system gets mixed up, and is about equally likely to be in any of its possible states, regardless of what state it was initially in. We define some number

[1]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.



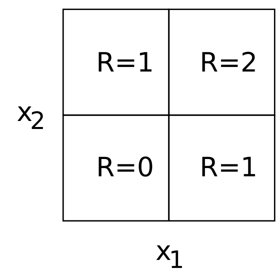
c / Earth orbit is becoming cluttered with space junk, and the pieces can be thought of as the "molecules" comprising an exotic kind of gas. These image shows the evolution of a cloud of debris arising from a 2007 Chinese test of an anti-satellite rocket. Panels 1-4 show the cloud five minutes, one hour, one day, and one month after the impact. The entropy seems to have maximized by panel 4.

that describes an interesting property of the whole system, say the number of atoms in the right half of the box, R. A high-entropy value of R is one like $R = 3$, which allows many possible states. We are far more likely to encounter $R = 3$ than a low-entropy value like $R = 0$ or $R = 6$.

5.4.2 Phase space

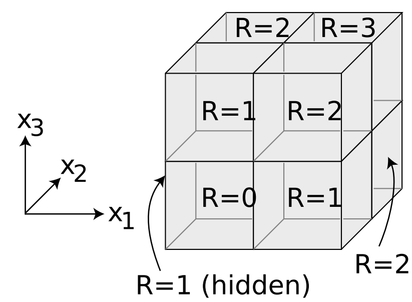
There is a problem with making this description of entropy into a mathematical definition. The problem is that it refers to the number of possible states, but that number is theoretically infinite. To get around the problem, we coarsen our description of the system. For the atoms in figure a, we don't really care exactly where each atom is. We only care whether it is in the right side or the left side. If a particular atom's left-right position is described by a coordinate x, then the set of all possible values of x is a line segment along the x axis, containing an infinite number of points. We break this line segment down into two halves, each of width Δx, and we consider two different values of x to be variations on the same state if they both lie in the same half. For our present purposes, we can also ignore completely the y and z coordinates, and all three momentum components, p_x, p_y, and p_z.

Now let's do a real calculation. Suppose there are only two atoms in the box, with coordinates x_1 and x_2. We can give all the relevant information about the state of the system by specifying one of the cells in the grid shown in figure d. This grid is known as the *phase space* of the system. The lower right cell, for instance, describes a state in which atom number 1 is in the right side of the box and atom number 2 in the left. Since there are two possible states with $R = 1$ and only one state with $R = 2$, we are twice as likely to observe $R = 1$, and $R = 1$ has higher entropy than $R = 2$.

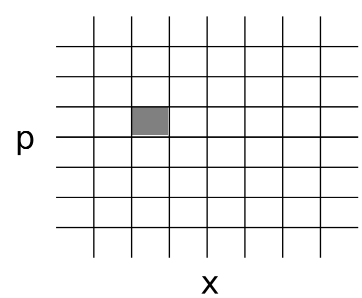


d / The phase space for two atoms in a box.

Figure e shows a corresponding calculation for three atoms, which makes the phase space three-dimensional. Here, the $R = 1$ and 2 states are three times more likely than $R = 0$ and 3. Four atoms would require a four-dimensional phase space, which exceeds our ability to visualize. Although our present example doesn't require it, a phase space can describe momentum as well as position, as shown in figure f. In general, a phase space for a monoatomic gas has six dimensions per atom (one for each coordinate and one for each momentum component).



e / The phase space for three atoms in a box.



f / A phase space for a single atom in one dimension, taking momentum into account.

5.4.3 Microscopic definitions of entropy and temperature

Two more issues need to be resolved in order to make a microscopic definition of entropy.

First, if we defined entropy as the number of possible states, it would be a multiplicative quantity, not an additive one: if an ice cube in a glass of water has M_1 states available to it, and the number of states available to the water is M_2, then the number of possible states of the whole system is the product $M_1 M_2$. To get around this problem, we take the natural logarithm of the number of states, which makes the entropy additive because of the property of the logarithm $\ln(M_1 M_2) = \ln M_1 + \ln M_2$.

g / Ludwig Boltzmann's tomb, inscribed with his equation for entropy.

The second issue is a more trivial one. The concept of entropy was originally invented as a purely macroscopic quantity, and the macroscopic definition $\Delta S = Q/T$, which has units of J/K, has a different calibration than would result from defining $S = \ln M$. The calibration constant we need turns out to be simply the Boltzmann constant, k.

Microscopic definition of entropy: The entropy of a system is $S = k \ln M$, where M is the number of available states.[2]

This also leads to a more fundamental definition of temperature. Two systems are in thermal equilibrium when they have maximized their combined entropy through the exchange of energy. Here the energy possessed by one part of the system, E_1 or E_2, plays the same role as the variable R in the examples of free expansion above. A maximum of a function occurs when the derivative is zero, so the maximum entropy occurs when

$$\frac{\mathrm{d}\left(S_1 + S_2\right)}{\mathrm{d}E_1} = 0 \qquad .$$

We assume the systems are only able to exchange heat energy with each other, $\mathrm{d}E_1 = -\mathrm{d}E_2$, so

$$\frac{\mathrm{d}S_1}{\mathrm{d}E_1} = \frac{\mathrm{d}S_2}{\mathrm{d}E_2} \qquad ,$$

and since the energy is being exchanged in the form of heat we can make the equations look more familiar if we write $\mathrm{d}Q$ for an amount of heat to be transferred into either system:

$$\frac{\mathrm{d}S_1}{\mathrm{d}Q_1} = \frac{\mathrm{d}S_2}{\mathrm{d}Q_2} \qquad .$$

In terms of our previous definition of entropy, this is equivalent to $1/T_1 = 1/T_2$, which makes perfect sense since the systems are in thermal equilibrium. According to our new approach, entropy has

[2]This is the same relation as the one on Boltzmann's tomb, just in a slightly different notation.

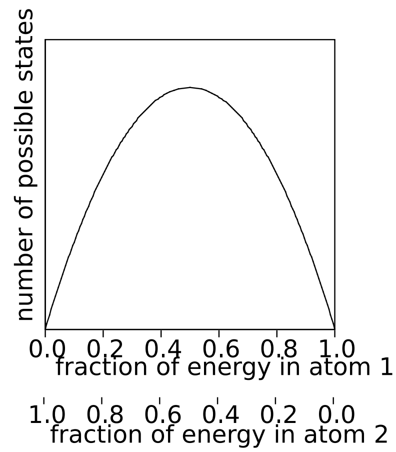
already been defined in a fundamental manner, so we can take this as a definition of temperature:

$$\frac{1}{T} = \frac{\mathrm{d}S}{\mathrm{d}Q} \qquad ,$$

where dS represents the increase in the system's entropy from adding heat dQ to it.

Examples with small numbers of atoms

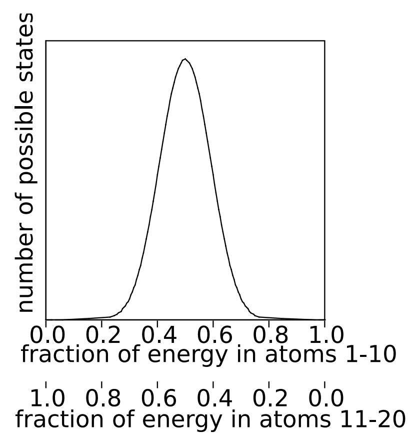
Let's see how this applies to an ideal, monoatomic gas with a small number of atoms. To start with, consider the phase space available to one atom. Since we assume the atoms in an ideal gas are noninteracting, their positions relative to each other are really irrelevant. We can therefore enumerate the number of states available to each atom just by considering the number of momentum vectors it can have, without considering its possible locations. The relationship between momentum and kinetic energy is $E = (p_x^2+p_y^2+p_z^2)/2m$, so if for a fixed value of its energy, we arrange all of an atom's possible momentum vectors with their tails at the origin, their tips all lie on the surface of a sphere in phase space with radius $|\mathbf{p}| = \sqrt{2mE}$. The number of possible states for that atom is proportional to the sphere's surface area, which in turn is proportional to the square of the sphere's radius, $|\mathbf{p}|^2 = 2mE$.



h / A two-atom system has the highest number of available states when the energy is equally divided. Equal energy division is therefore the most likely possibility at any given moment in time.

Now consider two atoms. For any given way of sharing the energy between the atoms, $E = E_1 + E_2$, the number of possible combinations of states is proportional to E_1E_2. The result is shown in figure h. The greatest number of combinations occurs when we divide the energy equally, so an equal division gives maximum entropy.

By increasing the number of atoms, we get a graph whose peak is narrower, i. With more than one atom in each system, the total energy is $E = (p_{x,1}^2 + p_{y,1}^2 + p_{z,1}^2 + p_{x,2}^2 + p_{y,2}^2 + p_{z,2}^2 + ...)/2m$. With n atoms, a total of $3n$ momentum coordinates are needed in order to specify their state, and such a set of numbers is like a single point in a $3n$-dimensional space (which is impossible to visualize). For a given total energy E, the possible states are like the surface of a $3n$-dimensional sphere, with a surface area proportional to p^{3n-1}, or $E^{(3n-1)/2}$. The graph in figure i, for example, was calculated according to the formula $E_1^{29/2}E_2^{29/2} = E_1^{29/2}(E-E_1)^{29/2}$.



i / When two systems of 10 atoms each interact, the graph of the number of possible states is narrower than with only one atom in each system.

Since graph i is narrower than graph h, the fluctuations in energy sharing are smaller. If we inspect the system at a random moment in time, the energy sharing is very unlikely to be more lopsided than a 40-60 split. Now suppose that, instead of 10 atoms interacting with 10 atoms, we had a 10^{23} atoms interacting with 10^{23} atoms. The graph would be extremely narrow, and it would be a statistical certainty that the energy sharing would be nearly perfectly equal. This is why we never observe a cold glass of water to change itself into an ice cube sitting in some warm water!

By the way, note that although we've redefined temperature, these examples show that things are coming out consistent with the old definition, since we saw that the old definition of temperature could be described in terms of the average energy per atom, and here we're finding that equilibration results in each subset of the atoms having an equal share of the energy.

Entropy of a monoatomic ideal gas

Let's calculate the entropy of a monoatomic ideal gas of n atoms. This is an important example because it allows us to show that our present microscopic treatment of thermodynamics is consistent with our previous macroscopic approach, in which temperature was defined in terms of an ideal gas thermometer.

The number of possible locations for each atom is $V/\Delta x^3$, where Δx is the size of the space cells in phase space. The number of possible combinations of locations for the atoms is therefore $(V/\Delta x^3)^n$.

The possible momenta cover the surface of a $3n$-dimensional sphere, whose surface area is proportional to $E^{(3n-1)/2}$. In terms of phase-space cells, this area corresponds to $E^{(3n-1)/2}/\Delta p^{3n-1}$ possible combinations of momenta, multiplied by some constant of proportionality which depends on m, the atomic mass, and n, the number of atoms. To avoid having to calculate this constant of proportionality, we limit ourselves to calculating the part of the entropy that does not depend on n, so the resulting formula will not be useful for comparing entropies of ideal gas samples with different numbers of atoms.

The final result for the number of available states is

$$M = \left(\frac{V}{\Delta x^3}\right)^n \frac{E^{(3n-1)/2}}{\Delta p^{3n-1}} \qquad , \qquad \text{[function of } n\text{]}$$

so the entropy is

$$S = nk\ln V + \frac{3}{2}nk\ln E + (\text{function of } \Delta x,\ \Delta p, \text{ and } n) \qquad ,$$

where the distinction between n and $n-1$ has been ignored. Using $PV = nkT$ and $E = (3/2)nkT$, we can also rewrite this as

$$S = \frac{5}{2}nk\ln T - nk\ln P + \ldots, \qquad \text{[entropy of a monoatomic ideal gas]}$$

where "..." indicates terms that may depend on Δx, Δp, and n, but that have no effect on comparisons of gas samples with the same number of atoms.

Self-Check

Why does it make sense that the temperature term has a positive sign in the above example, while the pressure term is negative? Why does it make sense that the whole thing is proportional to n? ▷ Answer, p. 707

To show consistency with the macroscopic approach to thermodynamics, we need to show that these results are consistent with the behavior of an ideal-gas thermometer. Using the new definition $1/T = \mathrm{d}S/\mathrm{d}Q$, we have $1/T = \mathrm{d}S/\mathrm{d}E$, since transferring an amount of heat $\mathrm{d}Q$ into the gas increases its energy by a corresponding amount. Evaluating the derivative, we find $1/T = (3/2)nk/E$, or $E = (3/2)nkT$, which is the correct relation for a monoatomic ideal gas.

5.4.4 The arrow of time, or "This way to the Big Bang"

What about the second law of thermodynamics? The second law defines a forward direction to time, "time's arrow." The microscopic treatment of entropy, however, seems to have mysteriously sidestepped that whole issue. A graph like figure b on page 246, showing a fluctuation away from equilibrium, would look just as natural if we flipped it over to reverse the direction of time. After all, the basic laws of physics are conservation laws, which don't distinguish between past and future. Our present picture of entropy suggests that we restate the second law of thermodynamics as follows: low-entropy states are short-lived. An ice cube can't exist forever in warm water. We no longer have to distinguish past from future.

But how do we reconcile this with our strong psychological sense of the direction of time, including our ability to remember the past but not the future? Why do we observe ice cubes melting in water, but not the time-reversed version of the same process?

The answer is that there is no past-future asymmetry in the laws of physics, but there is a past-future asymmetry in the universe. The universe started out with the Big Bang. (Some of the evidence for the Big Bang theory is given on page 281.) The early universe had a very low entropy, and low-entropy states are short-lived. What does "short-lived" mean here, however? Hot coffee left in a paper cup will equilibrate with the air within ten minutes or so. Hot coffee in a thermos bottle maintains its low-entropy state for much longer, because the coffee is insulated by a vacuum between the inner and outer walls of the thermos. The universe has been mostly vacuum for a long time, so it's well insulated. Also, it takes billions of years for a low-entropy normal star like our sun to evolve into the high-entropy cinder known as a white dwarf.

The universe, then, is still in the process of equilibrating, and all the ways we have of telling the past from the future are really just ways of determining which direction in time points toward the Big Bang, i.e. which direction points to lower entropy. The psychological arrow of time, for instance, is ultimately based on the thermodynamic arrow. In some general sense, your brain is like a computer, and computation has thermodynamic effects. In even the most efficient possible computer, for example, erasing one bit of memory decreases its entropy from $k \ln 2$ (two possible states) to $k \ln 1$ (one state), for a drop of about 10^{-23} J/K. One way of determining the direction of the psychological arrow of time is that forward in psychological time is the direction in which, billions of years from now, all consciousness will have ceased; if consciousness was to exist forever in the universe, then there would have to be a never-ending decrease in the universe's entropy. This can't happen, because low-entropy states are short-lived.

Relating the direction of the thermodynamic arrow of time to the existence of the Big Bang is a satisfying way to avoid the paradox of how the second law can come from basic laws of physics that don't distinguish past from future. There is a remaining mystery, however: why did our universe have a Big Bang that was low in entropy? It could just as easily have been a maximum-entropy state, and in fact the number of possible high-entropy Big Bangs is vastly greater than the number of possible low-entropy ones. The question, however, is probably not one that can be answered using the methods of science. All we can say is that if the universe had started with a maximum-entropy Big Bang, then we wouldn't be here to wonder about it.

5.4.5 Quantum mechanics and zero entropy

The previous discussion would seem to imply that absolute entropies are never well defined, since any calculation of entropy will always end up having terms that depend on Δp and Δx. For instance, we might think that cooling an ideal gas to absolute zero would give zero entropy, since there is then only one available momentum state, but there would still be many possible position states. We'll see later in this book, however, that the quantum mechanical uncertainty principle makes it impossible to know the location and position of a particle simultaneously with perfect accuracy. The best we can do is to determine them with an accuracy such that the product $\Delta p \Delta x$ is equal to a constant called Planck's constant. According to quantum physics, then, there is a natural minimum size for rectangles in phase space, and entropy can be defined in absolute terms. Another way of looking at it is that according to quantum physics, the gas as a whole has some well-defined ground state, which is its state of minimum energy. When the gas is cooled to absolute zero, the scene is not at all like what we would picture in classical physics, with a lot of atoms lying around motionless. It might, for instance, be a strange quantum-mechanical state called the Bose-Einstein condensate, which was achieved for the first time recently with macroscopic amounts of atoms. Classically, the gas has many possible states available to it at zero temperature, since the positions of the atoms can be chosen in a variety of ways. The classical picture is a bad approximation under these circumstances, however. Quantum mechanically there is only one ground state, in which each atom is spread out over the available volume in a cloud of probability. The entropy is therefore zero at zero temperature. This fact, which cannot be understood in terms of classical physics, is known as the third law of thermodynamics.

5.4.6 Summary of the laws of thermodynamics

Here is a summary of the laws of thermodynamics:

The zeroth law of thermodynamics (page 233) If object A is at the same temperature as object B, and B is at the same temperature as C, then A is at the same temperature as C.

The first law of thermodynamics (page 228) Energy is conserved.

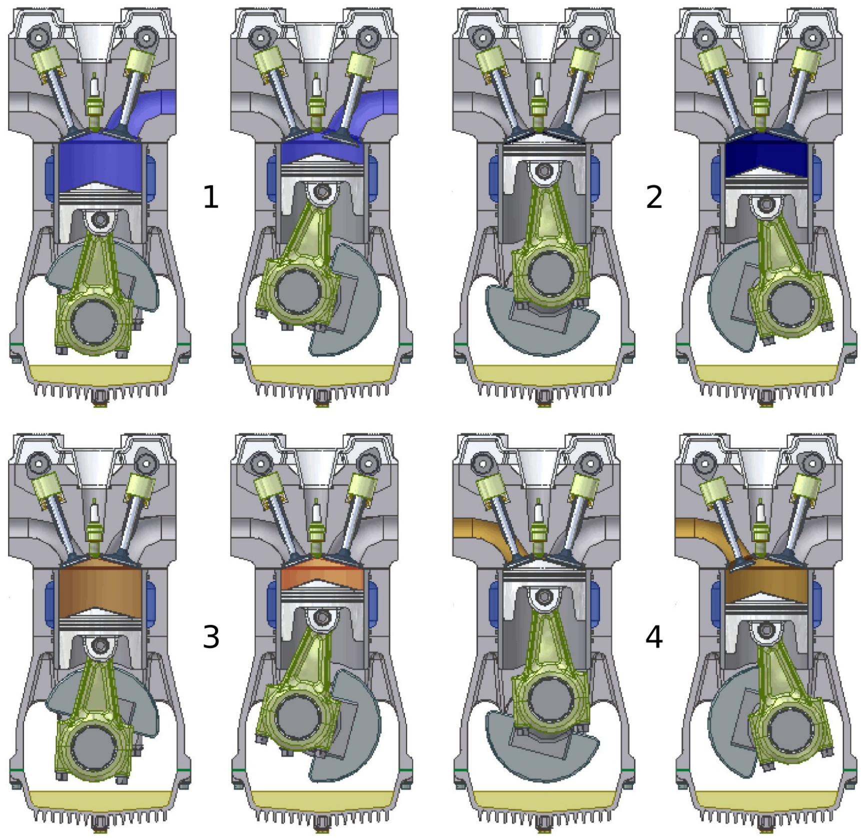
The second law of thermodynamics (page 243) The entropy of a closed system always increases, or at best stays the same: $\Delta S \geq 0$.

The third law of thermodynamics (page 253) The entropy of a system approaches zero as its temperature approaches absolute zero.

From a modern point of view, only the first law deserves to be called a fundamental law of physics. Once Boltmann discovered the microscopic nature of entropy, the zeroth and second laws could be understood as statements about probability: a system containing a large number of particles is overwhelmingly likely to do a certain thing, simply because the number of possible ways to do it is extremely large compared to the other possibilities. The third law is also now understood to be a consequence of more basic physical principles, but to explain the third law, it's not sufficient simply to know that matter is made of atoms: we also need to understand the quantum-mechanical nature of those atoms, discussed in chapter 12. Historically, however, the laws of thermodynamics were discovered in the eighteenth century, when the atomic theory of matter was generally considered to be a hypothesis that couldn't be tested experimentally. Ideally, with the publication of Boltzmann's work on entropy in 1877, the zeroth and second laws would have been immediately demoted from the status of physical laws, and likewise the development of quantum mechanics in the 1920's would have done the same for the third law.

5.5 More about Heat Engines (Optional)

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 j shows one full cycle

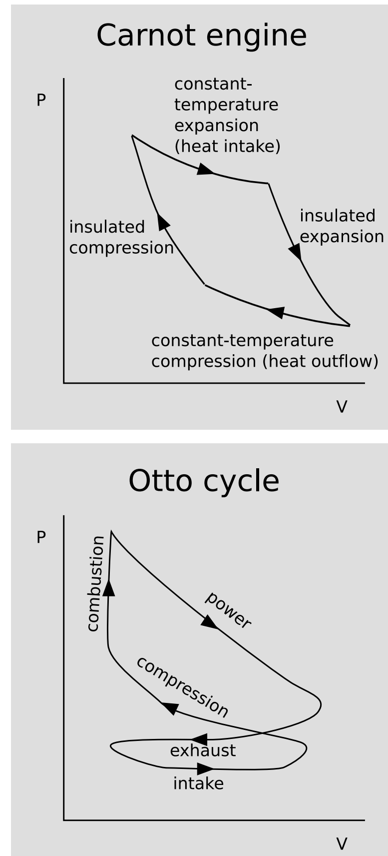


j / The Otto cycle. 1. In the exhaust stroke, the piston expels the burned air-gas mixture left over from the preceding cycle. 2. In the intake stroke, the piston sucks in fresh air-gas mixture. 3. In the compression stroke, the piston compresses the mixture, and heats it. 4. At the beginning of the power stroke, the spark plug fires, causing the air-gas mixture to burn explosively and heat up much more. The heated mixture expands, and does a large amount of positive mechanical work on the piston.

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 combustion: 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).



a / P-V diagrams for a Carnot engine and an Otto engine.

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 $\mathrm{d}W = F\mathrm{d}x$ for mechanical work can be rewritten as $\mathrm{d}W = P\mathrm{d}V$ 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, $\mathrm{d}x$ is positive, and the gas does positive work. When the gas is being compressed, $\mathrm{d}x$ is negative, and the gas does negative work, i.e., it absorbs energy. 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. The reason the gas cools is that it is doing work on the piston, and since it's insulated, it can't replenish that energy from the hot reservoir anymore. By conservation of energy, the energy it's giving away via the piston must be matched by a corresponding reduction in the random kinetic energies of its own molecules. This process *is* affected by whether the gas is monoatomic, diatomic, or polyatomic. The atoms in a monoatomic gas can only have kinetic energy due to their motion in space,

$$\begin{aligned} KE &= \frac{1}{2}m|\mathbf{v}|^2 \\ &= \frac{1}{2}mv_x^2 + \frac{1}{2}mv_y^2 + \frac{1}{2}mv_z^2 \qquad , \end{aligned}$$

along the three coordinate axes. We say that is has three *degrees of freedom.* In a diatomic gas, however, it is possible for the molecule to rotate end over end. This represents a two more degrees of freedom, since there are two axes about which it can rotate. When a diatomic gas expands, it acts like a person who is starving to death, but discovers hidden reserves of strength. The energy taken from it can come from the three types of kinetic energy due to motion through space, but it can also be supplied by reducing its rotational energy. It can be shown that, at equilibrium, every degree of freedom has an equal share of energy. (On the average, it's $kT/2$ per degree of freedom for each molecule.) Therefore, a diatomic gas undergoing insulated expansion only supplies $3/5$ of the energy from its motion through space, the other $2/5$ coming from its rotation. This means that the reduction in the speed of the motion through space is not as severe, and therefore the P-V curve is gentler. A polyatomic gas has not just two but three axes about which it can rotate, so its P-V curve is more gentle still. In general, we have $P \propto V^{-\gamma}$ for the insulated expansion of an ideal gas, where $\gamma = 5/3$, $7/5$, or $4/3$ in the case of a monoatomic, diatomic, or polyatomic gas, respectively. (The monoatomic case is analyzed on page 696.) 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.

The compression ratio *example 20*

Operating along a constant-temperature stroke, the amount of mechanical work done by a heat engine can be calculated as follows:

$$PV = nkT$$

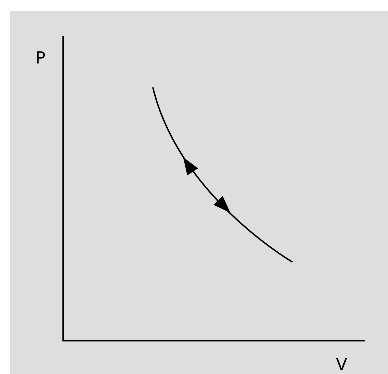
Setting $c = nkT$ to simplify the writing,

$$\begin{aligned}
P &= cV^{-1} \\
W &= \int_{V_i}^{V_f} P \, \mathrm{d}V \\
&= c \int_{V_i}^{V_f} V^{-1} \, \mathrm{d}V \\
&= c \ln V_f - c \ln V_i \\
&= c \ln(V_f / V_i)
\end{aligned}$$

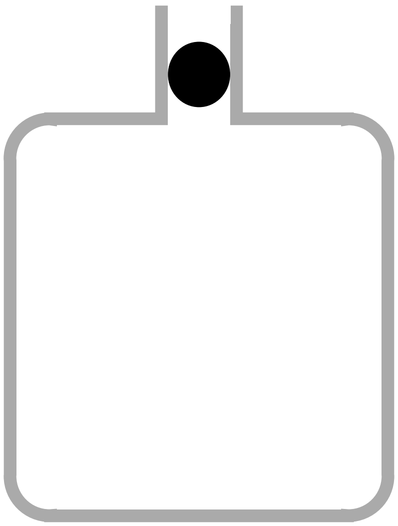
The ratio V_f / V_i is called the *compression ratio* of the engine, and higher values result in more power along this stroke. Along an insulated stroke, we have $P \propto V^{-\gamma}$, with $\gamma \neq 1$, so the result for the work no longer has this perfect mathematical property of depending only 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.

Sound *example 21*

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.



b / Example 21,



c / Example 22.

Measuring γ using the "spring of air" *example 22*

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.

$$\begin{aligned}
\omega^2 &= \frac{k}{m} \\
&= -\frac{1}{m}\frac{dF}{dy}\bigg|_{y_o} \\
&= -\frac{A}{m}\frac{dP}{dy}\bigg|_{y_o} \\
&= -\frac{A^2}{m}\frac{dP}{dV}\bigg|_{V_o}
\end{aligned}$$

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,

$$\omega^2 = \gamma \frac{A^2}{m}\frac{P_o}{V_o}$$

What is remarkable about this experiment is that although the technology needed to construct it has been been available for centuries, it allows us to find out the shape of the molecules in a gas! That is, with the proper interpretation, it answers a question that, even as late as Boltzmann's lifetime, would have been looked down on as being unscientific, in the same class as debates about how many angels could dance on the head of a pin.

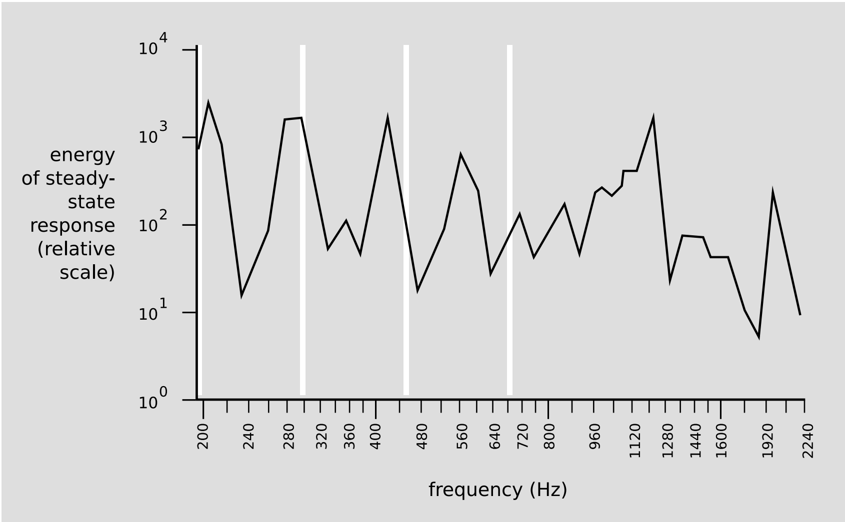
The Helmholtz resonator *example 23*

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 22, 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

$$\omega^2 = \gamma \frac{P_o}{\rho}\frac{A}{LV_o} \qquad .$$

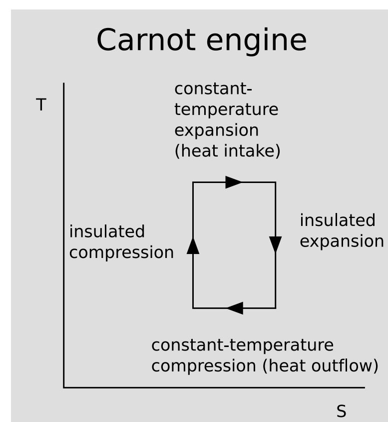
This is known as a Helmholtz resonator. As shown in figure d, a violin or an acoustic guitar has a Helmholtz resonance, since air can move in and out through the f-holes. Problem 10 is a more quantitative exploration of this.

We have already seen, based on the microscopic nature of entropy, that any Carnot engine has the same efficiency, and the ar-



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.

gument 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 20? 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. An infinitesimal transfer of heat $\mathrm{d}Q$ gives rise to a change in entropy $\mathrm{d}S = \mathrm{d}Q/T$, 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.



e / A T-S diagram for a Carnot engine.

This chapter is summarized on page 728. Notation and terminology are tabulated on pages 718-719.

Problems

The symbols $\surd$, , etc. are explained on page 262.

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 ($0\,^\circ\mathrm{C}$ and 101 kPa). $\surd$

2 A gas in a cylinder expands its volume by an amount $\mathrm{d}V$, pushing out a piston. Show that the work done by the gas on the piston is given by $\mathrm{d}W = P\mathrm{d}V$.

3 (a) A helium atom contains 2 protons, 2 electrons, and 2 neutrons. Find the mass of a helium atom. $\surd$
(b) Find the number of atoms in 1.0 kg of helium. $\surd$
(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.) $\surd$

4 Refrigerators, air conditioners, and heat pumps are heat engines that work in reverse. You put in mechanical work, and the effect is to take heat out of a cooler reservoir and deposit heat in a warmer one: $Q_L + W = Q_H$. As with the heat engines discussed previously, the efficiency is defined as the energy transfer you want (Q_L for a refrigerator or air conditioner, Q_H for a heat pump) divided by the energy transfer you pay for (W).

Efficiencies are supposed to be unitless, but the efficiency of an air conditioner is normally given in terms of an EER rating (or a more complex version called an SEER). The EER is defined as Q_L/W, but expressed in the barbaric units of 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 $80\,^\circ F$ ($27\,^\circ\mathrm{C}$) inside and $95\,^\circ F$ ($35\,^\circ\mathrm{C}$) outside.

(a) What would be the EER rating of a reversed Carnot engine used as an air conditioner? $\surd$
(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? $\surd$

5 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 18 on page 67, 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. $\surd$

(b) Show that your answer from part a has the right units for pressure.
(c) Evaluate the result numerically. √
(d) Given that the earth's atmosphere is on the order of one thousandth the earth's radius, and that the density of the earth is several thousand times greater than the density of the lower atmosphere, check that your result is of a reasonable order of magnitude.

6 (a) Determine the ratio between the escape velocities from the surfaces of the earth and the moon. √
(b) The temperature during the lunar daytime gets up to about 130 °C. In the extremely thin (almost nonexistent) lunar atmosphere, estimate how the typical velocity of a molecule would compare with that of the same type of molecule in the earth's atmosphere. Assume that the earth's atmosphere has a temperature of 0 °C. √
(c) Suppose you were to go to the moon and release some fluorocarbon gas, with molecular formula C_nF_{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. √

7 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). √

8 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?

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 °C and an ambient temperature of 22 °C. (c) Interpret your answer. ▷ Answer, p. 712

10 Example 23 on page 258 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. ▷ Answer, p. 712

Key to symbols:
easy   typical   challenging   difficult   very difficult
√ An answer check is available at www.lightandmatter.com.

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

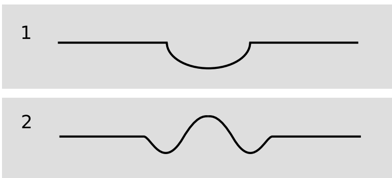
Chapter 6
Waves

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.

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]

6.1 Free Waves

6.1.1 Wave motion



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

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

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

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

1. Superposition

If you watched the water in the cup carefully, you noticed the ghostlike behavior of the reflected ripples coming back toward the center of the cup and the outgoing ripples that hadn't yet been reflected: they passed right through each other. This is the first, and the most profound, difference between wave motion and the mo-

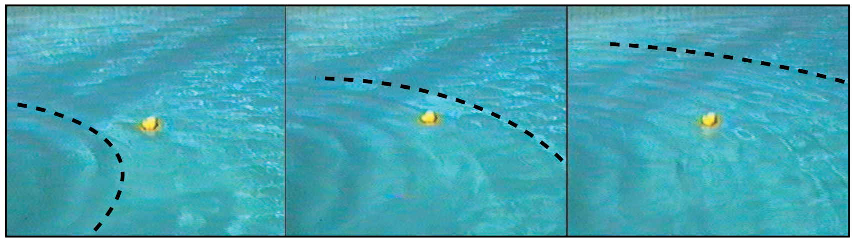
[1]Speaking more carefully, I should say that every basic building block of light and matter has both wave and particle properties.

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

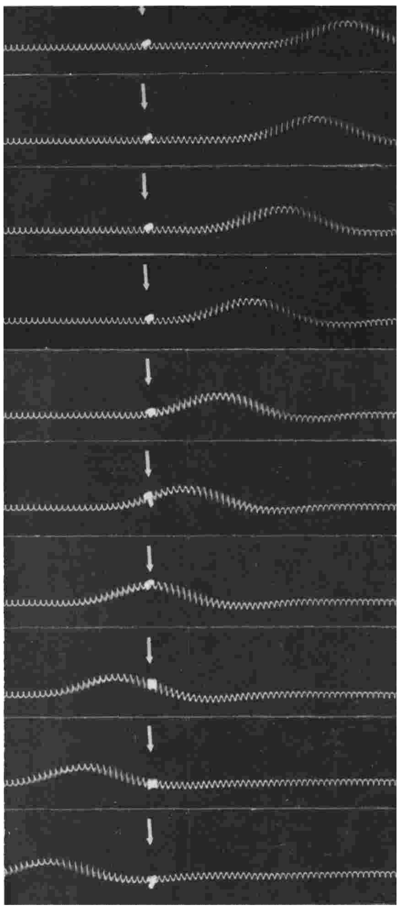
tion of objects: waves do not display any repulsion of each other analogous to the normal forces between objects that come in contact. Two wave patterns can therefore overlap in the same region of space, as shown in figure b. Where the two waves coincide, they add together. For instance, suppose that at a certain location in at a certain moment in time, each wave would have had a crest 3 cm above the normal water level. The waves combine at this point to make a 6-cm crest. We use negative numbers to represent depressions in the water. If both waves would have had a troughs measuring -3 cm, then they combine to make an extra-deep -6 cm trough. A $+3$ cm crest and a -3 cm trough result in a height of zero, i.e. the waves momentarily cancel each other out at that point. This additive rule is referred to as the principle of superposition, "superposition" being merely a fancy word for "adding."

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

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



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



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

2. The medium is not transported with the wave.

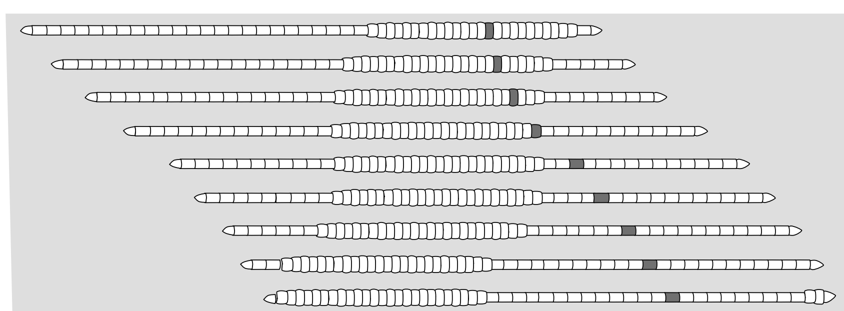
The sequence of three photos in figure c shows a series of water waves before it has reached a rubber duck (left), having just passed the duck (middle) and having progressed about a meter beyond the duck (right). The duck bobs around its initial position, but is not carried along with the wave. This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

Self-Check

In figure d, you can detect the side-to-side motion of the spring because the spring appears blurry. At a certain instant, represented by a single photo, how would you describe the motion of the different parts of the spring? Other than the flat parts, do any parts of the spring have zero velocity? ▷ Answer, p. 707

A worm *example 1*

The worm in the figure is moving to the right. The wave pattern, a pulse consisting of a compressed area of its body, moves to the left. In other words, the motion of the wave pattern is in the opposite direction compared to the motion of the medium.



Surfing *example 2*

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

e / Example 2. The surfer is dragging his hand in the water.

3. A wave's velocity depends on the medium.

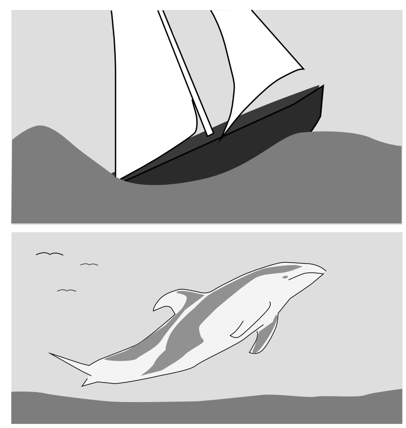
A material object can move with any velocity, and can be sped up or slowed down by a force that increases or decreases its kinetic energy. Not so with waves. The speed of a wave, depends on the properties of the medium (and perhaps also on the shape of the wave, for certain types of waves). Sound waves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

f / Example 3: a breaking wave.

Breaking waves *example 3*

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

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



g / Example 4. The boat has run up against a limit on its speed because it can't climb over its own wave. Dolphins get around the problem by leaping out of the water.

Hull speed *example 4*

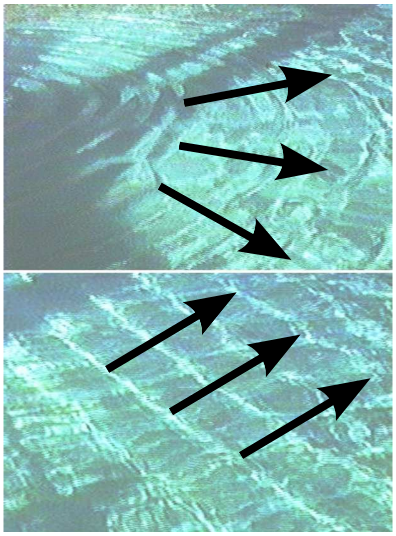
The speeds of most boats, and of some surface-swimming animals, are limited by the fact that they make a wave due to their motion through the water. The boat in figure g is going at the same speed as its own waves, and can't go any faster. No matter how hard the boat pushes against the water, it can't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

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

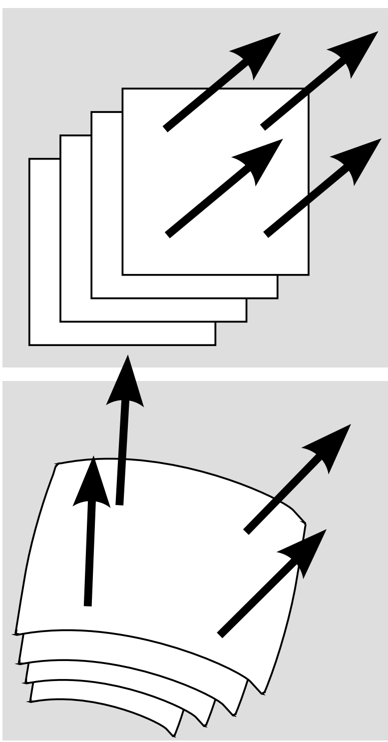
Wave patterns

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

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



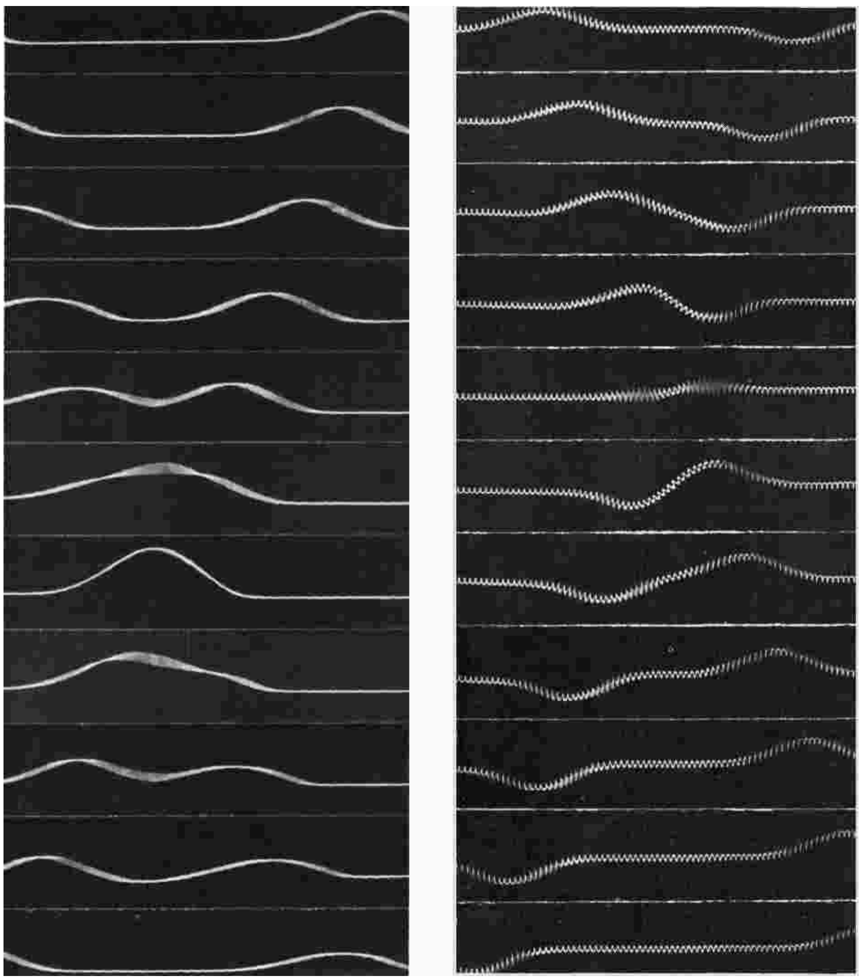
h / Circular and linear wave patterns.



i / Plane and spherical wave patterns.

Discussion Questions

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



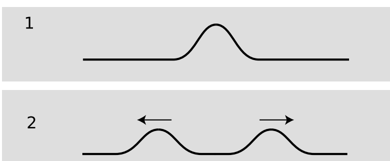
j / Discussion question A.

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

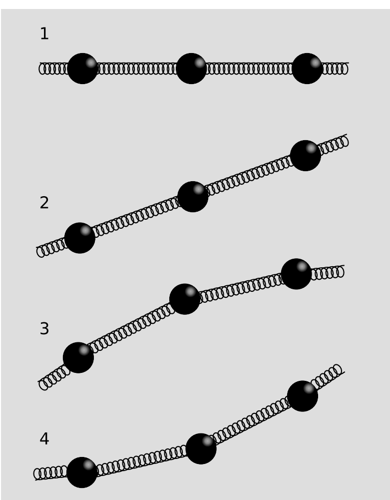
C A traveling wave pulse is moving to the right on a string. Sketch the velocity vectors of the various parts of the string. Now do the same for a pulse moving to the left.

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

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



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



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

6.1.2 Waves on a string

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

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

Intuitive ideas

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

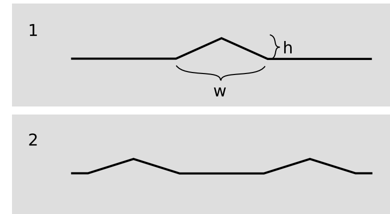
We can gain insight by modeling the string as a series of masses connected by springs, m. (In the actual string the mass and the springiness are both contributed by the molecules themselves.) If we look at various microscopic portions of the string, there will be some areas that are flat, 1, some that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

Approximate treatment

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

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

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



n / A triangular pulse spreads out.

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

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

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

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

Our final result for the speed of the pulses is

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

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

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

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

Treatment using calculus

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

$$\mathrm{d}F = T\,\frac{\partial^2 y}{\partial x^2}\mathrm{d}x \quad .$$

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

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

The second derivative with respect to time is related to the second derivative with respect to position. This is no more than a fancy mathematical statement of the intuitive fact developed above, that the string accelerates so as to flatten out its curves.

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

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

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

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

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

6.1.3 Sound and light waves

Sound waves

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

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

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

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

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

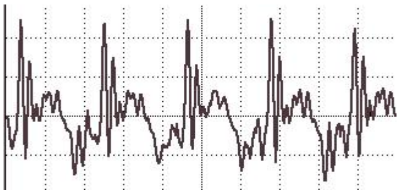
If sound has all the properties we expect from a wave, then what type of wave is it? It is a series of compressions and expansions of the air. Even for a very loud sound, the increase or decrease compared to normal atmospheric pressure is no more than a part per million, so our ears are apparently very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

Light waves

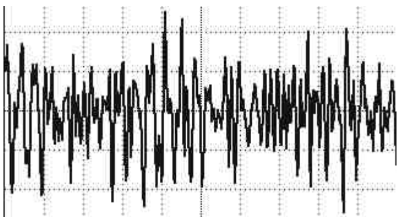
Entirely similar observations lead us to believe that light is a wave, although the concept of light as a wave had a long and tortuous history. It is interesting to note that Isaac Newton very influentially advocated a contrary idea about light. The belief that matter was made of atoms was stylish at the time among radical thinkers (although there was no experimental evidence for their existence), and it seemed logical to Newton that light as well should be made of tiny particles, which he called corpuscles (Latin for "small objects"). Newton's triumphs in the science of mechanics, i.e. the study of matter, brought him such great prestige that nobody bothered to question his incorrect theory of light for 150 years. One persuasive proof that light is a wave is that according to Newton's theory, two intersecting beams of light should experience at least some disruption because of collisions between their corpuscles. Even if the corpuscles were extremely small, and collisions therefore very infrequent, at least some dimming should have been measurable. In fact, very delicate experiments have shown that there is no dimming.

The wave theory of light was entirely successful up until the 20th century, when it was discovered that not all the phenomena of light could be explained with a pure wave theory. It is now believed that both light and matter are made out of tiny chunks which have both wave and particle properties. For now, we will content ourselves with the wave theory of light, which is capable of explaining a great many things, from cameras to rainbows.

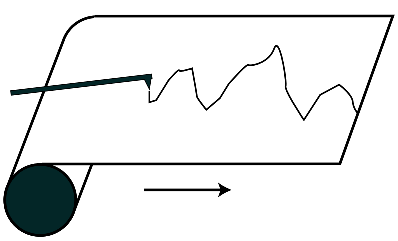
If light is a wave, what is waving? What is the medium that wiggles when a light wave goes by? It isn't air. A vacuum is impenetrable to sound, but light from the stars travels happily through zillions of miles of empty space. Light bulbs have no air inside them, but that doesn't prevent the light waves from leaving the filament. For a long time, physicists assumed that there must be a mysterious medium for light waves, and they called it the ether (not to be confused with the chemical). Supposedly the ether existed everywhere in space, and was immune to vacuum pumps. The details of the story are more fittingly reserved for later in this course, but the end result was that a long series of experiments failed to detect any evidence for the ether, and it is no longer believed to exist. Instead, light can be explained as a wave pattern made up of electrical and magnetic fields.



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



p / A similar graph for a nonperiodic wave, “sh.”



q / A strip chart recorder.

6.1.4 Periodic waves

Period and frequency of a periodic wave

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

$$T = \text{period} = \text{seconds per cycle}$$
$$f = \text{frequency} = 1/T = \text{cycles per second}$$
$$\omega = \text{angular frequency} = 2\pi f = \text{radians per second}$$

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

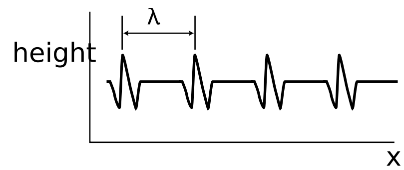
The period of a sound wave correlates with our sensory impression of musical pitch. A high frequency (short period) is a high note. The sounds that really define the musical notes of a song are only the ones that are periodic. It is not possible to sing a nonperiodic sound like “sh” with a definite pitch.

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

Graphs of waves as a function of position

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

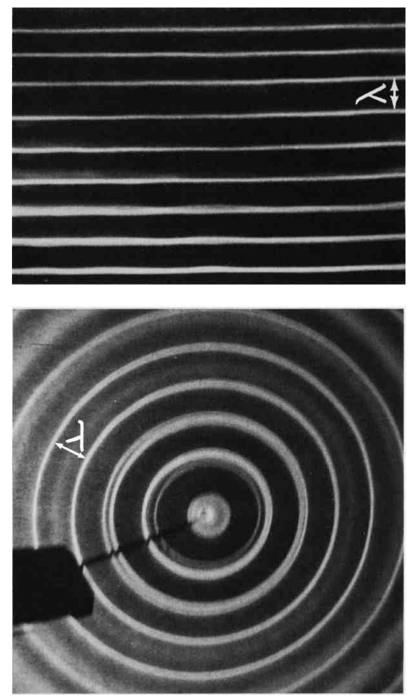
An easy way to visualize this is in terms of a strip chart recorder, an obsolescing device consisting of a pen that wiggles back and forth as a roll of paper is fed under it. It can be used to record a person’s electrocardiogram, or seismic waves too small to be felt as a noticeable earthquake but detectable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



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

Wavelength

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



s / Wavelengths of linear and circular waves. (*PSSC Physics*)

Wave velocity related to frequency and wavelength

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

$$v = \lambda/T \qquad .$$

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

$$v = f\lambda \qquad .$$

Wavelength of radio waves *example 5*

▷ The speed of light is 3.0×10^8 m/s. What is the wavelength of the radio waves emitted by KLON, a station whose frequency is 88.1 MHz?

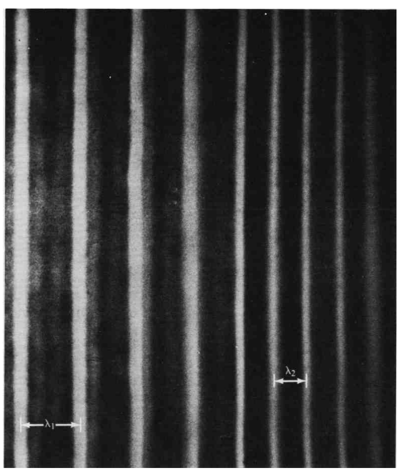
▷ Solving for wavelength, we have

$$\begin{aligned} \lambda &= v/f \\ &= (3.0 \times 10^8\ \text{m/s})/(88.1 \times 10^6\ \text{s}^{-1}) \\ &= 3.4\ \text{m} \end{aligned}$$

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

The equation $v = f\lambda$ defines a fixed relationship between any two of the variables if the other is held fixed. The speed of radio waves

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



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

in air is almost exactly the same for all wavelengths and frequencies (it is exactly the same if they are in a vacuum), so there is a fixed relationship between their frequency and wavelength. Thus we can say either "Are we on the same wavelength?" or "Are we on the same frequency?"

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

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

A note on dispersive waves

The discussion of wave velocity given here is actually a little bit of an oversimplification for a wave whose velocity depends on its frequency and wavelength. Such a wave is called a dispersive wave. Nearly all the waves we deal with in this course are nondispersive, but the issue becomes important in chapter 12, where it is discussed in detail.

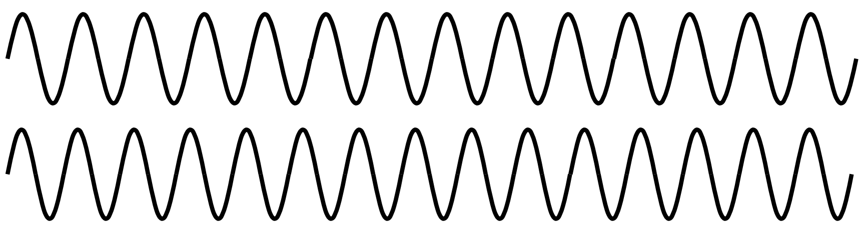
Sinusoidal waves

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

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

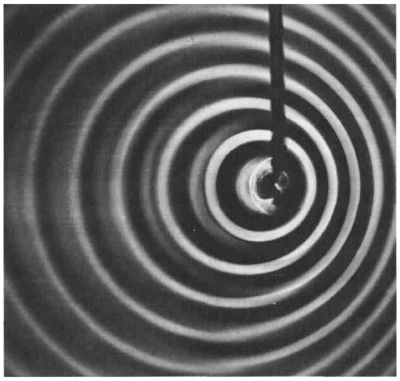
Discussion Questions

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



v / Discussion question A.

6.1.5 The Doppler effect



w / 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.

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

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

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

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

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

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

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

Doppler-shifted sound from a race car *example 6*

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

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

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

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

$$1/0.85 = 1.18,$$

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

Doppler shift of the light emitted by a race car *example 7*

▷ What is the percent shift in the wavelength of the light waves emitted by a race car's headlights?

▷ Looking up the speed of light in the back of the book, $v = 3.0\times10^8$ m/s, we find

$$1-\frac{u}{v}=0.99999983 \qquad ,$$

i.e. the percentage shift is only 0.000017%.

The second example shows that under ordinary earthbound circumstances, Doppler shifts of light are negligible because ordinary things go so much slower than the speed of light. It's a different story, however, when it comes to stars and galaxies, and this leads us to a story that has profound implications for our understanding of the origin of the universe.

The Big Bang

As soon as astronomers began looking at the sky through telescopes, they began noticing certain objects that looked like clouds in deep space. The fact that they looked the same night after night meant that they were beyond the earth's atmosphere. Not knowing what they really were, but wanting to sound official, they called them "nebulae," a Latin word meaning "clouds" but sounding more impressive. In the early 20th century, astronomers realized that although some really were clouds of gas (e.g. the middle "star" of Orion's sword, which is visibly fuzzy even to the naked eye when conditions are good), others were what we now call galaxies: virtual island universes consisting of trillions of stars (for example the Andromeda Galaxy, which is visible as a fuzzy patch through binoculars). Three hundred years after Galileo had resolved the Milky Way into individual stars through his telescope, astronomers realized that the universe is made of galaxies of stars, and the Milky Way is simply the visible part of the flat disk of our own galaxy, seen from inside.

x / The galaxy M100 in the constellation Coma Berenices. Under higher magnification, the milky clouds reveal themselves to be composed of trillions of stars.

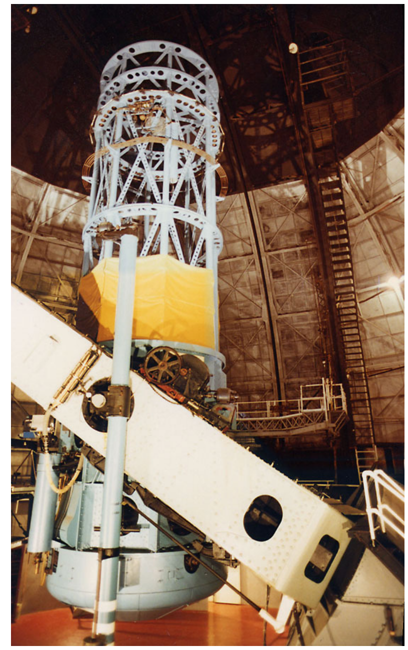
y / How do astronomers know what mixture of wavelengths a star emitted originally, so that they can tell how much the Doppler shift was? This image (obtained by the author with equipment costing about $5, and no telescope) shows the mixture of colors emitted by the star Sirius. (If you have the book in black and white, blue is on the left and red on the right.) The star appears white or bluish-white to the eye, but any light looks white if it contains roughly an equal mixture of the rainbow colors, i.e. of all the pure sinusoidal waves with wavelengths lying in the visible range. Note the black "gap teeth." These are the fingerprint of hydrogen in the outer atmosphere of Sirius. These wavelengths are selectively absorbed by hydrogen. Sirius is in our own galaxy, but similar stars in other galaxies would have the whole pattern shifted toward the red end, indicating they are moving away from us.

This opened up the scientific study of cosmology, the structure and history of the universe as a whole, a field that had not been seriously attacked since the days of Newton. Newton had realized that if gravity was always attractive, never repulsive, the universe would have a tendency to collapse. His solution to the problem was to posit a universe that was infinite and uniformly populated with matter, so that it would have no geometrical center. The gravitational forces in such a universe would always tend to cancel out by symmetry, so there would be no collapse. By the 20th century, the belief in an unchanging and infinite universe had become conventional wisdom in science, partly as a reaction against the time that had been wasted trying to find explanations of ancient geological phenomena based on catastrophes suggested by biblical events like Noah's flood.

In the 1920's astronomer Edwin Hubble began studying the Doppler shifts of the light emitted by galaxies. A former college football player with a serious nicotine addiction, Hubble did not set out to change our image of the beginning of the universe. His autobiography seldom even mentions the cosmological discovery for which he is now remembered. When astronomers began to study the Doppler shifts of galaxies, they expected that each galaxy's direction and velocity of motion would be essentially random. Some would be approaching us, and their light would therefore be Doppler-shifted to the blue end of the spectrum, while an equal number would be expected to have red shifts. What Hubble discovered instead was that except for a few very nearby ones, all the galaxies had red shifts, indicating that they were receding from us at a hefty fraction of the speed of light. Not only that, but the ones farther away were receding more quickly. The speeds were directly proportional to their distance from us.

Did this mean that the earth (or at least our galaxy) was the center of the universe? No, because Doppler shifts of light only depend on the relative motion of the source and the observer. If we see a distant galaxy moving away from us at 10% of the speed of light, we can be assured that the astronomers who live in that galaxy will see ours receding from them at the same speed in the opposite direction. The whole universe can be envisioned as a rising loaf of raisin bread. As the bread expands, there is more and more space between the raisins. The farther apart two raisins are, the greater the speed with which they move apart.

The universe's expansion is presumably decelerating because of gravitational attraction among the galaxies. We do not presently know whether there is enough mass in the universe to cause enough attraction to halt the expansion eventually. But perhaps more interesting than the distant future of the universe is what its present expansion implies about its past. Extrapolating backward in time using the known laws of physics, the universe must have been denser and denser at earlier and earlier times. At some point, it must have been extremely dense and hot, and we can even detect the radiation from this early fireball, in the form of microwave radiation that permeates space. The phrase Big Bang was originally coined by the doubters of the theory to make it sound ridiculous, but it stuck, and today essentially all astronomers accept the Big Bang theory based on the very direct evidence of the red shifts and the cosmic microwave background radiation.



z / The telescope at Mount Wilson used by Hubble.

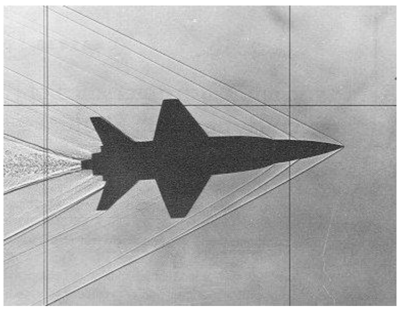
Finally it should be noted what the Big Bang theory is not. It is not an explanation of *why* the universe exists. Such questions belong to the realm of religion, not science. Science can find ever simpler and ever more fundamental explanations for a variety of phenomena, but ultimately science takes the universe as it is according to observations.

Furthermore, there is an unfortunate tendency, even among many scientists, to speak of the Big Bang theory was a description of the very first event in the universe, which caused everything after it. Although it is true that time may have had a beginning (Einstein's theory of general relativity admits such a possibility), the methods of science can only work within a certain range of conditions such as temperature and density. Beyond a temperature of about 10^9 K, the random thermal motion of subatomic particles becomes so rapid that its velocity is comparable to the speed of light. Early enough in the history of the universe, when these temperatures existed, Newtonian physics becomes less accurate, and we must describe nature using the more general description given by Einstein's theory of relativity, which encompasses Newtonian physics as a special case. At even higher temperatures, beyond about 10^{33} degrees, physicists know that Einstein's theory as well begins to fall apart, but we don't know how to construct the even more general theory of nature that would work at those temperatures. No matter how far physics progresses, we will never be able to describe nature at infinitely high temperatures, since there is a limit to the temperatures we can explore by experiment and observation in order to guide us to the right theory. We are confident that we understand the basic physics involved in the evolution of the universe starting a few minutes after the Big Bang, and we may be able to push back to milliseconds or microseconds after it, but we cannot use the methods of science to deal with the beginning of time itself.

A note on Doppler shifts of light

If Doppler shifts depend only on the relative motion of the source and receiver, then there is no way for a person moving with the source and another person moving with the receiver to determine who is moving and who isn't. Either can blame the Doppler shift entirely on the other's motion and claim to be at rest herself. This is entirely in agreement with the principle stated originally by Galileo that all motion is relative.

On the other hand, a careful analysis of the Doppler shifts of water or sound waves shows that it is only approximately true, at low speeds, that the shifts just depend on the relative motion of the source and observer. For instance, it is possible for a jet plane to keep up with its own sound waves, so that the sound waves appear to stand still to the pilot of the plane. The pilot then knows she is moving at exactly the speed of sound. The reason this doesn't disprove the relativity of motion is that the pilot is not really determining her absolute motion but rather her motion relative to the air, which is the medium of the sound waves.



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

Einstein realized that this solved the problem for sound or water waves, but would not salvage the principle of relative motion in the case of light waves, since light is not a vibration of any physical medium such as water or air. Beginning by imagining what a beam of light would look like to a person riding a motorcycle alongside it, Einstein eventually came up with a radical new way of describing the universe, in which space and time are distorted as measured by observers in different states of motion. As a consequence of this Theory of Relativity, he showed that light waves would have Doppler shifts that would exactly, not just approximately, depend only on the relative motion of the source and receiver.

ab / This fighter jet has just accelerated past the speed of sound. The sudden decompression of the air causes water droplets to condense, forming a cloud.

Discussion Questions

A If an airplane travels at exactly the speed of sound, what would be the wavelength of the forward-emitted part of the sound waves it emitted? How should this be interpreted, and what would actually happen? What happens if it's going faster than the speed of sound? Can you use this to explain what you see in figures aa and ab?

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

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

6.2 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.

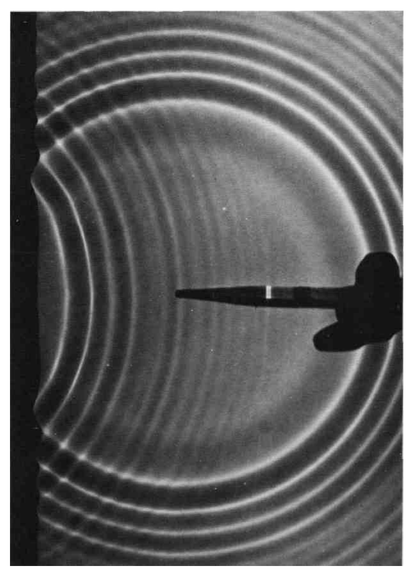
Since speech sounds are instinctive for us, we seldom think about them consciously. How do we do control sound waves so skillfully? Mostly we do it by changing the shape of a connected set of hollow cavities in our chest, throat, and head. Somehow by moving the boundaries of this space in and out, we 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.



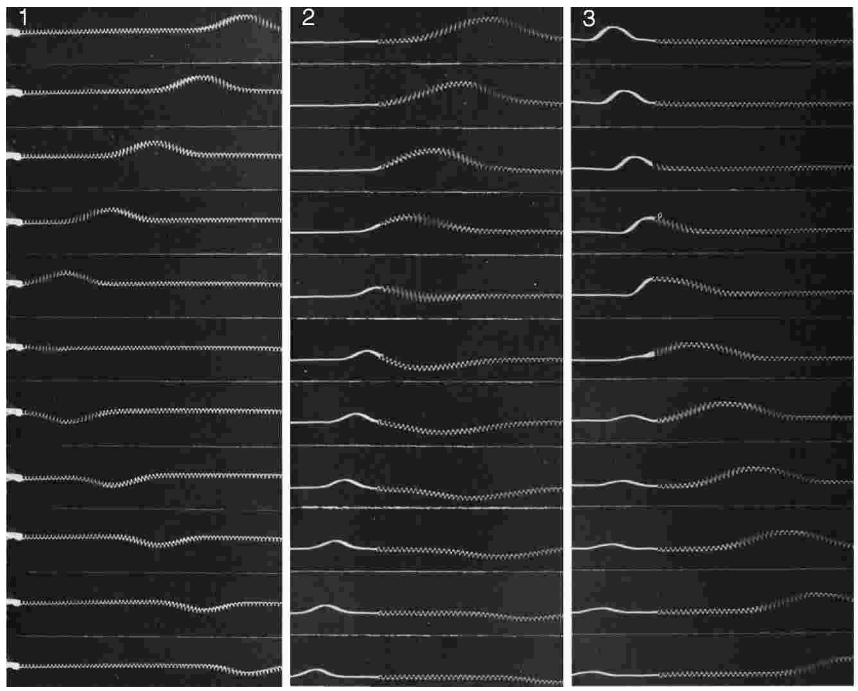
b / Circular water waves are reflected from a boundary on the left. (*PSSC Physics*)

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.

Self-Check
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, p. 707

Fish have internal ears. *example 8*
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.

Whale songs traveling long distances *example 9*

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.

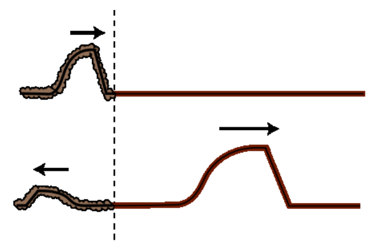
Long-distance radio communication *example 10*

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, p. 707

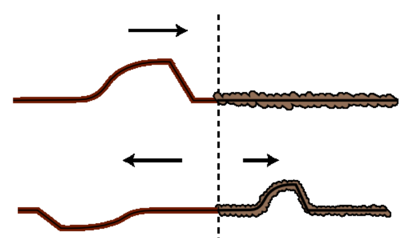
The use of the word "reflection" naturally brings to mind the creation of an image by a mirror, but this might be confusing, because we do not normally refer to "reflection" when we look at surfaces that are not shiny. Nevertheless, reflection is how we see the surfaces of all objects, not just polished 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.



d / An uninverted reflection. The reflected pulse is reversed front to back, but is not upside-down.

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.

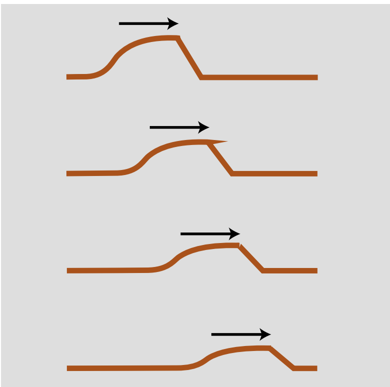


e / An inverted reflection. The reflected pulse is reversed both front to back and top to bottom.

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.

Self-Check
As a wave undergoes absorption, it loses energy. Does this mean that it slows down? ▷ Answer, p. 707

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).

Soundproofing *example 11*

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.

Radio transmission *example 12*

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

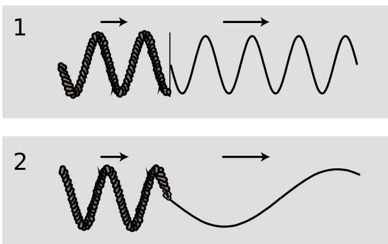
A A sound wave that underwent a pressure-inverting reflection would have its compressions converted to expansions and vice versa. How would its energy and frequency compare with those of the original sound? Would it sound any different? What happens if you swap the two wires where they connect to a stereo 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.

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

$$1+R=T \qquad .$$

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

$$1-R=\frac{v_1}{v_2}T \qquad .$$

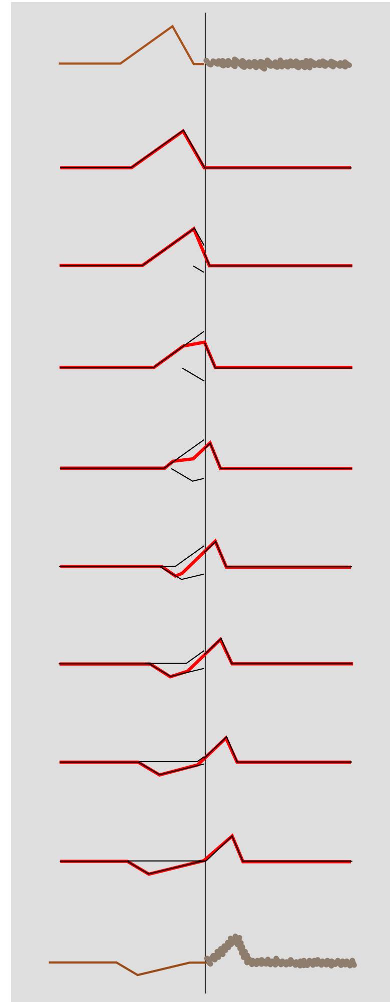
Solving the two equations for the unknowns R and T gives

$$R=\frac{v_2-v_1}{v_2+v_1}$$

and

$$T=\frac{2v_2}{v_2+v_1} \qquad .$$

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.

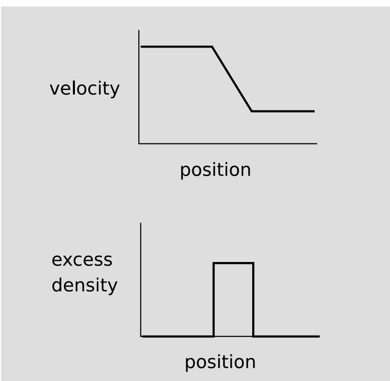


h / A pulse being partially reflected and partially transmitted at the boundary between two strings in which the wave speed is different. The top drawing shows the pulse heading to the right, toward the heaver string. For clarity, all but the first and last drawings are schematic. Once the reflected 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.

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.



i / A wave pattern in freeway traffic.

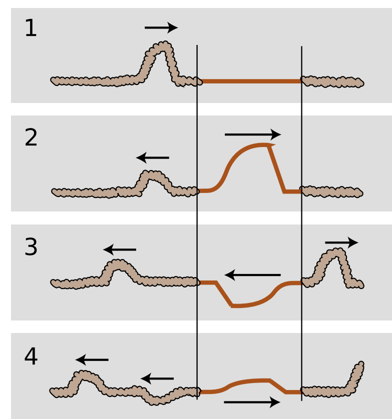
Let's start by considering wave disturbances of freeway traffic. Anyone who has driven frequently on crowded freeways has observed the phenomenon in which one driver taps the brakes, starting a chain reaction that travels backward down the freeway as each person in turn exercises caution in order to avoid rear-ending anyone. The reason why this type of wave is relevant is that it gives a simple, easily visualized example of 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.

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.

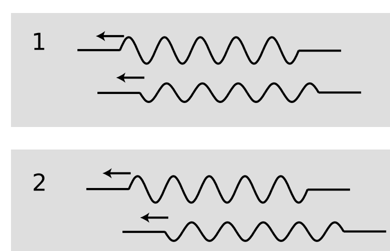
Now let's consider reflections. If we observe the freeway wave in a mirror, the high-density area will still appear high in density, but velocity in the opposite direction will now be described by a negative number. A person observing the mirror image will draw the same density graph, but the velocity 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.

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. In the other, the electric field is uninverted and the magnetic field inverted.



j / A pulse encounters two boundaries.



k / A sine wave has been reflected at two different boundaries, and the two reflections interfere.

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. Figure j/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, j/3. In figure j/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 j/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, k/1, or out of phase, k/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, k/2, with cancellation resulting in a very weak reflected wave.

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 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, p. 707

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.

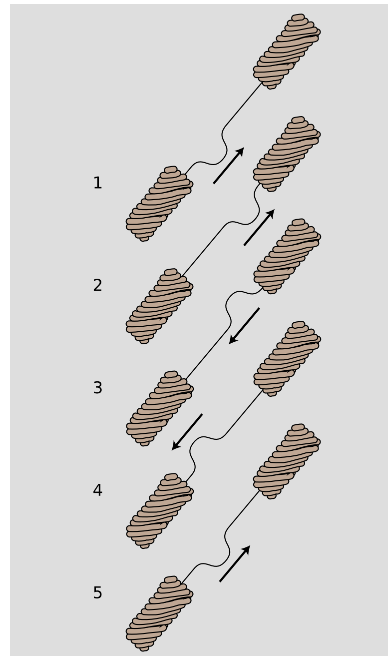
Discussion Questions

A Is it possible to get *complete* destructive interference in an optical coating, at least for light of one specific wavelength?

B 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?

C 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?

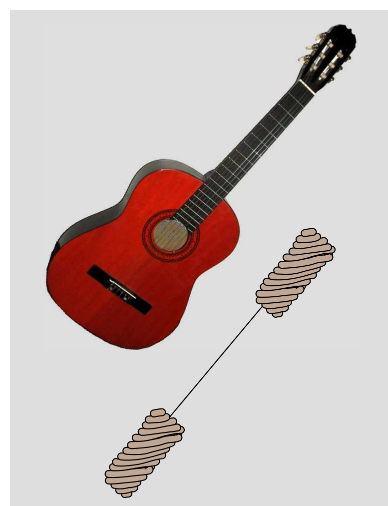
D An observation like the one described in the previous discussion question was used by Newton as evidence *against* the wave theory of light! If Newton didn't know about inverting and noninverting reflections, what would have seemed inexplicable to him about the region where the air layer had zero or nearly zero thickness?



l / A pulse bounces back and forth.

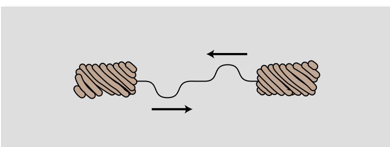
6.2.4 Waves bounded on both sides

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.



m / 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.

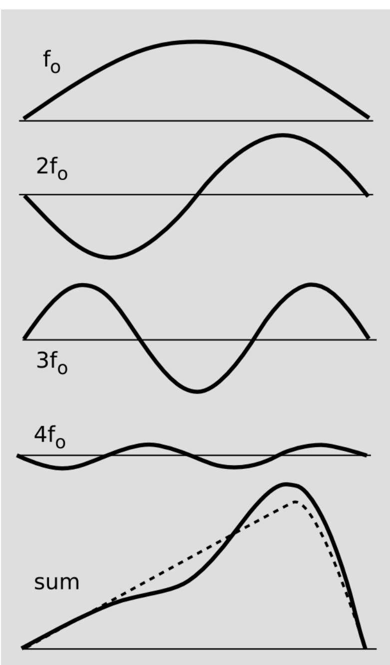


n / The period of this double-pulse pattern is half of what we'd otherwise expect.

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, l/1. After reflection from the top end, it is inverted, l/3. Now something interesting happens: figure l/5 is identical to figure l/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 l/1 to l/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, p. 708



o / Any wave can be made by superposing sine waves.

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.)

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 (g). 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.

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. The figure on the left 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.

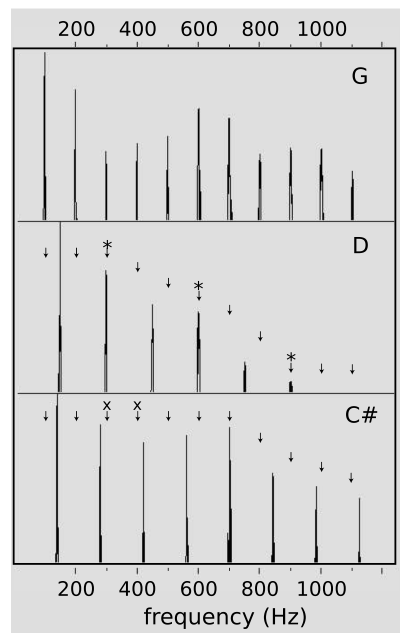
We thus arrive at the following very general conclusion. Whenever a wave pattern exists in a medium bounded on both sides by media in which the wave speed is very different, the motion can be broken down into the motion of a (theoretically infinite) series of sine waves, with frequencies f_o, $2f_o$, $3f_o$, ... 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

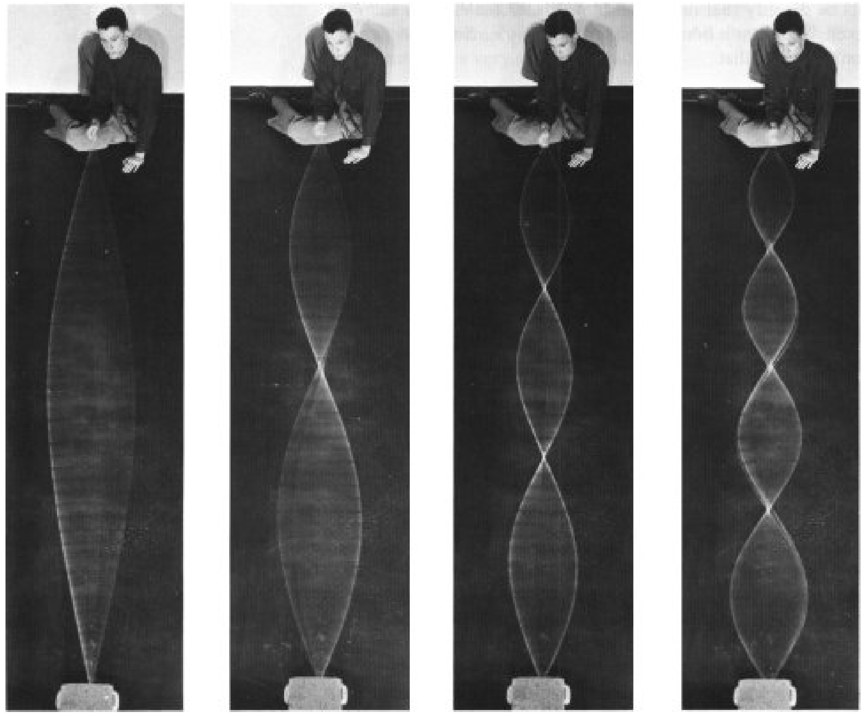
Many musicians claim to be able to pick out by ear several of the frequencies $2f_o$, $3f_o$, ..., called overtones or *harmonics* of the fundamental f_o, but they are kidding themselves. In reality, the overtone series has two important roles in music, neither of which depends on this fictitious ability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

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

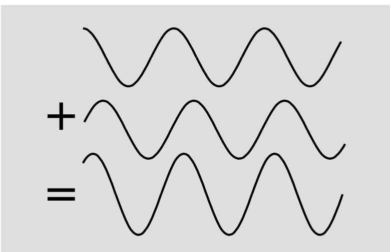
Second, although the ear cannot separate the individual harmonics of a single musical tone, it is very sensitive to clashes between the overtones of notes played simultaneously, i.e. in harmony. We tend to perceive a combination of notes as being dissonant if they have overtones that are close but not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

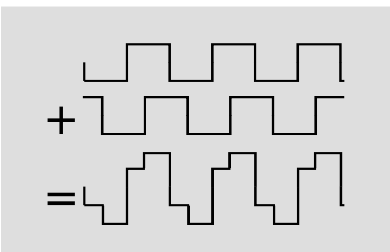


p / Graphs of loudness versus frequency for the vowel "ah," sung as three different musical notes. G is consonant with D, since every overtone of G that is close to an overtone of D (marked "*") is at exactly the same frequency. G and C♯ are dissonant together, since some of the overtones of G (marked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

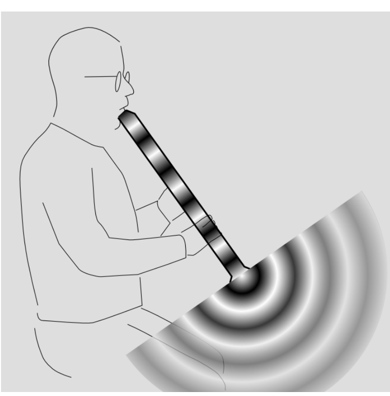


q / Standing waves on a rope. (*PSSC Physics.*)





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.



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

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.

Standing waves

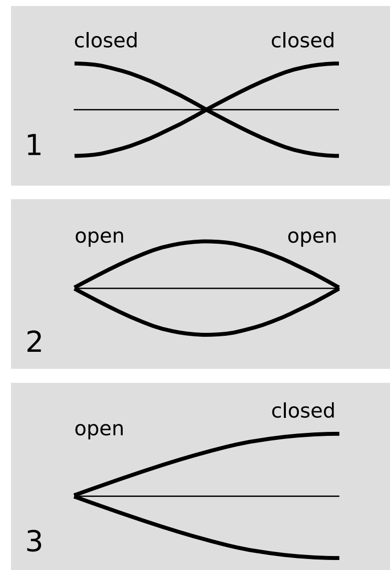
Figure q shows sinusoidal wave patterns made by shaking a rope. I used to enjoy doing this at the bank with the pens on chains, back in the days when people actually went to the bank. You might think that I and the person in the photos had to practice for a long time in order to get such nice sine waves. 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 or the human vocal tract 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.



t / 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.

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 t/1.

Figure s 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, t/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.

u / A concert flute is an asymmetric air column, open at one end and closed at the other.

Finally, we can have an asymmetric tube: closed at one end and open at the other. A common example is the concert flute, u. 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 t/3.

If both ends are open (as in the flute) or both ends closed (as in some organ pipes), then the standing wave pattern must be symmetric. The lowest-frequency wave fits half a wavelength in the tube, t/2-3.

If both ends are open (as in the flute) or both ends closed (as in some organ pipes), then the standing wave pattern must be symmetric. The lowest-frequency wave fits half a wavelength in the tube.

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, p. 708

The speed of sound *example 13*

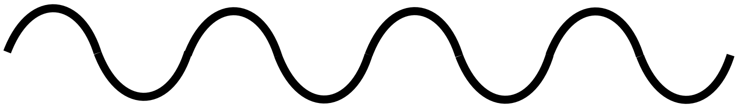
We can get a rough and ready derivation of the equation for the speed of sound by analyzing the standing waves in a cylindrical air column as a special type of Helmholtz resonance (example 23 on page 258), in which the cavity happens to have the same cross-sectional area as the neck. Roughly speaking, 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 $v = f\lambda = (\ldots)\sqrt{\gamma P_o/\rho}$, where the factor $(\ldots)$ represents numerical stuff that we can't possibly hope to have gotten right with such a crude argument. The correct result is in fact $v = \sqrt{\gamma P_o/\rho}$. 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 γ.

This chapter is summarized on page 729. Notation and terminology are tabulated on pages 718-719.

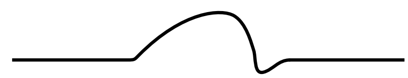
Problems

The symbols √, , etc. are explained on page 304.

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

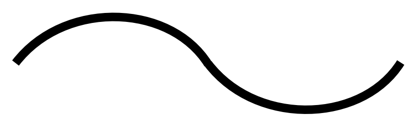


Trace this graph onto another piece of paper, and then sketch below it the corresponding graphs that would be obtained if
(a) the amplitude and frequency were doubled while the velocity remained the same;
(b) the frequency and velocity were both doubled while the amplitude remained unchanged;
(c) the wavelength and amplitude were reduced by a factor of three while the velocity was doubled.
Problem by Arnold Arons.



Problem 2

2 (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.
Problem by Arnold Arons.



Problem 3.

3 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.
Problem by Arnold Arons.

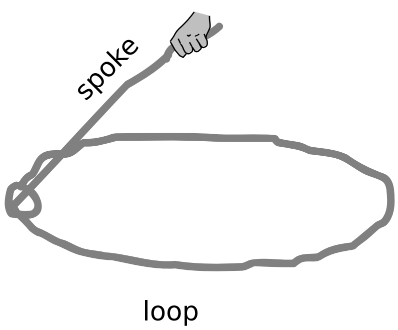
4 Find an equation for the relationship between the Doppler-shifted frequency of a wave and the frequency of the original wave, for the case of a stationary observer and a source moving directly toward or away from the observer.

5 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.

6 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. ▷ Answer, p. 712
(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.

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

8 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?



Problem 9.

9 The simplest trick with a lasso is to spin a flat loop in a horizontal plane. The whirling loop of a lasso is kept under tension mainly due to its own rotation. Although the spoke's force on the loop has an inward component, we'll ignore it. The purpose of this problem, which is based on one by A.P. French, is to prove a cute fact about wave disturbances moving around the loop. As far as I know, this fact has no practical implications for trick roping! Let the loop have radius r and mass per unit length μ, and let its angular velocity be ω.
(a) Find the tension, T, in the loop in terms of r, μ, and ω. Assume the loop is a perfect circle, with no wave disturbances on it yet. ▷ Hint, p. 704 ▷ Answer, p. 712
(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.

10 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.
(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. ✓

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. ▷ Hint, p. 704 √

13 A B-flat clarinet (the most common kind) produces its lowest note, at about 230 Hz, when half of a wavelength fits inside its tube. Compute the length of the clarinet. ▷ Answer, p. 712

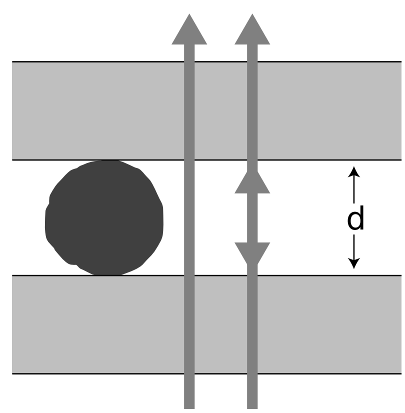
14 (a) A good tenor saxophone player can play all of the following notes without changing her fingering, simply by altering the tightness of her lips: E♭ (150 Hz), E♭ (300 Hz), B♭ (450 Hz), and E♭ (600 Hz). How is this possible?
(I'm not asking you to analyze the coupling between the lips, the reed, 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?

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 five or six harmonics of C and G, and determine whether these two notes will be consonant or dissonant.
(b) Do the same for C and B♭. (Recall that harmonics that differ by about 1-10% cause dissonance.)

C	261.6 Hz
D	293.7
E	329.6
F	349.2
G	392.0
A	440.0
B ♭	466.2

Problem 15.

16 (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. ▷ Answer, p. 712
(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.) ▷ Answer, p. 712



Problem 17.

17 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 a 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, what is the phase relationship between the two rays? (Recall that the phase can be inverted by a reflection.) With $d = 0$, is the interference constructive or destructive?
(b) If d is slowly increased from zero, what is the first value of d for which the interference is the same as at $d = 0$? 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.

Key to symbols:
easy   typical   challenging   difficult   very difficult
$\surd$ An answer check is available at www.lightandmatter.com.

Chapter 7
Relativity

Complaining about the educational system is a national sport among professors in the U.S., and I, like my colleagues, am often tempted to imagine a golden age of education in our country's past, or to compare our system unfavorably with foreign ones. Reality intrudes, however, when my immigrant students recount the overemphasis on rote memorization in their native countries and the philosophy that what the teacher says is always right, even when it's wrong.

Albert Einstein's education in late-nineteenth-century Germany was neither modern nor liberal. He did well in the early grades (the myth that he failed his elementary-school classes comes from a misunderstanding based on a reversal of the German numerical grading scale), but in high school and college he began to get in trouble for what today's edspeak calls "critical thinking."

a / Albert Einstein.

Indeed, there was much that deserved criticism in the state of physics at that time. There was a subtle contradiction between Maxwell's theory of electromagnetism and Galileo's principle that all motion is relative. Einstein began thinking about this on an intuitive basis as a teenager, trying to imagine what a light beam would look like if you could ride along beside it on a motorcycle at the speed of light. Today we remember him most of all for his radical and far-reaching solution to this contradiction, his theory of relativity, but in his student years his insights were greeted with derision from his professors. One called him a "lazy dog." Einstein's distaste for authority was typified by his decision as a teenager to renounce his German citizenship and become a stateless person, based purely on his opposition to the militarism and repressiveness of German society. He spent his most productive scientific years in Switzerland and Berlin, first as a patent clerk but later as a university professor. He was an outspoken pacifist and a stubborn opponent of World War I, shielded from retribution by his eventual acquisition of Swiss citizenship.

As the epochal nature of his work began to become evident, some liberal Germans began to point to him as a model of the "new German," but with the Nazi coup d'etat, staged public meetings began to be held at which Nazi scientists criticized the work of this ethnically Jewish (but spiritually nonconformist) giant of science. Einstein was on a stint as a visiting professor at Caltech when Hitler was appointed chancellor, and never returned to the Nazi state. World War II convinced Einstein to soften his strict pacifist stance, and he signed a secret letter to President Roosevelt urging research into the building of a nuclear bomb, a device that could not have been imagined without his theory of relativity. He later wrote, however, that when Hiroshima and Nagasaki were bombed, it made him wish he could burn off his own fingers for having signed the letter.

This chapter and the next are specifically about Einstein's theory of relativity, but Einstein also began a second, parallel revolution in physics known as the quantum theory, which stated, among other things, that certain processes in nature are inescapably random. Ironically, Einstein was an outspoken doubter of the new quantum ideas, being convinced that "the Old One [God] does not play dice with the universe," but quantum and relativistic concepts are now thoroughly intertwined in physics. The remainder of this book beyond the present pair of chapters is an introduction to the quantum theory, but we will continually be led back to relativistic ideas.

The structure of this chapter

From the modern point of view, electricity and magnetism becomes much simpler and easier to understand if it is encountered after relativity. Most schools' curricula, however, place electricity and magnetism before relativity. In such a curriculum, section 7.1 should be covered before electricity and magnetism, and then later in the course one can go back and cover all of chapter 7. This chapter is also designed so that it can be read without having previously covered waves.

7.1 Basic Relativity

Absolute, true, and mathematical time . . . flows at a constant rate without relation to anything external. . . Absolute space. . . without relation to anything external, remains always similar and immovable.

Isaac Newton (tr. Andrew Motte)

7.1.1 The principle of relativity

Galileo's most important physical discovery was that motion is relative. With modern hindsight, we restate this in a way that shows what made the teenage Einstein suspicious:

The principle of Galilean relativity: Matter obeys the same laws of physics in any inertial frame of reference, regardless of the frame's orientation, position, or constant-velocity motion.

Note that it only refers to matter, not light.

Einstein's professors taught that light waves obeyed an entirely different set of rules than material objects. They believed that light waves were a vibration of a mysterious medium called the ether, and that the speed of light should be interpreted as a speed relative to this ether. Thus although the cornerstone of the study of matter had for two centuries been the idea that motion is relative, the science of light seemed to contain a concept that a certain frame of reference was in an absolute state of rest with respect to the ether, and was therefore to be preferred over moving frames.

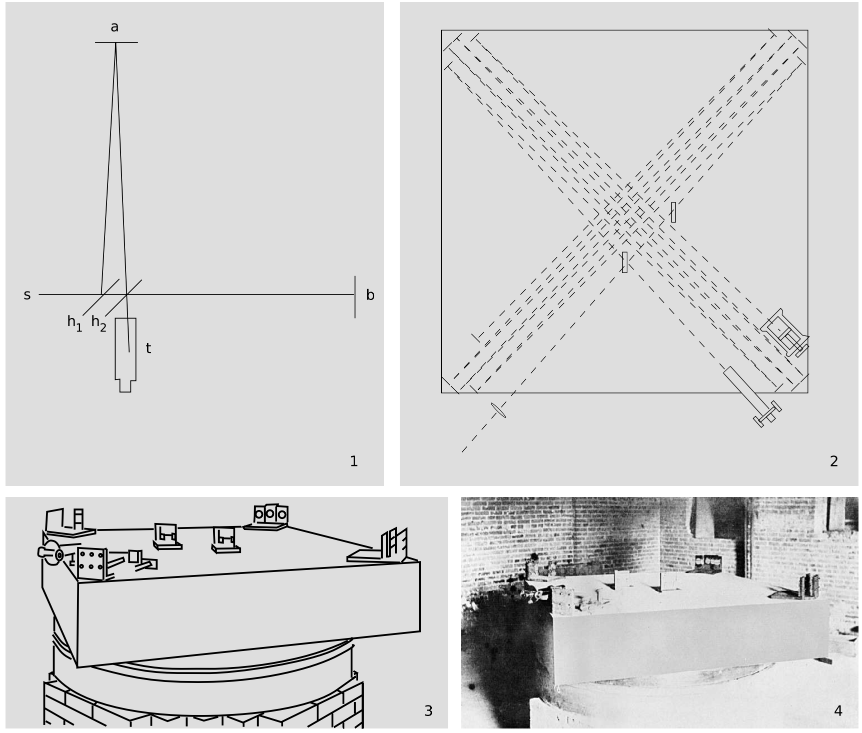
Now let's think about Albert Einstein's daydream of riding a motorcycle alongside a beam of light. In cyclist Albert's frame of reference, the light wave appears to be standing still. However, James Clerk Maxwell had already constructed a highly successful mathematical description of light waves as patterns of electric and magnetic fields. Einstein on his motorcycle can stick measuring instruments into the wave to monitor the electric and magnetic fields, and they will be constant at any given point. But an electromagnetic wave pattern standing frozen in space like this violates Maxwell's equations and cannot exist. Maxwell's equations say that light waves always move with the same velocity, notated c, equal to 3.0×10^8 m/s. Einstein could not tolerate this disagreement between the treatment of relative and absolute motion in the theories of matter on the one hand and light on the other. He decided to rebuild physics with a single guiding principle:

Einstein's principle of relativity: Both light and matter obey the same laws of physics in any inertial frame of reference, regardless of the frame's orientation, position, or constant-velocity motion.

7.1.2 Distortion of time and space

This is hard to swallow. 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. How, then, can a beam of light have the same speed as seen by someone who is chasing the beam?

In fact the strange constancy of the speed of light had shown up in the now-famous Michelson-Morley experiment of 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 in the ether hypothesis, 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

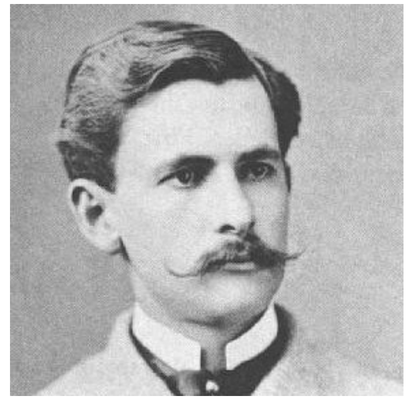


a / The Michelson-Morley experiment, shown in photographs, and drawings from the original 1887 paper. 1. A simplified drawing of the apparatus. A beam of light from the source, s, is partially reflected and partially transmitted by the half-silvered mirror h_1. The two half-intensity parts of the beam 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. Note that it is underground, in a room with solid brick walls.

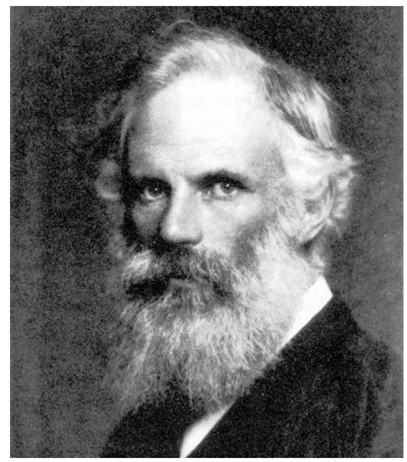
it through the ether. They were surprised when they found that the expected 0.01% change in the speed of light did not occur.

Although the Michelson-Morley experiment was nearly two decades in the past by the time Einstein published his first paper on relativity in 1905, it's unclear how much it influenced Einstein. Michelson and Morley themselves were uncertain about whether the result was to be trusted, or whether systematic and random errors were masking a real effect from the ether. There were a variety of competing theories, each of which could claim some support from the shaky data. Some physicists believed that the ether could be dragged along by matter moving through it, which inspired variations on the experiment that were conducted on mountaintops in thin-walled buildings, b, or with one arm of the appartus out in the open, and the other surrounded by massive lead walls. In the standard sanitized textbook version of the history of science, every scientist does his experiments without any preconceived notions about the truth, and any disagreement is quickly settled by a definitive experiment. In reality, this period of confusion about the Michelson-Morley experiment lasted for four decades, and a few reputable skeptics, including Miller, continued to believe that Einstein was wrong, and kept trying different variations of the experiment as late as the 1920's. Most of the remaining doubters were convinced by an extremely precise version of the experiment performed by Joos in 1930, although you can still find kooks on the internet who insist that Miller was right, and that there was a vast conspiracy to cover up his results.

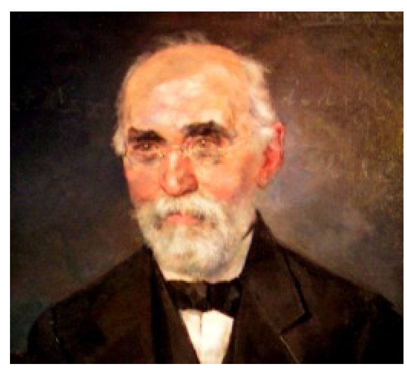
b / Dayton Miller thought that the result of the Michelson-Morley experiment could be explained because the ether had been pulled along by the dirt, and the walls of the laboratory. This motivated him to carry out a series of experiments at the top of Mount Wilson, in a building with thin walls.



c / Albert Michelson, in 1887, the year of the Michelson-Morley experiment.



d / George FitzGerald, 1851-1901.



e / Hendrik Lorentz, 1853-1928.

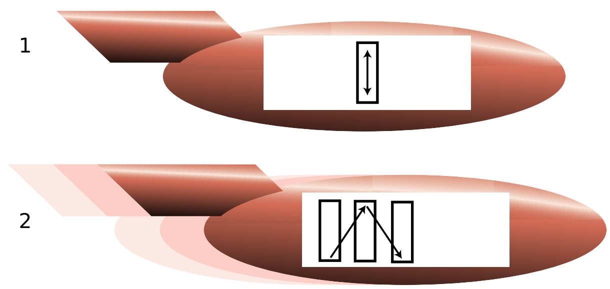
Before Einstein, some physicists who did believe the negative result of the Michelson-Morley experiment came up with explanations that preserved the ether. In the period from 1889 to 1895, Hendrik Lorentz and George FitzGerald suggested that the negative result of the Michelson-Morley experiment could be explained if the earth, and every physical object on its surface, was contracted slightly by the strain of the earth's motion through the ether.[1]

How did Einstein explain this strange refusal of light waves to obey the usual rules of addition and subtraction of velocities due to relative motion? He had the originality and bravery to suggest a radical solution. He decided that space and time must be stretched and compressed as seen by observers in different frames of reference. Since velocity equals distance divided by time, an appropriate distortion of time and space could cause the speed of light to come out the same in a moving frame. This conclusion could have been reached by the physicists of two generations before, on the day after Maxwell published his theory of light, but the attitudes about absolute space and time stated by Newton were so strongly ingrained that such a radical approach did not occur to anyone before Einstein.

[1] See discussion question F on page 316, and homework problem 22

Time distortion

Consider the situation shown in figure f. Aboard a rocket ship we have a tube with mirrors at the ends. If we let off a flash of light at the bottom of the tube, it will be reflected back and forth between the top and bottom. It can be used as a clock: by counting the number of times the light goes back and forth we get an indication of how much time has passed. (This may not seem very practical, but a real atomic clock does work on essentially the same principle.) Now imagine that the rocket is cruising at a significant fraction of the speed of light relative to the earth. Motion is relative, so for a person inside the rocket, f/1, there is no detectable change in the behavior of the clock, just as a person on a jet plane can toss a ball up and down without noticing anything unusual. But to an observer in the earth's frame of reference, the light appears to take a zigzag path through space, f/2, increasing the distance the light has to travel.



f / A light beam bounces between two mirrors in a spaceship.

If we didn't believe in the principle of relativity, we could say that the light just goes faster according to the earthbound observer. Indeed, this would be correct if the speeds were not close to the speed of light, and if the thing traveling back and forth was, say, a ping-pong ball. But according to the principle of relativity, the speed of light must be the same in both frames of reference. We are forced to conclude that time is distorted, and the light-clock appears to run more slowly than normal as seen by the earthbound observer. In general, a clock appears to run most quickly for observers who are in the same state of motion as the clock, and runs more slowly as perceived by observers who are moving relative to the clock.

We can easily calculate the size of this time-distortion effect. In the frame of reference shown in figure f/1, moving with the spaceship, let t_1 be the time required for the beam of light to move from the bottom to the top. An observer on the earth, who sees the situation shown in figure f/2, disagrees, and says this motion took a longer time t_2. Let v be the velocity of the spaceship relative to the earth. In frame 2, the light beam travels along the hypotenuse of a right triangle whose base has length

$$\text{base} = vt_2 \qquad .$$

Observers in the two frames of reference agree on the vertical distance traveled by the beam, i.e. the height of the triangle perceived in frame 2, and an observer in frame 1 says that this height is the distance covered by a light beam in time t_1, so the height is

$$\text{height} = ct_1 \quad .$$

The hypotenuse of this triangle is the distance the light travels in frame 2,

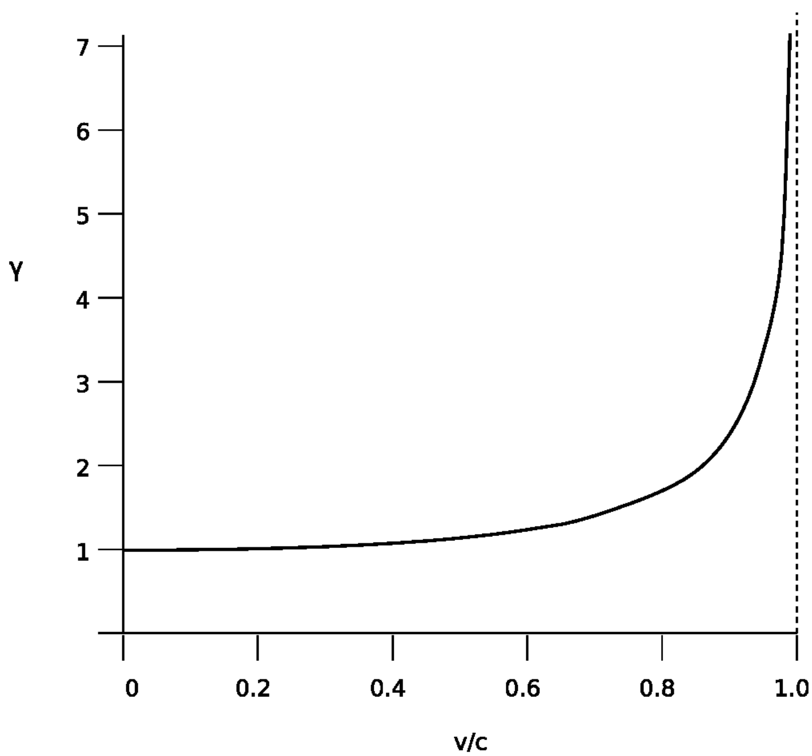
$$\text{hypotenuse} = ct_2 \quad .$$

Using the Pythagorean theorem, we can relate these three quantities, and solving for t_2 we find

$$t_2 = \frac{t_1}{\sqrt{1-(v/c)^2}} \quad .$$

The amount of distortion is given by the factor $1/\sqrt{1-(v/c)^2}$, and this quantity appears so often that we give it a special name, γ (Greek letter gamma),

$$\gamma = \frac{1}{\sqrt{1-(v/c)^2}} \quad . \qquad \text{[definition of the } \gamma \text{ factor]}$$



g / The behavior of the γ factor.

Self-Check
What is γ when v=0? What does this mean? ▷ Answer, p. 708

Distortion of space

The speed of light is supposed to be the same in all frames of reference, and a speed is a distance divided by the time. We can't change time without changing distance, since then the speed couldn't come out the same. A rigorous treatment requires some delicacy, but we postpone that to section 7.2 and state for now the apparently reasonable result that if time is distorted by a factor of γ, then lengths must also be distorted according to the same ratio. An object in motion appears longest to someone who is at rest with respect to it, and is shortened along the direction of motion as seen by other observers.

7.1.3 Applications

Nothing can go faster than the speed of light.

What happens if we want to send a rocket ship off at, say, twice the speed of light, $v = 2c$? Then γ will be $1/\sqrt{-3}$. But your math teacher has always cautioned you about the severe penalties for taking the square root of a negative number. The result would be physically meaningless, so we conclude that no object can travel faster than the speed of light. Even travel exactly at the speed of light appears to be ruled out for material objects, since then γ would be infinite.

Einstein had therefore found a solution to his original paradox about riding on a motorcycle alongside a beam of light, resulting in a violation of Maxwell's theory of electromagnetism. The paradox is resolved because it is impossible for the motorcycle to travel at the speed of light.

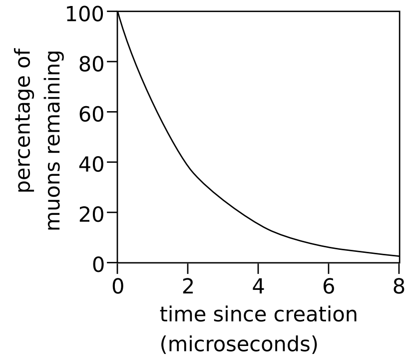
Most people, when told that nothing can go faster than the speed of light, immediately begin to imagine methods of violating the rule. For instance, it would seem that by applying a constant force to an object for a long time, we could give it a constant acceleration, which would eventually cause it to go faster than the speed of light. We will take up these issues in section 7.3.

Cosmic-ray muons

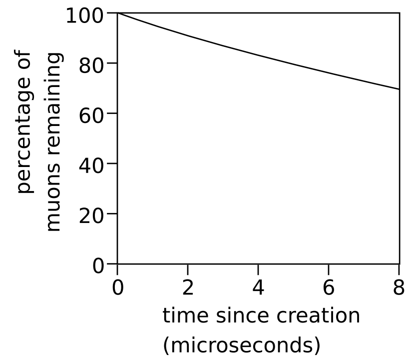
A classic experiment to demonstrate time distortion uses observations of cosmic rays. Cosmic rays are protons and other atomic nuclei from outer space. When a cosmic ray happens to come the way of our planet, the first earth-matter it encounters is an air molecule in the upper atmosphere. This collision then creates a shower of particles that cascade downward and can often be detected at the earth's surface. One of the more exotic particles created in these cosmic ray showers is the muon (named after the Greek letter mu, μ). The reason muons are not a normal part of our environment is that a muon is radioactive, lasting only 2.2 microseconds on the average before changing itself into an electron and two neutrinos. A muon can therefore be used as a sort of clock, albeit a self-destructing and somewhat random one! Figures h and i show the average rate at which a sample of muons decays, first for muons created at rest and then for high-velocity muons created in cosmic-ray showers. The second graph is found experimentally to be stretched out by a factor of about ten, which matches well with the prediction of relativity theory:

$$\begin{aligned}\gamma &= 1/\sqrt{1-(v/c)^2}\\ &= 1/\sqrt{1-(0.995)^2}\\ &\approx 10\end{aligned}$$

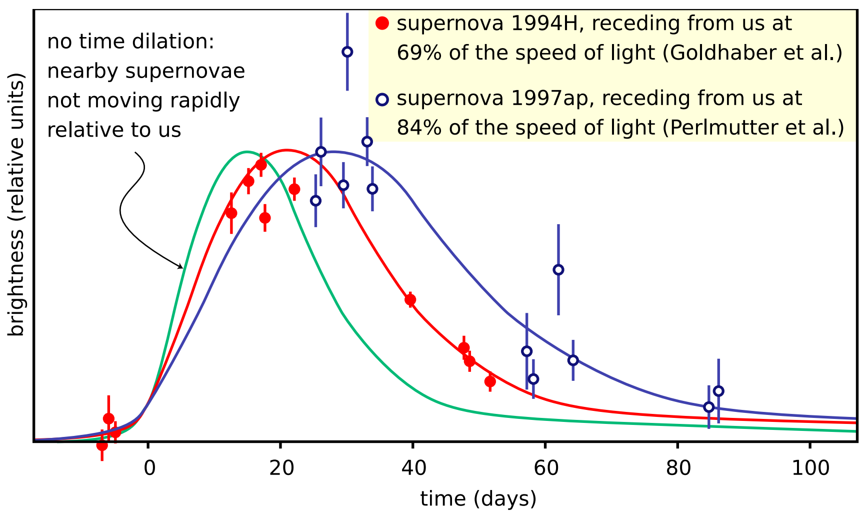
Since a muon takes many microseconds to pass through the atmosphere, the result is a marked increase in the number of muons that reach the surface.



h / Decay of muons created at rest with respect to the observer.



i / Decay of muons moving at a speed of 0.995c with respect to the observer.



j / Light curves of supernovae, showing a time-dilation effect for supernovae that are in motion relative to us.

Time dilation for objects larger than the atomic scale

Our world is (fortunately) not full of human-scale objects moving at significant speeds compared to the speed of light. For this reason, it took over 80 years after Einstein's theory was published before anyone could come up with a conclusive example of drastic time dilation that wasn't confined to cosmic rays or particle accelerators. Recently, however, astronomers have found definitive proof that entire stars undergo time dilation. The universe is expanding in the aftermath of the Big Bang, so in general everything in the universe is getting farther away from everything else. One need only find an astronomical process that takes a standard amount of time, and then observe how long it appears to take when it occurs in a part of the universe that is receding from us rapidly. A type of exploding star called a type Ia supernova fills the bill, and technology is now sufficiently advanced to allow them to be detected across vast distances. Figure j shows convincing evidence for time dilation in the brightening and dimming of two distant supernovae.

The twin paradox

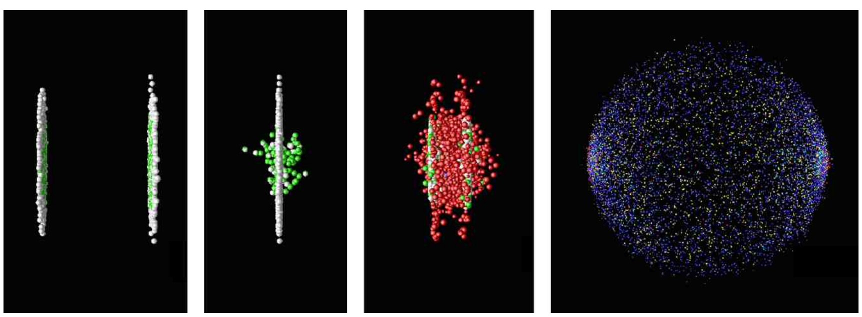
A natural source of confusion in understanding the time-dilation effect is summed up in the so-called twin paradox, which is not really a paradox. Suppose there are two teenaged twins, and one stays at home on earth while the other goes on a round trip in a spaceship at relativistic speeds (i.e. speeds comparable to the speed of light, for which the effects predicted by the theory of relativity are important). When the traveling twin gets home, he has aged only a few years, while his brother is now old and gray. (Robert Heinlein even wrote a science fiction novel on this topic, although it is not one of his better stories.)

The "paradox" arises from an incorrect application of the principle of relativity to a description of the story from the traveling twin's point of view. From his point of view, the argument goes, his homebody brother is the one who travels backward on the receding earth, and then returns as the earth approaches the spaceship again, while in the frame of reference fixed to the spaceship, the astronaut twin is not moving at all. It would then seem that the twin on earth is the one whose biological clock should tick more slowly, not the one on the spaceship. The flaw in the reasoning is that the principle of relativity only applies to frames that are in motion at constant velocity relative to one another, i.e., inertial frames of reference. The astronaut twin's frame of reference, however, is noninertial, because his spaceship must accelerate when it leaves, decelerate when it reaches its destination, and then repeat the whole process again on the way home. Their experiences are not equivalent, because the astronaut twin feels accelerations and decelerations. A correct treatment requires some mathematical complication to deal with the changing velocity of the astronaut twin, but the result is indeed that it's the traveling twin who is younger when they are reunited.

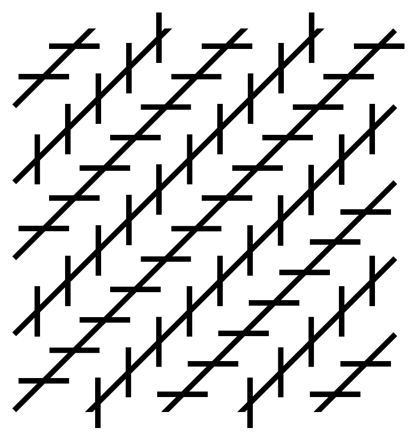
The twin "paradox" really isn't a paradox at all. It may even be a part of your ordinary life. The effect was first verified experimentally by synchronizing two atomic clocks in the same room, and then sending one for a round trip on a passenger jet. (They bought the clock its own ticket and put it in its own seat.) The clocks disagreed when the traveling one got back, and the discrepancy was exactly the amount predicted by relativity. The effects are strong enough to be important for making the global positioning system (GPS) work correctly. If you've ever taken a GPS receiver with you on a hiking trip, then you've used a device that has the twin "paradox" programmed into its calculations. Your handheld GPS box talks to a system onboard a satellite, and the satellite is moving fast enough that its time dilation is an important effect. So far no astronauts have gone fast enough to make time dilation a dramatic effect in terms of the human lifetime. The effect on the Apollo astronauts, for instance, was only a fraction of a second, since their speeds were still fairly small compared to the speed of light. (As far as I know, none of the astronauts had twin siblings back on earth!)

An example of length contraction

Figure k 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.



k / Colliding nuclei show relativistic length contraction.



Discussion question B

Discussion Questions

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

B A question that students often struggle with is whether time and space can really be distorted, or whether it just seems that way. Compare with optical illusions or magic tricks. How could you verify, for instance, that the lines in the figure are actually parallel? Are relativistic effects the same or not?

C If you were in a spaceship traveling at the speed of light (or extremely close to the speed of light), would you be able to see yourself in a mirror?

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

E What would the shapes of the two nuclei in the RHIC experiment 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?

F If you stick a piece of foam rubber out the window of your car while driving down the freeway, the wind may compress it a little. Does it make sense to interpret the relativistic length contraction as a type of strain that pushes an object's atoms together like this? How does this relate to discussion question E?

7.2 The Lorentz transformation

7.2.1 Coordinate transformations in general

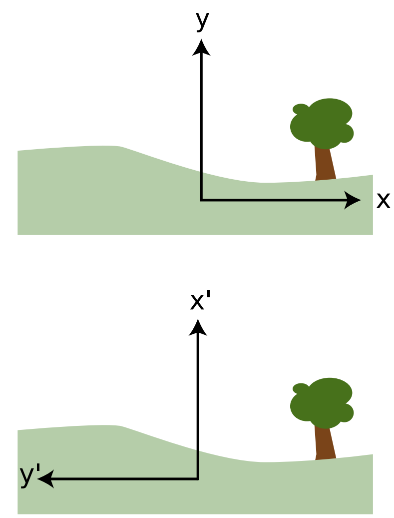
In section 7.1 the emphasis was on demonstrating some of the fundamental relativistic phenomena, without getting tangled up in too much mathematics. However, the issues that were glossed over there would come back to bite us if we never examined them carefully, and we haven't yet seen the full extent of relativity's attack on the traditional and intuitive concepts of space and time.

Rotation

For guidance, let's look at the mathematical treatment of the part of the principle of relativity that states that the laws of physics are the same regardless of the orientation of the coordinate system. Suppose that two observers are in frames of reference that are at rest relative to each other, and they set up coordinate systems with their origins at the same point, but rotated by 90 degrees, as in figure a. To go back and forth between the two systems, we can use the equations

$$\begin{aligned} x' &= y \\ y' &= -x \end{aligned}$$

A set of equations such as this one for changing from one system of coordinates to another is called a coordinate transformation, or just a transformation for short.



a / Two observers describe the same landscape with different coordinate systems.

Similarly, if the coordinate systems differed by an angle of 5 degrees, we would have

$$\begin{aligned} x' &= (\cos 5\,^\circ)x + (\sin 5\,^\circ)y \\ y' &= (-\sin 5\,^\circ)x + (\cos 5\,^\circ)y \end{aligned}$$

Since cos $5\,^\circ$=0.997 is very close to one, and sin $5\,^\circ$=0.087 is close to zero, the rotation through a small angle has only a small effect, which makes sense. The equations for rotation are always of the form

$$\begin{aligned} x' &= (\text{constant }\#1)x + (\text{constant }\#2)y \\ y' &= (\text{constant }\#3)x + (\text{constant }\#4)y \qquad . \end{aligned}$$

Galilean transformation for frames moving relative to each other

Einstein wanted to see if he could find a rule for changing between coordinate systems that were moving relative to each other. As a second warming-up example, let's look at the transformation between frames of reference in relative motion according to Galilean relativity, i.e. without any distortion of space and time. Suppose the x' axis is moving to the right at a velocity v relative to the x axis. The transformation is simple:

$$\begin{aligned} x' &= x - vt \\ t' &= t \end{aligned}$$

Again we have an equation with constants multiplying the variables, but now the variables are distance and time. The interpretation of the $-vt$ term is the observer moving with the origin x' system sees a steady reduction in distance to an object on the right and at rest in the x system. In other words, the object appears to be moving according to the x' observer, but at rest according to x. The fact that the constant in front of x in the first equation equals one tells us that there is no distortion of space according to Galilean relativity, and similarly the second equation tells us there is no distortion of time.

7.2.2 Derivation of the Lorentz transformation

Guided by analogy, Einstein decided to look for a transformation between frames in relative motion that would have the form

$$\begin{aligned} x' &= Ax + Bt \\ t' &= Cx + Dt \qquad . \end{aligned}$$

(Any form more complicated than this, for example equations including x^2 or t^2 terms, would violate the part of the principle of relativity that says the laws of physics are the same in all locations.) For historical reasons, this is called a Lorentz transformation. The constants A, B, C, and D would depend only on the relative velocity, v, of the two frames. Galilean relativity had been amply verified by experiment for values of v much less than the speed of light, so at low speeds we must have $A \approx 1$, $B \approx -v$, $C \approx 0$, and $D \approx 1$. For high speeds, however, the constants A and D would start to become measurably different from 1, providing the distortions of time and space needed so that the speed of light would be the same in all frames of reference.

Self-Check

What units would the constants *A*, *B*, *C*, and *D* need to have? ▷ Answer, p. 708

Natural units

Despite the reputation for difficulty of Einstein's theories, the derivation of Einstein's transformations is fairly straightforward. The algebra, however, can appear more cumbersome than necessary unless we adopt a choice of units that is better adapted to relativity than the metric units of meters and seconds. The form of the transformation equations shows that time and space are not even entirely separate entities. Life is easier if we adopt a new set of units:

-Time is measured in seconds.
-Distance is also measured in units of seconds. A distance of one second is how far light travels in one second of time.

In these units, the speed of light equals one by definition:

$$c = \frac{1 \text{ second of distance}}{1 \text{ second of time}} = 1$$

All velocities are represented by unitless numbers in this system, so for example $v = 0.5$ would describe motion at half the speed of light.

Converting a formula from ordinary units to natural units *example 1*
In ordinary units, the equation for the Lorentz factor γ is

$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} \qquad .$$

Suppose we want to reexpress this in natural units. One way of doing it would be to redo the derivation on page 310, but with the simplifying assumption of $c = 1$. However, this would just mean eliminating any c that appears in a multiplication or division, so rather than retracing the trail of breadcrumbs, we can just eliminate the c's from the final result:

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

Converting a formula from natural units to ordinary units *example 2*
In reality, the reason for using natural units in the first place is to make derivations simpler. Therefore a much more common situation is that you get a formula in natural units as the result of some symbolic calculation, but then you need to convert it to ordinary units in order to plug in numbers that you have in ordinary units. Working in the opposite direction, we observe that the equation $\gamma = 1/\sqrt{1-v^2}$ doesn't make any sense in metric units, because you can't take a unitless number like 1 and subtract from it a quantity that has units of $\mathrm{m}^2/\mathrm{s}^2$. That would be like subtracting three gallons from seven miles! Even if we don't remember how the formula was derived, we know that the derivation in natural units and the derivation in ordinary units could only have differed by the presence or absence of factors of c in various places. Therefore, we know we can recover the result in metric units simply by inserting factors of c wherever they're needed in order to turn the nonsense into sense. One way of doing this would be to divide the v^2 term by c^2, which makes it into a unitless quantity that it's possible to subtract from 1. The result is $\gamma = 1/\sqrt{1-v^2/c^2}$. (It might seem like the result wouldn't be unique, since we could instead fix the 1 by multiplying it by c^2, giving $c^2 - v^2$ inside the square root. However, the units of the right-hand side of the equation would then be s/m, so we'd also need to change the left-hand side to γ/c, and then the result would be exactly equivalent $\gamma = 1/\sqrt{1-v^2/c^2}$.)

A black hole *example 3*
Here's an example where you don't even have the option of rederiving the equation from scratch. A black hole of mass m has an invisible spherical boundary of radius r surrounding it, and any object that comes in closer than that can never escape. The radius is given in natural units by $r = 2Gm$, where G is Newton's gravitational constant. All of this can be partly, but not completely, explained using special relativity. For instance, we can try calculating the distance at which escape velocity becomes greater than the speed of light. However, this would be a swindle, because special relativity doesn't include gravity — to get a correct relativistic treatment of gravity, we'd need general relativity, which is beyond the scope of this book. (One way you can tell that

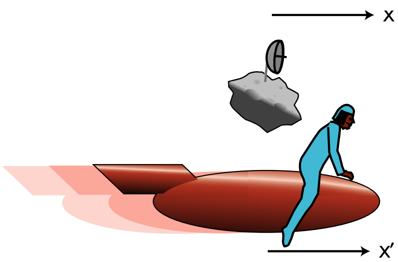
the naive calculation using escape velocity isn't really correct is that it makes it sound as though an object could still be hoisted out of a black hole on a cable, but that's actually not true according to general relativity.) But even though you don't know enough physics to derive the equation correctly from scratch, you can still convert it to metric units. The units of G are $\text{m}^3/\text{kg}\cdot\text{s}^2$, so the units of Gm are m^3/s^2. This doesn't equal meters, so the equation $r = 2Gm$ is nonsense if you interpret it directly in metric units. However, the units do work if you change it to $r = 2Gm/c^2$, so that's what the equation must be in metric units.

Derivation of the Lorentz transformation

We now want to find out how the constants A, B, C, and D in the transformation equations

$$x' = Ax + Bt$$
$$t' = Cx + Dt$$

depend on velocity. For vividness, we imagine that the x, t frame is defined by an asteroid at $x = 0$, and the x', t' frame by a rocket ship at $x' = 0$. The rocket ship is coasting at a constant speed v relative to the asteroid, and as it passes the asteroid they synchronize their clocks to read $t = 0$ and $t' = 0$.



b / The x,t frame is defined from the asteroid, and the x',t' frame from the astronaut.

Asteroid time as perceived by the rocket

In section 7.1, we've already found that a clock seems to run more slowly by a factor of γ to an observer in motion with respect to the clock. A clock on the asteroid has $x = 0$, so if the rocket pilot monitors the ticking of a clock on the asteroid via radio signals, the Lorentz transformation gives $t' = Dt$. The idea of time running more slowly by a factor of γ is expressed by $t' = \gamma t$, so we have

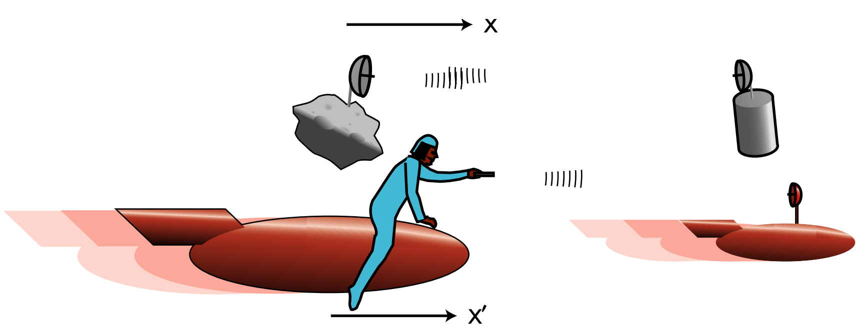
$$D = \gamma \qquad .$$

Asteroid's motion as seen by the rocket

Straightforward algebra can be used to reverse the transformation equations so that they give x and t in terms of x' and t'. The result for x is $x = (Dx' - Bt')/(AD - BC)$. The asteroid's frame of reference has its origin defined by the asteroid itself, so the asteroid is always at $x = 0$. In the rocket's frame, the asteroid falls behind according to the equation $x' = -vt'$, and substituting this into the equation for x gives $0 = (-Dvt' - Bt')/(AD - BC)$. This requires us to have $B = -vD$, or

$$B = -v\gamma \qquad .$$

So far in this derivation, we've been able to avoid talking about events that happen in different places and at different times, but we won't be able to avoid that anymore. We need to compare the perception of space and time by observers on the rocket and the asteroid, but this can be a bit tricky because our usual ideas about measurement contain hidden assumptions. If, for instance, we want to measure the length of a box, we imagine we can lay a ruler down on it, take in the scene visually, and take the measurement using the ruler's scale on the right side of the box while the left side of the box is simultaneously lined up with the butt of the ruler. The assumption that we can take in the whole scene at once with our eyes is, however, based on the assumption that light travels with infinite speed to our eyes. Since we will be dealing with relative motion at speeds comparable to the speed of light, we have to spell out our methods of measuring distance.



c / To discuss distances and time intervals between different events, we imagine that each frame of reference has observers in more than one place.

We will therefore imagine an explicit procedure for the asteroid and the rocket pilot to make their distance measurements: they send electromagnetic signals (light or radio waves) back and forth to their own remote stations. For instance the asteroid's station will send it a message to tell it the time at which the rocket went by. The asteroid's station is at rest with respect to the asteroid, and the rocket's is at rest with respect to the rocket (and therefore in motion with respect to the asteroid).

The measurement of time is likewise fraught with danger if we are careless, which is why we have had to spell out procedures for the synchronization of clocks between the asteroid and the rocket. The asteroid must also synchronize its clock with its remote stations's clock by adjusting them until flashes of light released by both the asteroid and its station at equal clock readings are received on the opposite sides at equal clock readings. The rocket pilot must go through the same kind of synchronization procedure with her remote station.

Rocket's motion as seen by the asteroid

The origin of the rocket's x', t' frame is defined by the rocket itself, so the rocket always has $x' = 0$. Let the asteroid's remote station be at position x in the asteroid's frame. The asteroid sees the rocket travel at speed v, so the asteroid's remote station sees the rocket pass it when x equals vt. The equation $x' = Ax + Bt$ then becomes $0 = Avt + Bt$, which implies a relationship between A and B: $B = -Av$. (In the Galilean version, we had $B = -v$ and

$A = 1$.) Thus,

$$A = \gamma \qquad .$$

This boils down to a statement that length contraction occurs in the same proportion as time dilation, as we'd already argued less rigorously.

Agreement on the speed of light

Suppose the rocket pilot releases a flash of light in the forward direction as she passes the asteroid at $t = t' = 0$. As seen in the asteroid's frame, we might expect this pulse to travel forward faster than normal because it was emitted by the moving rocket, but the principle of relativity tells us this is not so. The flash reaches the asteroid's remote station when x equals ct, and since we are working in natural units, this is equivalent to $x = t$. The speed of light must be the same in the rocket's frame, so we must also have $x' = t'$ when the flash gets there. Setting $x' = Ax + Bt$ equal to $t' = Cx + Dt$ and substituting in $x = t$, we find $A + B = C + D$, so we deduce $C = B + A - D = B$, or

$$C = -v\gamma \qquad .$$

We have now arrived at the correct relativistic equation for transforming between frames in relative motion. For completeness, I will include, without proof, the trivial transformations of the y and z coordinates.

$$\begin{aligned} x' &= \gamma x - v\gamma t \\ t' &= -v\gamma x + \gamma t \\ y' &= y \\ z' &= z \end{aligned}$$

These equations are valid provided that (1) the two coordinate systems coincide at $t = t' = 0$; and (2) the observer in the x', y', z', t' frame is moving at velocity v relative to the x, y, z, t frame, and the motion is in the direction of the x axis.

Self-Check

What happens to the Lorentz transformation in the case where *v* equals zero? ▷ Answer, p. 708

We now turn to some subversive consequences of these equations.

7.2.3 Spacetime

No absolute time

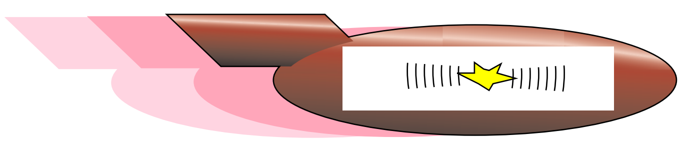
The fact that the equation for time is not just $t' = t$ tells us we're not in Kansas anymore — Newton's concept of absolute time is dead. One way of understanding this is to think about the steps described for synchronizing the four clocks:

(1) The asteroid's clock — call it A1 — was synchronized with the clock on its remote station, A2.

(2) The rocket pilot synchronized her clock, R1, with A1, at the moment when she passed the asteroid.

(3) The clock on the rocket's remote station, R2, was synchronized with R1.

Now if A2 matches A1, A1 matches R1, and R1 matches R2, we would expect A2 to match R2. This cannot be so, however. The rocket pilot released a flash of light as she passed the asteroid. In the asteroid's frame of reference, that light had to travel the full distance to the asteroid's remote station before it could be picked up there. In the rocket pilot's frame of reference, however, the asteroid's remote station is rushing at her, perhaps at a sizeable fraction of the speed of light, so the flash has less distance to travel before the asteroid's station meets it. Suppose the rocket pilot sets things up so that R2 has just enough of a head start on the light flash to reach A2 at the same time the flash of light gets there. Clocks A2 and R2 cannot agree, because the time required for the light flash to get there was different in the two frames. Thus, two clocks that were initially in agreement will disagree later on.



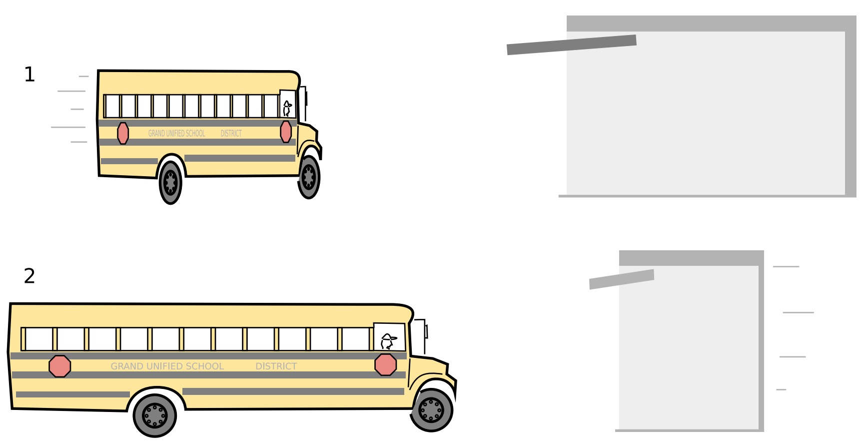
d / Different observers don't agree that the flashes of light hit the front and back of the ship simultaneously.

No simultaneity

Part of the concept of absolute time was the assumption that it was valid to say things like, "I wonder what my uncle in Beijing is doing right now." In the nonrelativistic world-view, clocks in Los Angeles and Beijing could be synchronized and stay synchronized, so we could unambiguously define the concept of things happening simultaneously in different places. It is easy to find examples, however, where events that seem to be simultaneous in one frame of reference are not simultaneous in another frame. In figure d, a flash of light is set off in the center of the rocket's cargo hold. According to a passenger on the rocket, the flashes have equal distances to travel to reach the front and back walls, so they get there simultaneously. But an outside observer who sees the rocket cruising by at high speed will see the flash hit the back wall first, because the wall is rushing up to meet it, and the forward-going part of the flash hit the front wall later, because the wall was running away from it. Only when the relative velocity of two frames is small compared to the speed of light will observers in those frames agree on the simultaneity of events.

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 paradox arises when we shut the door and then quickly slam on the brakes of the bus. An observer in the garage's frame of reference will claim that the bus fit in the garage because of its contracted length. 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. The person in the garage's frame can shut the door at an instant he perceives to be simultaneous with the front bumper's arrival 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.



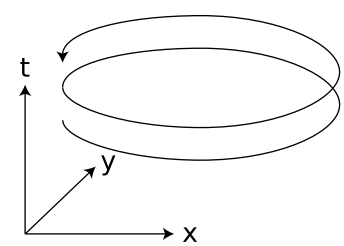
e / In the garage's frame of reference, 1, the bus is moving, and can fit in the garage. In the bus's frame of reference, the garage is moving, and can't hold the bus.

Spacetime

We consider x, y, and z to be three axes in space. There can be no physical distinction between them, since rotation transformations like the ones given on page 317 can interchange or mix together the three coordinates. One observer can say that two different points in space have the same value of x, but a different observer working in a differently oriented coordinate system would say they have different x values; this shows that the x axis can't be singled out from the others in any physically meaningful way.

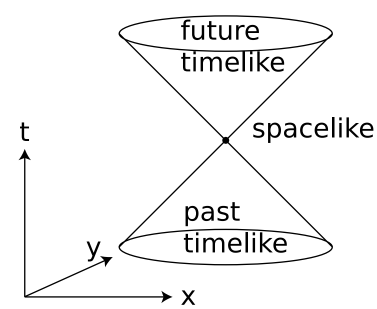
In relativity, Lorentz transformations can mix the *space* variables with the *time* variable, and different observers will not necessarily agree on whether two events are simultaneous, i.e. on whether they have the same t. Thus there is no unique, physically meaningful way to defined a time axis and set it apart from the space axes. The three space coordinates and the time coordinate are really just four coordinates that allow us to describe points in a four-dimensional space, which we call *spacetime*.

What does this mean? We can't visualize four dimensions. One technique for visualizing spacetime is to ignore one of the space dimensions, cutting the total number back down to three. For instance, the earth orbits the sun within a certain plane, so if we define x and y axes within the orbital plane, the z axis is not very interesting, and we can ignore it for purposes of describing the earth's motion. If we visualize x, y, and t in three dimensions, the points in spacetime visited by the earth form a helical curve, f. The earth stays "above" the circle defined by its orbit in the x-y plane. The earth will visit the same x-y point over and over, but it never visits the same spacetime point again, because t changes by one year over each orbit. Every point in spacetime is called an *event*.



f / The world-line of the earth as it orbits the sun.

It might seem that the mixing of space and time is so insane that virtually anything can happen, but there is a good way of bringing order to the madness. Continuing with the same mode of visualization, imagine that at a certain moment at a certain point in space (at a certain event in spacetime), a flash of light is emitted. The light pulse travels outward in all directions, forming an expanding spherical shell. If we ignore one of the space dimensions, this becomes an expanding circle, like a ripple on a pond. In x-y-t space, the ripple becomes a cone, g. If we are present at the emission of the light pulse, then events inside this *light cone* are ones that we may be able to observe in the future. For instance, if we just stay put, we will be present at every event that lies along the axis of the cone. If we were to move off at 99.99999% of the speed of light, we could witness a bunch of the events along a world line lying just inside the cone. Events in the region of spacetime outside the light cone, however, are ones we can never experience firsthand. For instance, consider an event that is happening right now, in a galaxy far, far away. This event lies in the spacelike region outside the light cone,



g / The light cone.

directly away from the tip of the cone in the direction perpendicular to its axis. We can't travel a large distance in an instant, since it's impossible to go faster than light, so we can never get to this event. The spacelike region consists of points whose distance from us in space is greater (in natural units) than their distance from us in time, so we can never visit them without traveling faster than light.

The great thing about the light cone is that everyone has to agree on it. A Lorentz transformation will skew and distort all of spacetime, but it will leave the light cone alone, since the light cone is defined by a flash of light, and all observers agree on the speed of light.

It's also possible to make a light cone that extends backward into the past. These are events that we can remember or get information about, but that we can never visit again because they lie in our past.

The spacetime interval

The light cone is helpful because it stays the same when you do a Lorentz transformation. Is there anything else that stays the same? For guidance, consider rotations. In a rotation, distances and angles stay the same. Now if you were an ant living on a telephone wire, you'd only know about one dimension, and the only type of rotation you'd be able to understand would be a 180-degree flip, which wouldn't change the lengths of line segments. But suppose there was some unsuspected second dimension to space. A one-meter line segment, with $\Delta x = 1.0$ m, that was rotated 60 degrees into this second dimension would then have $\Delta x = 0.5$ m. The existence of the newly discovered dimension has broken the rule that rotations don't change values of Δx. However, an ant named Pythagoras might realize that there was a new way to redefine distance, as $\sqrt{\Delta x^2 + \Delta y^2}$, so distances would stay the same in two-dimensional space. For reasons that will become apparent shortly, it turns out to be more convenient to work with the square of the distance, $\Delta x^2 + \Delta y^2$, which we call an *interval*. Generalizing to three dimensions is a snap: we just define the interval as $\Delta x^2 + \Delta y^2 + \Delta z^2$.

Now what about the generalization to four dimensions? The quantity $\Delta x^2 + \Delta y^2 + \Delta z^2$ doesn't stay the same when we do a Lorentz transformation — distances get contracted. We might try $\Delta x^2 + \Delta y^2 + \Delta z^2 + \Delta t^2$ (in natural units), but that wouldn't work, as you can easily verify by trying an example. We already know that the light cone stays the same under a Lorentz transformation, and the light cone is defined by the equation (distance)/(time)=(speed of light), or (distance)=(time) in natural units, which is equivalent to

$$\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} = \Delta t$$

or

$$\Delta x^2 + \Delta y^2 + \Delta z^2 - \Delta t^2 = 0 \qquad .$$

With this motivation, we define[2] the interval between two events in spacetime as $\Delta x^2 + \Delta y^2 + \Delta z^2 - \Delta t^2$. With this definition, the interval stays the same under a Lorentz transformation. Events with a zero interval between them are in each other's light cones. A positive interval indicates a spacelike relationship, and a negative one shows a timelike relationship. The possibility of a negative result is the reason for working with the quantity $\Delta x^2 + \Delta y^2 + \Delta z^2 - \Delta t^2$ rather than $\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2 - \Delta t^2}$.

The interval and the correspondence principle *example 4*

What happens if we try to interpret the interval in a nonrelativistic context? The equation for the interval is expressed in natural units. Suppose we take two events in the everyday world. For instance, my dog barks, and I turn around to look at why he's barking. In natural units, both the Δt and the Δx between these two events would be expressed in units of seconds. The Δt is something like a second or half a second. The Δx would be a few meters in SI units, but in natural units, it converts to the time light would take to travel a few meters, which would be on the order of 10^{-8} s. We find, then, that the Δx makes a negligibly small contribution to the interval compared to the Δt. In other words, we never encounter spacelike or lightlike intervals in our everyday experience; all the intervals we experience directly are timelike. Nonrelativistically, when nothing is traveling at an appreciable fraction of the speed of light, the interval is essentially the same as Δt^2.

How are we to interpret this? The interval is the same in all frames of reference, and in nonrelativistic situations, this means that Δt must be the same for all observers. But this is exactly what we expect in Galilean relativity. In the Galilean transformations, the absolute nature of time is expressed in the equation $t' = t$. This is an example of the correspondence principle, which states that when a new physical theory supersedes an old one, it must be consistent with the old one within the old one's domain of validity.

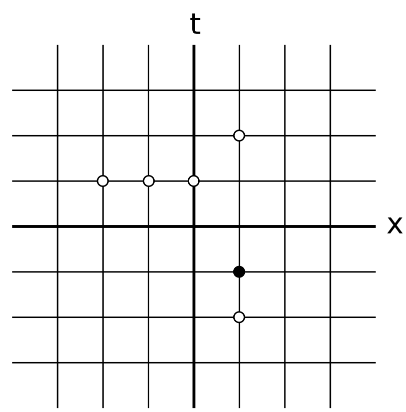
Discussion Questions

A The graphs for discussion questions A and B represent spacetime with one space dimension. Each square on the graph is one light-year wide and one year tall. In A, the dots represent civilizations in different times and on different planets. The planets all happen to lie along the same line, so we don't need y or z coordinates to show their locations. None of the planets are in rapid motion relative to each other, so we don't have to worry right now about whose frame of reference the graph depicts — they all agree.

(1) How many different planets are represented on the graph?

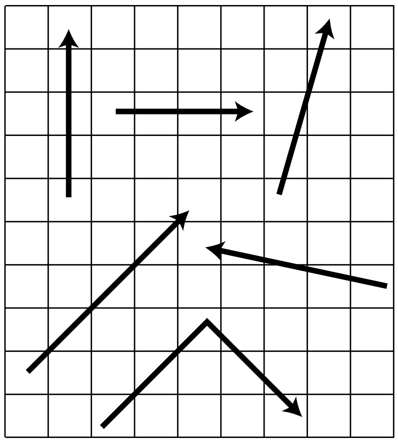
(2) The black dot represents a historian. Draw the historian's light cone.

(3) The historian is interested in getting information that has been preserved by civilizations in her past, and also hopes to preserve information about her own civilization for those in her future. By what methods could this be accomplished for the events shown as white circles?



Discussion question A.

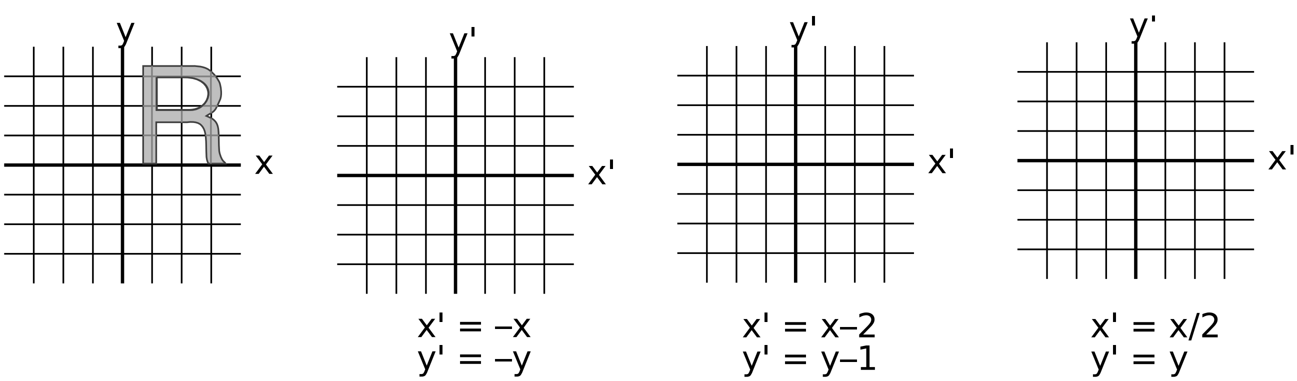
[2]The definition $\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$ is equally valid. It's just a matter of convention. You have to be careful when using the literature to make sure you don't mix equations that assume inconsistent choices of the signs.



Discussion question B.

B Which of the world lines are possible, and which are impossible? Could they represent light? Matter? What does this have to do with the light cone?

C The graph below is unlike the other ones we've been considering because it represents two dimensions of *space* at a certain instant. In each case, show what happens to the letter R when you do the transformation. Are the laws of physics the same in the x', y' coordinates? In other words, if the transformation is suddenly applied to your physics lab, will experiments still come out the same?



D The following graphs show one space dimension and one time dimension.

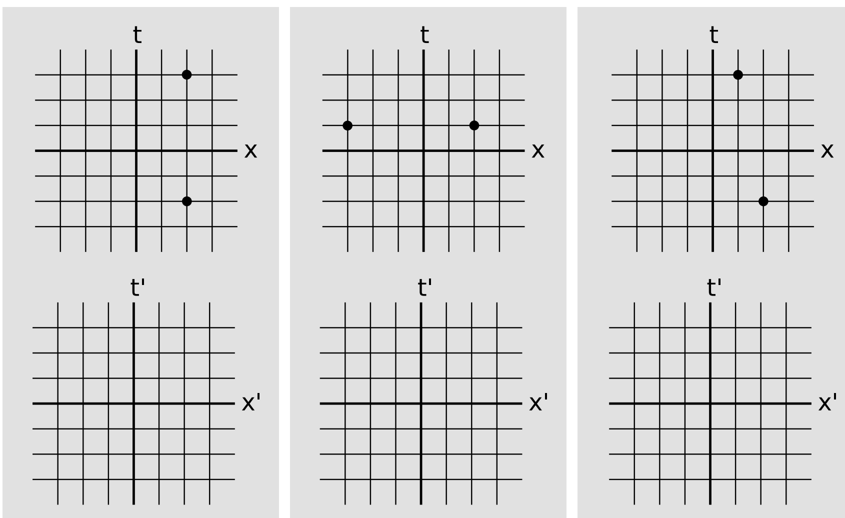
(1) In each case, apply the transformation

$$x' = x + (0.2)t$$
$$t' = t$$

to the indicated events.

(2) How would you interpret the meaning of the transformation?

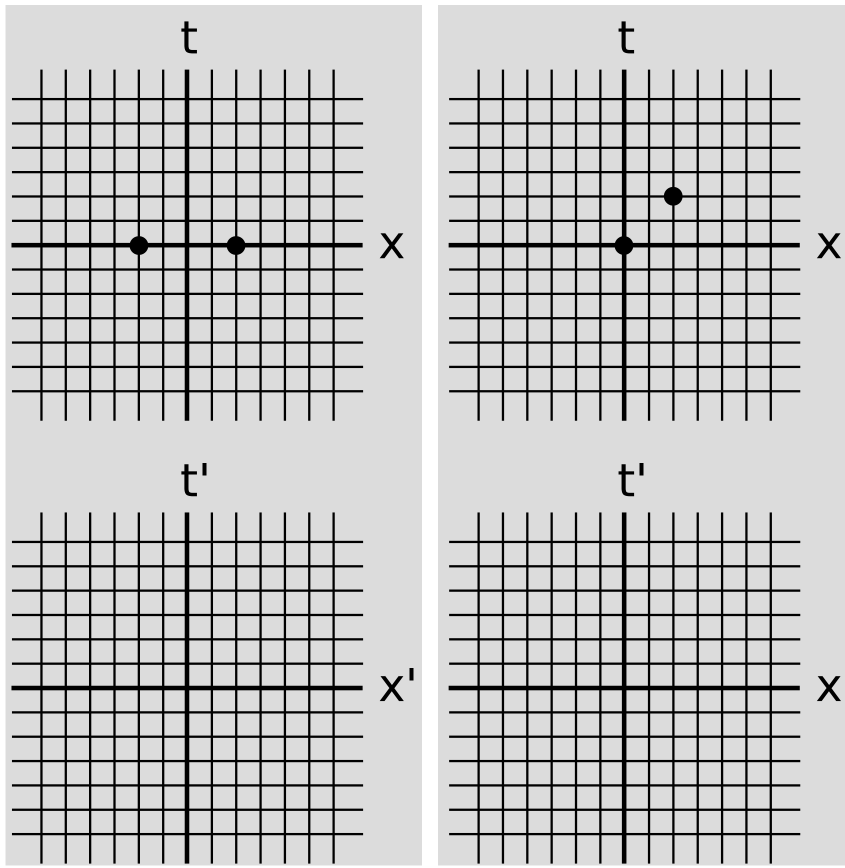
(3) In each case are there special relationships between the two events? Do observers in these two frames of reference agree on these relationships?



Discussion question D.

E This is similar to discussion question D, but with the following transformation:

$$x' = (1.67)x + (-1.34)t$$
$$t' = (-1.34)x + (1.67)t$$



h / Discussion question E.

7.3 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 conversion of mass to energy and energy to mass, as described by the famous equation $E = mc^2$. To cut down on the level of mathematical detail, I have relegated most of the derivations to page 697, presenting mainly the results and their physical explanations in this section.

7.3.1 Invariants

The discussion has the potential to become very confusing very quickly because some quantities, force for example, are perceived differently by observers in different frames, whereas in Galilean relativity they were the same in all frames of reference. To clear the smoke it will be helpful to start by identifying quantities that we can depend on not to be different in different frames. We have already seen how the principle of relativity requires that the speed of light is the same in all frames of reference. We say that c is invariant.

Another important invariant is mass. This makes sense, because the principle of relativity states that physics works the same in all reference frames. The mass of an electron, for instance, is the same everywhere in the universe, so its numerical value is one of the basic laws of physics. We should therefore expect it to be the same in all frames of reference as well. (Just to make things more confusing, about 50% of all books say mass is invariant, while 50% describe it as changing. It is possible to construct a self-consistent framework of physics according to either description. Neither way is right or wrong, the two philosophies just require different sets of definitions of quantities like momentum and so on. For what it's worth, Einstein eventually weighed in on the mass-as-an-invariant side of the argument. The main thing is just to be consistent.)

A third invariant is electrical charge. This has been verified to high precision because experiments show that an electric field does not produce any measurable force on a hydrogen atom. If charge varied with speed, then the electron, typically orbiting at about 1% of the speed of light, would not exactly cancel the charge of the proton, and the hydrogen atom would have a net charge.

7.3.2 Combination of velocities

The impossibility of motion faster than light is a radical difference between relativistic and nonrelativistic physics, and we can get at most of the issues in this section by considering the flaws in various plans for going faster than light. The simplest argument of this kind is as follows. Suppose Janet takes a trip in a spaceship, and accelerates until she is moving at $v = 0.9$ (90% of the speed of light in natural units) relative to the earth. She then launches a space probe in the forward direction at a speed $u = 0.2$ relative to her ship. Isn't the probe then moving at a velocity of 1.1 times the speed of light relative to the earth?

The problem with this line of reasoning is that the distance covered by the probe in a certain amount of time is shorter as seen by an observer in the earthbound frame of reference, due to length contraction. Velocities are therefore combined not by simple addition but by a more complex method, which we derive on page 697 by performing two transformations in a row. In our example, the first transformation would be from the earth's frame to Janet's, the second from Janet's to the probe's. The result is

$$v_{combined} = \frac{u+v}{1+uv} \qquad .$$

Janet's probe *example 5*

Applying the equation to Janet's probe, we find

$$\begin{aligned} v_{combined} &= \frac{0.9+0.2}{1+(0.9)(0.2)} \\ &= 0.93 \qquad , \end{aligned}$$

so it's still going quite a bit slower than the speed of light

Combination of velocities in unnatural units *example 6*

In a system of units, like the metric system, with $c \neq 1$, all our symbols for velocity should be replaced with velocities divided by c, so we have

$$\frac{v_{combined}}{c} = \frac{\frac{u}{c}+\frac{v}{c}}{1+\left(\frac{u}{c}\right)\left(\frac{v}{c}\right)} \qquad ,$$

or

$$v_{combined} = \frac{u+v}{1+uv/c^2} \qquad .$$

When u and v are both much less than the speed of light, the quantity uv/c^2 is very close to zero, and we recover the nonrelativistic approximation, $v_{combined} = u + v$.

The second example shows the correspondence principle at work: when a new scientific theory replaces an old one, the two theories must agree within their common realm of applicability.

7.3.3 Momentum and force

Momentum

We begin our discussion of relativistic momentum with another scheme for going faster than light. Imagine that a freight train moving at a velocity of 0.6 ($v = 0.6c$ in unnatural units) strikes a ping-pong ball that is initially at rest, and suppose that in this collision no kinetic energy is converted into other forms such as heat and sound. We can easily prove based on conservation of momentum that in a very unequal collision of this kind, the smaller object flies off with double the velocity with which it was hit. (This is because the center of mass frame of reference is essentially the same as the frame tied to the freight train, and in the center of mass frame both objects must reverse their initial momenta.) So doesn't the ping-pong ball fly off with a velocity of 1.2, i.e. 20% faster than the speed of light?

The answer is that since $p = mv$ led to this contradiction with the structure of relativity, $p = mv$ must not be the correct equation for relativistic momentum. Apparently $p = mv$ is only a low-velocity approximation to the correct relativistic result. We need to find a new expression for momentum that agrees approximately with $p = mv$ at low velocities, and that also agrees with the principle of relativity, so that if the law of conservation of momentum holds in one frame of reference, it also is obeyed in every other frame. A proof is given on page 697 that such an equation is

$$p = m\gamma v \quad , \quad \text{[relativistic equation for momentum]}$$

which differs from the nonrelativistic version only by the factor of γ. At low velocities γ is very close to 1, so $p = mv$ is approximately true, in agreement with the correspondence principle. At velocities close to the speed of light, γ approaches infinity, and so an object would need infinite momentum to reach the speed of light.

Force

What happens if you keep applying a constant force to an object, causing it to accelerate at a constant rate until it exceeds the speed of light? The hidden assumption here is that Newton's second law, $\mathbf{a} = \mathbf{F}/m$, is still true. It isn't. Experiments show that at speeds comparable to the speed of light, $\mathbf{a} = \mathbf{F}/m$ is wrong. The equation that still is true is

$$\mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} \quad .$$

You could apply a constant force to an object forever, increasing its momentum at a steady rate, but as the momentum approached infinity, the velocity would approach the speed of light. In general, a force produces an acceleration significantly less than $\mathbf{F}/m$ at relativistic speeds.

Would passengers on a spaceship moving close to the speed of light perceive every object as being more difficult to accelerate, as if it was more massive? No, because then they would be able to detect a change in the laws of physics because of their state of motion, which would violate the principle of relativity. The way out of this difficulty is to realize that force is not an invariant. What the passengers perceive as a small force causing a small change in momentum would look to a person in the earth's frame of reference like a large force causing a large change in momentum. As a practical matter, conservation laws are usually more convenient tools for relativistic problem solving than procedures based on the force concept.

7.3.4 Kinetic energy

Since kinetic energy equals $\frac{1}{2}mv^2$, wouldn't a sufficient amount of energy cause v to exceed the speed of light? You're on to my methods by now, so you know this is motivation for a redefinition of kinetic energy. The work-kinetic energy theorem is derived on page 697 using the correct relativistic treatment of force. The result is

$$K = m(\gamma - 1) \qquad . \qquad \text{[relativistic kinetic energy]}$$

Since γ approaches infinity as velocity approaches the speed of light, an infinite amount of energy would be required in order to make an object move at the speed of light.

Kinetic energy in unnatural units *example 7*

How can this equation be converted back into units in which the speed of light does not equal one? One approach would be to redo the derivation on page 697 in unnatural units. A far simpler approach is simply to add factors of c where necessary to make the metric units look consistent. Suppose we decide to modify the right side in order to make its units consistent with the energy units on the left. The ordinary nonrelativistic definition of kinetic energy as $(1/2)mv^2$ shows that the units on the left are

$$\text{kg}\cdot\frac{\text{m}^2}{\text{s}^2} \qquad .$$

The factor of $\gamma - 1$ is unitless, so the mass units on the right need to be multiplied by m^2/s^2 to agree with the left. This means that we need to multiply the right side by c^2:

$$K = mc^2(\gamma - 1)$$

This is beginning to resemble the famous $E = mc^2$ equation, which we will soon attack head-on.

The correspondence principle for kinetic energy *example 8*

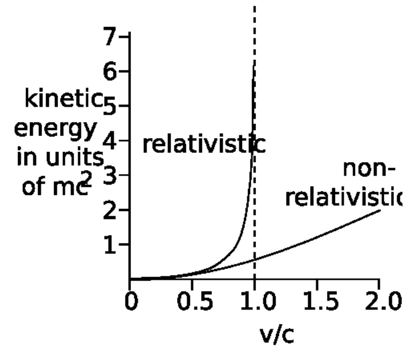
It's far from obvious that this result, even in its metric-unit form, reduces to the familiar $(1/2)mv^2$ at low speeds, as required by the correspondence principle. To show this, we need to find a low-velocity approximation for γ. In metric units, the equation for γ reads as

$$\gamma = \frac{1}{\sqrt{1-v^2/c^2}} \qquad .$$

Reexpressing this as $\left(1-v^2/c^2\right)^{-1/2}$, and making use of the approximation $(1+\epsilon)^p \approx 1+p\epsilon$ for small ϵ, the equation for gamma becomes

$$\gamma \approx 1 + \frac{v^2}{2c^2} \qquad ,$$

which can readily be used to show $mc^2(\gamma-1) \approx (1/2)mv^2$.



a / A comparison of the relativistic and nonrelativistic expressions for kinetic energy.

The large hadron collider *example 9*

▷ The Large Hadron[3] Collider (LHC), being built in Switzerland, is a ring with a radius of 4.3 km, designed to accelerate two counterrotating beams of protons to energies of 1.1×10^{-6} J per proton. (A microjoule is quite a healthy energy for a subatomic particle!) The ring has to be so big because the inward force from the accelerator's magnets would not be great enough to make the protons curve more tightly at top speed.

(a) What inward force must be exerted on each proton?

(b) In a purely Newtonian world where there were no relativistic effects, how much smaller could the LHC be if it was to produce proton beams moving at speeds close to the speed of light?

▷ (a) Since the protons have velocity vectors with constant magnitudes, γ is constant, so let's start by computing it. We'll work the whole problem in SI units, since none of the data are given in natural units. Looking up the mass of a proton, we have

$$\begin{aligned} mc^2 &= (1.7\times10^{-27}\ \text{kg})(3.0\times10^{8}\ \text{m/s})^2 \\ &= 1.5\times10^{-10}\ \text{J} \qquad . \end{aligned}$$

The kinetic energy is thousands of times greater than mc^2, so the protons go very close to the speed of light. Under these conditions there is no significant difference between γ and $\gamma-1$, so

$$\begin{aligned} \gamma &\approx K/mc^2 \\ &= 7.3\times10^{3} \end{aligned}$$

We analyze the circular motion in the laboratory frame of reference, since that is the frame of reference in which the LHC's magnets sit, and their fields were calibrated by instruments at rest with respect to them. The inward force required is

$$\begin{aligned} \mathbf{F} &= \mathrm{d}\mathbf{p}/\mathrm{d}t \\ &= \mathrm{d}(m\gamma\mathbf{v})/\mathrm{d}t \\ &= m\gamma\mathrm{d}\mathbf{v}/\mathrm{d}t \\ &= m\gamma\mathbf{a} \qquad . \end{aligned}$$

b / The Large Hadron Collider. The red circle shows the location of the underground tunnel which the LHC will share with a preexisting accelerator.

[3] "Hadron" refers to particles like protons and neutrons, which participate in nuclear forces.

Except for the factor of γ, this is the same result we would have had in Newtonian physics, where we already know the equation $a = v^2/r$ for the inward acceleration in uniform circular motion. Since the velocity is essentially the speed of light, we have $a = c^2/r$. The force required is

$$\begin{aligned} F &= m\gamma c^2/r \\ &= K/r \qquad . \qquad [\text{since } \gamma \approx \gamma - 1] \end{aligned}$$

This looks a little funny, but the units check out, since a joule is the same as a newton-meter. The result is

$$F = 2.6 \times 10^{-10} \text{ N}$$

(b) In a Newtonian universe,

$$\begin{aligned} F &= mv^2/r \\ &= mc^2/r \\ r &= mc^2/F \\ &= 0.59 \text{ m} \end{aligned}$$

In a nonrelativistic world, it would be a table-top accelerator! The energies and momenta, however, would be smaller.

7.3.5 Equivalence of mass and energy

The treatment of relativity so far has been purely mechanical, so the only form of energy we have discussed is kinetic. For example, the storyline for the introduction of relativistic momentum was based on collisions in which no kinetic energy was converted to other forms. We know, however, that collisions can result in the production of heat, which is a form of kinetic energy at the molecular level, or the conversion of kinetic energy into entirely different forms of energy, such as light or electrical energy.

Let's consider what happens if a blob of putty moving at velocity v hits another blob that is initially at rest, sticking to it, and as much kinetic energy as possible is converted into heat. (It is not possible for all the kinetic energy to be converted to heat, because then conservation of momentum would be violated.) The nonrelativistic result is that to obey conservation of momentum the two blobs must fly off together at $v/2$.

Relativistically, however, an interesting thing happens. A hot object has more momentum than a cold object! This is because the relativistically correct expression for momentum is $p = m\gamma v$, and the more rapidly moving molecules in the hot object have higher values of γ. There is no such effect in nonrelativistic physics, because the velocities of the moving molecules are all in random directions, so the random motion's contribution to momentum cancels out.

In our collision, the final combined blob must therefore be moving a little more slowly than the expected $v/2$, since otherwise the final momentum would have been a little greater than the initial momentum. To an observer who believes in conservation of momentum and knows only about the overall motion of the objects and not about their heat content, the low velocity after the collision would seem to be the result of a magical change in the mass, as if the mass of two combined, hot blobs of putty was more than the sum of their individual masses.

Heat energy is equivalent to mass.

Now we know that mass is invariant, and no molecules were created or destroyed, so the masses of all the molecules must be the same as they always were. The change is due to the change in γ with heating, not to a change in m. But how much does the mass *appear* to change? On page 697 we prove that the perceived change in mass exactly equals the change in heat energy between two temperatures, i.e. changing the heat energy by an amount E changes the effective mass of an object by E as well. This looks a bit odd because the natural units of energy and mass are the same. Converting back to ordinary units by our usual shortcut of introducing factors of c, we find that changing the heat energy by an amount E causes the apparent mass to change by $m = E/c^2$. Rearranging, we have the famous $E = mc^2$.

All energy is equivalent to mass.

But this whole argument was based on the fact that heat is a form of kinetic energy at the molecular level. Would $E = mc^2$ apply to other forms of energy as well? Suppose a rocket ship contains some electrical energy stored in a battery. If we believed that $E = mc^2$ applied to forms of kinetic energy but not to electrical energy, then we would have to believe that the pilot of the rocket could slow the ship down by using the battery to run a heater! This would not only be strange, but it would violate the principle of relativity, because the result of the experiment would be different depending on whether the ship was at rest or not. The only logical conclusion is that all forms of energy are equivalent to mass. Running the heater then has no effect on the motion of the ship, because the total energy in the ship was unchanged; one form of energy was simply converted to another.

A rusting nail — *example 10*

▷ A 50-gram iron nail is left in a cup of water until it turns entirely to rust. The energy released is about 0.5 MJ (megajoules). In theory, would a sufficiently precise scale register a change in mass? If so, how much?

▷ The energy will appear as heat, which will be lost to the environment. So the total mass-energy of the cup, water, and iron will indeed be lessened by 0.5 MJ. (If it had been perfectly insulated, there would have been no change, since the heat energy would have been trapped in the cup.) Converting to mass units, we have

$$\begin{aligned} m &= E/c^2 \\ &= (0.5 \times 10^6 \text{ J}) \ / \ (3.0 \times 10^8 \text{ m/s})^2 \\ &= 6 \times 10^{-12} \text{ J/(m}^2\text{/s}^2) \\ &= 6 \times 10^{-12} \ \left(\text{kg}\cdot(\text{m}^2/\text{s}^2)\right) \ / \ (\text{m}^2/\text{s}^2) \\ &= 6 \times 10^{-12} \text{ kg} \qquad , \end{aligned}$$

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

Energy participates in gravitational forces.

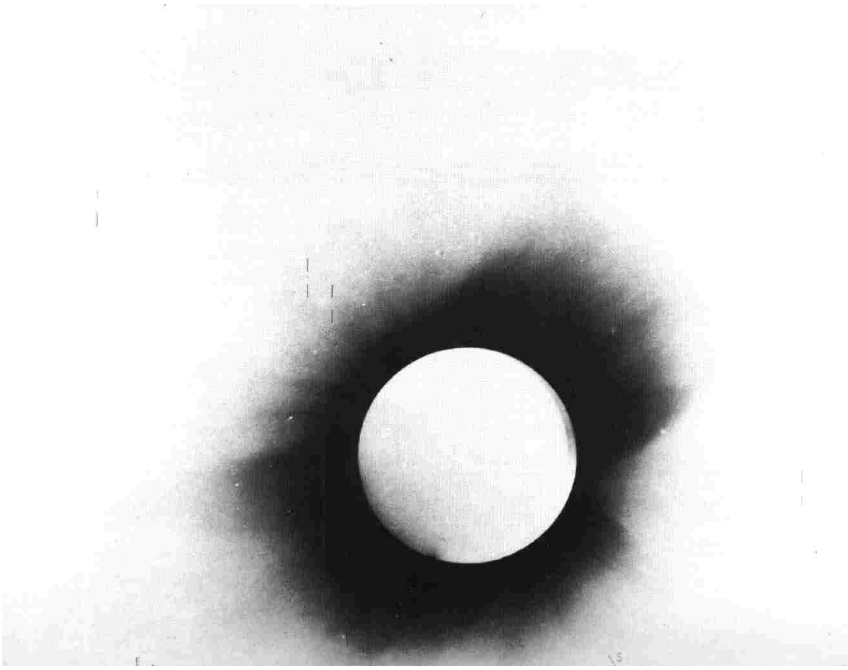
In the example we tacitly assumed that the increase in mass would show up on a scale, i.e. that its gravitational attraction with the earth would increase. Strictly speaking, however, we have only proved that energy relates to inertial mass, i.e. to phenomena like momentum and the resistance of an object to a change in its state of motion. Even before Einstein, however, experiments had shown to a high degree of precision that any two objects with the same inertial mass will also exhibit the same gravitational attractions, i.e. have the same gravitational mass. For example, the only reason that all objects fall with the same acceleration is that a more massive object's inertia is exactly in proportion to the greater gravitational forces in which it participates. We therefore conclude that energy participates in gravitational forces in the same way mass does. The total gravitational attraction between two objects is proportional not just to the product of their masses, $m_1 m_2$, as in Newton's law of gravity, but to the quantity $(m_1+E_1)(m_2+E_2)$. (Even this modification does not give a complete, self-consistent theory of gravity, which is only accomplished through the general theory of relativity.)

Gravity bending light *example 11*

Mass and energy are equivalent. The energy of a beam of light is equivalent to a certain amount of mass, and the beam is therefore deflected by a gravitational field. Einsteins prediction of this effect was verified in 1919 by astronomers who photographed stars in the dark sky surrounding the sun during an eclipse. (If there was no eclipse, the glare of the sun would prevent the stars from being observed.) Figure c is a photographic negative, so the circle that appears bright is actually the dark face of the moon, and the dark area is really the bright corona of the sun. The stars, marked by lines above and below them, appeared at positions slightly different than their normal ones, indicating that their light had been bent by the suns gravity on its way to our planet.

Black holes *example 12*

A star with sufficiently strong gravity can prevent light from leaving. Quite a few black holes have been detected via their gravitational forces on neighboring stars or clouds of dust.



c / Example 11.

Creation and destruction of particles

Since mass and energy are beginning to look like two sides of the same coin, it may not be so surprising that nature displays processes in which particles are actually destroyed or created; energy and mass are then converted back and forth on a wholesale basis. This means that in relativity there are no separate laws of conservation of energy and conservation of mass. There is only a law of conservation of mass plus energy (referred to as mass-energy). In natural units, $E+m$ is conserved, while in ordinary units the conserved quantity is $E+mc^2$.

Electron-positron annihilation — *example 13*

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

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

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

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

Note that the idea of mass as an invariant is separate from the idea that mass is not separately conserved. Invariance is the statement that all observers agree on a particle's mass regardless of their motion relative to the particle. Mass may be created or destroyed if particles are created or destroyed, and in such a situation mass invariance simply says that all observers will agree on how much mass was created or destroyed.

This chapter is summarized on page 730. Notation and terminology are tabulated on pages 718-719.

Problems

The symbols ✓, ■, etc. are explained on page 345.

1 As of 2006, the best atomic clocks have accuracies of about one part in 10^{15}. How does this compare with the time dilation effect produced if the clock takes a trip aboard a jet moving at 300 m/s? Would the effect be measurable? ▷ Hint, p. 705 ■

2 (a) Find an expression for v in terms of γ
(b) Using your result from part a, show that for very large values of γ, v gets close to the speed of light. ✓ ■

3 (a) Reexpress the Lorentz transformation equations using ordinary metric units where $c \neq 1$. The point here is to practice the technique for converting *any* formula from natural units to metric units, by inserting factors of c wherever necessary in order to make the units make sense, as in the examples 2 and 3 on page 319. That means you shouldn't go back and redo the whole derivation from scratch. ✓
(b) Show that for speeds that are small compared to the speed of light, these equations are identical to the Galilean equations. ■

4 (a) Make up a numerical example of two events, and show that the if we defined the spacetime interval as $\Delta x^2 + \Delta y^2 + \Delta z^2 + \Delta t^2$, we would not get consistent results when we Lorentz-transformed the events into a different frame of reference.
(b) Show that, for the particular example you chose in part a, the quantity $\Delta x^2 + \Delta y^2 + \Delta z^2 - \Delta t^2$ does come out the same in both frames.
(c) Ignoring the y and z space dimensions, prove that $\Delta x^2 - \Delta t^2$ stays the same under a Lorentz transformation for motion along the x axis. You're proving this in general now, not just checking it for one numerical example.
(d) Reexpress the definition $\Delta x^2 + \Delta y^2 + \Delta z^2 - \Delta t^2$ of the spacetime interval in unnatural units, where $c \neq 1$. ■

5 Make up a numerical example, in a particular frame of reference, of two events with a spacelike interval between them. Make event 2 occur after event 1. Now show by using a Lorentz transformation that you can find another frame of reference in which event 2 occurs before event 1.

Remark: To get from event 1 to event 2, or vice versa, you would have to travel faster than light. Therefore there can't be a cause-and-effect relationship between the two events, and it doesn't really matter which one we consider to have happened first. On the other hand, if faster-than-light travel was possible, then time travel paradoxes would be possible in this kind of situation. For example, event 2 could be your birth, and event 1 could be when you kill your own grandmother before she has any children. ■

6 (a) A spacecraft traveling at 1.0000×10^7 m/s relative to the earth releases a probe in the forward direction at a relative speed

of 2.0000×10^7 m/s. How fast is the probe moving relative to the earth? How does this compare with the nonrelativistic result? ✓
(b) Repeat the calculation, but with both velocities equal to $c/2$. How does this compare with the nonrelativistic result? ✓

7 (a) Show that when two velocities are combined relativistically, and one of them equals the speed of light, the result also equals the speed of light.
(b) Explain why it has to be this way based on the principle of relativity. (Note that it doesn't work to say that it has to be this way because motion faster than c is impossible. That isn't what the principle of relativity says, and it also doesn't handle the case where the velocities are in opposite direction.)

8 Cosmic-ray particles with relativistic velocities are continually bombarding the earth's atmosphere. They are protons and other atomic nuclei. Suppose a carbon nucleus (containing six protons and six neutrons) arrives with an energy of 10^{-7} J, which is unusually high, but not unheard of. By what factor is its length shortened as seen by an observer in the earth's frame of reference?

▷ Hint, p. 705 ✓

9 (a) A free neutron (as opposed to a neutron bound into an atomic nucleus) is unstable, and decays radioactively into a proton, an electron, and a particle called a neutrino. (This process can also occur for a neutron in a nucleus, but then other forms of mass-energy are involved as well.) The masses are as follows:

neutron	1.67495×10^{-27} kg
proton	1.67265×10^{-27} kg
electron	0.00091×10^{-27} kg
neutrino	negligible

Find the energy released in the decay of a free neutron. ✓
(b) We might imagine that a proton could decay into a neutron, a positron, and a neutrino. 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!)

10 (a) Find a relativistic equation for the velocity of an object in terms of its mass and momentum (eliminating γ). Work in natural units. ✓
(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.

11 (a) Prove the equation $E^2 - p^2 = m^2$ for a material object, where $E = m\gamma$ is the total mass-energy.
(b) Using this result, show that an object with zero mass must move at the speed of light.
(c) This equation can be applied more generally, to light for instance. Use it to find the momentum of a beam of light having energy E.

$\checkmark$

(d) Convert your answer from the previous part into ordinary units. $\checkmark$

▷ Answer, p. 712

12 Starting from the equation $v_{combined}\gamma_{combined} = (v_1 + v_2)\gamma_1\gamma_2$ derived on page 697, complete the proof of $v_{combined} = (v_1+v_2)/(1+v_1v_2)$.

13 A source of light with frequency f is moving toward an observer at velocity v (or away from the observer if v is negative). Find the relativistically correct equation for the Doppler shift of the light.

▷ Hint, p. 705

14 An antielectron collides with an electron that is at rest. (An antielectron is a form of antimatter that is just like an electron, but with the opposite charge.) The antielectron and electron annihilate each other and produce two gamma rays. (A gamma ray is a form of light. It has zero mass.) Gamma ray 1 is moving in the same direction as the antielectron was initially going, and gamma ray 2 is going in the opposite direction. Throughout this problem, you should work in natural units and use the notation E to mean the total mass-energy of a particle, i.e. its mass plus its kinetic energy. Find the energies of the two gamma-rays, E_1 and E_2, in terms of m, the mass of an electron or antielectron, and E_o, the initial mass-energy of the antielectron. You'll need the result of problem 11a.

15 (a) Use the result of problem 11d to show that if light with power P is reflected perpendicularly from a perfectly reflective surface, the force on the surface is $2P/c$.
(b) Estimate the maximum mass of a thin film that is to be levitated by a 100-watt lightbulb. ▷ Solution, p. 715

16 When an object moves at a speed extremely close to the speed of light, we refer to its motion as "ultrarelativistic." Find an approximation for the γ of an object in ultrarelativistic motion at a velocity of $(1-\epsilon)c$, where ϵ is small. This approximation can be useful in cases where ϵ is so small that your calculator would round off the expression $\sqrt{1-v^2/c^2}$ to zero, giving a $\gamma = \infty$. $\checkmark$

17 Our sun lies at a distance of 26,000 light years from the center of the galaxy, where there are some spectactular sights to see, including a supermassive black hole that is rapidly eating up the surrounding interstellar gas and dust. Rich tourist Bill Gates IV buys a spaceship, and heads for the galactic core at a speed of 99.99999% of the speed of light.
(a) According to observers on Earth, how long does it take before he gets back? (Ignore the short time he actually spends sightseeing at the core.) $\checkmark$
(b) In Bill's frame of reference, how much time passes? $\checkmark$

(c) When you compare your answer to part b with the round-trip distance, do you conclude that Bill considers himself to be moving faster than the speed of light? If so, how do you reconcile this with relativity? If not, then resolve the apparent paradox.

18 A velocity of 4/5 the speed of light results in $\gamma = 5/3$, which is a nice simple fraction: one integer divided by another. Find one or more additional examples like this (not the trivial cases $v = 0$ or $-4/5$).

19 Expand the equation $K = m(\gamma - 1)$ in a Taylor series, and find the first two nonvaninishing terms. Show that the first term is the classical expression for kinetic energy.

20 Expand the relativistic equation for momentum in a Taylor series, and find the first two nonvaninishing terms. Show that the first term is the classical expression.

21 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) Describe the motion of the other two ships according to Alice, who is aboard ship A.
(b) Give the description according to Betty, whose frame of reference is ship B.
(c) Do the same for Cathy, aboard ship C.

22 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.

23 Radiocative particle a decays, annihilating itself and producing two particles b and c, of unequal mass. Consider this process in the frame of reference in which particle a was at rest before the decay.
(a) In the special case where very little energy is released in the decay, and particles b and c have nonrelativistic speeds, prove using classical physics that the particle with the lower mass must have the higher kinetic energy.
(b) Find an expression for the mass-energy E_c of particle c, in terms of the masses m_a, m_b, and m_c. Hint: work in natural units, and make use of the result of problem 11a. √
(c) Show that the units of your answer make sense.
(d) Show that your expression has the correct behavior in the case of $m_b = m_c$.
(e) A process of this type is the decay of a K^+ particle into a π^+ and a π_0 (called pions). The masses are 493.7, 139.6, and 135.0 MeV, respectively. (MeV are a unit of energy, but in natural units, they can also be a unit of mass.) Find the mass-energies and kinetic energies of the two pions, and verify that the nonrelativistic prediction of part (a) is still correct, even in the fully relativistic case.

24 (a) Find an expression, in natural units, for the velocity of a

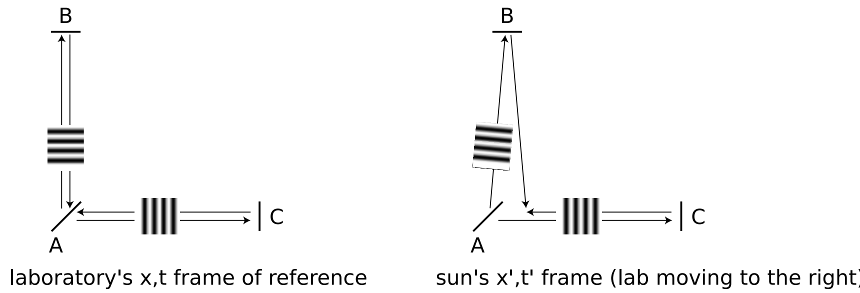
particle having mass m and mass-energy E.
(b) Show that the units of your equation make sense.
(c) Your answer involves a square root, which could be either the positive or the negative root. Explain what this represents physically, and why it makes sense.
(d) Discuss the limit of $E \gg m$, both mathematically and physically.
(e) Rewrite your expression in SI units. Don't rederive it from scratch. Simply determine how it needs to be altered by inserting factors of c in order to make the units work out in the SI.

25 In the equation for the relativistic addition of velocities u and v, consider the limit in which u approaches 1, but v simultaneously approaches -1. Give both a physical and a mathematical interpretation.

Key to symbols:
easy typical challenging difficult very difficult
$\surd$ An answer check is available at www.lightandmatter.com.

Exercises

Exercise 7A: The Michelson-Morley Experiment



In this exercise you will analyze the Michelson-Morley experiment, and find what the results should have been according to Galilean relativity and Einstein's theory of relativity. A beam of light coming from the west (not shown) comes to the half-silvered mirror A. Half the light goes through to the east, is reflected by mirror C, and comes back to A. The other half is reflected north by A, is reflected by B, and also comes back to A. When the beams reunite at A, part of each ends up going south, and these parts interfere with one another. If the time taken for a round trip differs by, for example, half the period of the wave, there will be destructive interference.

The point of the experiment was to search for a difference in the experimental results between the daytime, when the laboratory was moving west relative to the sun, and the nighttime, when the laboratory was moving east relative to the sun. Galilean relativity and Einstein's theory of relativity make different predictions about the results. According to Galilean relativity, the speed of light cannot be the same in all reference frames, so it is assumed that there is one special reference frame, perhaps the sun's, in which light travels at the same speed in all directions; in other frames, Galilean relativity predicts that the speed of light will be different in different directions, e.g. slower if the observer is chasing a beam of light. There are four different ways to analyze the experiment:

- *Laboratory's frame of reference, Galilean relativity.* This is not a useful way to analyze the experiment, since one does not know how fast light will travel in various directions.

- *Sun's frame of reference, Galilean relativity.* We assume that in this special frame of reference, the speed of light is the same in all directions: we call this speed c. In this frame, the laboratory moves with velocity v, and mirrors A, B, and C move while the light beam is in flight.

- *Laboratory's frame of reference, Einstein's theory of relativity.* The analysis is extremely simple. Let the length of each arm be L. Then the time required to get from A to either mirror is L/c, so each beam's round-trip time is $2L/c$.

- *Sun's frame of reference, Einstein's theory of relativity.* We analyze this case by starting with the laboratory's frame of reference and then transforming to the sun's frame.

Groups 1-4 work in the sun's frame of reference according to Galilean relativity.

Group 1 finds time AC. Group 2 finds time CA. Group 3 finds time AB. Group 4 finds time BA.

Groups 5 and 6 transform the lab-frame results into the sun's frame according to Einstein's theory.

Group 5 transforms the x and t when ray ACA gets back to A into the sun's frame of reference, and group 6 does the same for ray ABA.

Discussion:

Michelson and Morley found no change in the interference of the waves between day and night. Which version of relativity is consistent with their results?

What does each theory predict if v approaches c?

What if the arms are not exactly equal in length?

Does it matter if the "special" frame is some frame other than the sun's?

Exercise 7B: Sports in Slowlightland

In Slowlightland, the speed of light is 20 mi/hr $\approx$ 32 km/hr $\approx$ 9 m/s. Think of an example of how relativistic effects would work in sports. Things can get very complex very quickly, so try to think of a simple example that focuses on just one of the following effects:

- relativistic momentum
- relativistic kinetic energy
- relativistic addition of velocities
- time dilation and length contraction
- Doppler shifts of light
- equivalence of mass and energy
- time it takes for light to get to an athlete's eye
- deflection of light rays by gravity

Chapter 8
Atoms and Electromagnetism

8.1 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 fuzzyheaded 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*

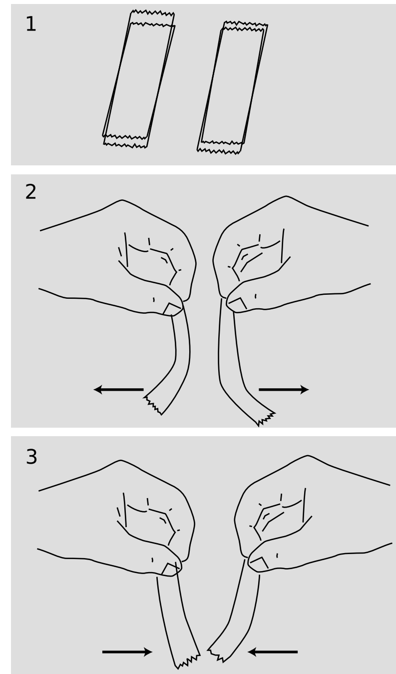
Newton's quest

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.

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:



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.

- 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 pointlike 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 pointlike, 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, p. 708

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

$$|\mathbf{F}| = k\frac{|q_1||q_2|}{r^2} \qquad ,$$

where the constant k equals $9.0\times10^9\ \text{N}\cdot\text{m}^2/\text{C}^2$. 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

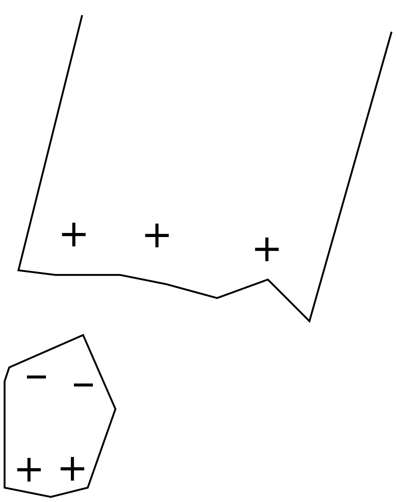
An even more fundamental reason for using positive and negative signs for electrical charge is that experiments show that charge is conserved according to this definition: in any closed system, the total amount of charge is a constant. This is why we observe that rubbing initially uncharged substances 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.

b / A charged piece of tape attracts uncharged pieces of paper from a distance, and they leap up to it.

Electrical forces involving neutral objects

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 is between the paper and the tape is now stronger than the repulsion, because the negatively charged end is closer to the tape.

Self-Check
What would have happened if the tape was negatively charged? ▷ Answer, p. 708



c / The paper has zero total charge, but it does have charged particles in it that can move.

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.

A magnetic compass — *example 1*

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.

A television tube — *example 2*

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

A 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?

B 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?

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 1899 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 elements

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.

$$\frac{m_{\text{He}}}{m_{\text{H}}} = 3.97$$

$$\frac{m_{\text{Ne}}}{m_{\text{H}}} = 20.01$$

$$\frac{m_{\text{Sc}}}{m_{\text{H}}} = 44.60$$

d / Examples of masses of atoms compared to that of hydrogen. Note how some, but not all, are close to integers.

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.

1 H																		2 He
3 Li	4 Be												5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg												13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc		22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y		40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	57 La	*	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	89 Ac	**	104 Rf	105 Ha	106	107	108	109	110	111	112	113	114	115	116	117	118

*	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
**	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

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

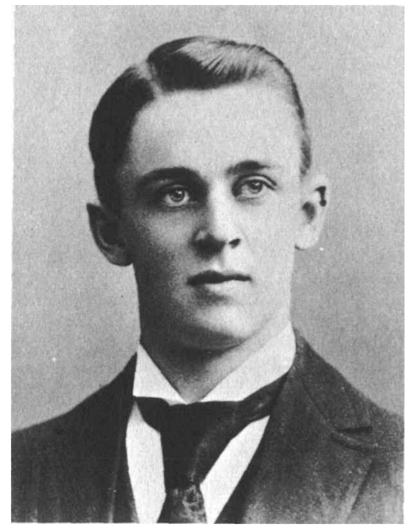
A How could knowledge of the size of an individual aluminum atom be used to infer an estimate of its mass, or vice versa?

B How could one test Einstein's interpretation of Brownian motion by observing it at different temperatures?

8.1.4 Quantization of charge

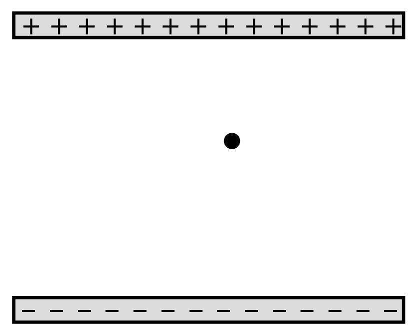
Proving that atoms actually existed was a big accomplishment, but demonstrating their existence was different from understanding their properties. Note that the Brown-Einstein observations had nothing at all to do with electricity, and yet we know that matter is inherently electrical, and we have been 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.

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.



f / A young Robert Millikan.(*Contemporary*)

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



g / A simplified diagram of Millikan's apparatus.

$$bvr^2 - mg = 0 \qquad ,$$

and setting the known density of oil equal to the drop's mass divided by its volume gives a second equation,

$$\rho = \frac{m}{\frac{4}{3}\pi r^3} \qquad .$$

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,

$$\frac{F}{q} = \text{constant} \qquad .$$

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 $/(1.64$ $\times 10^{-19}$ C)
-1.970×10^{-18}	-12.02
-0.987×10^{-18}	-6.02
-2.773×10^{-18}	-16.93

h / A few samples of Millikan's data.

The table on the left 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 $e = 1.60 \times 10^{-19}$ 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

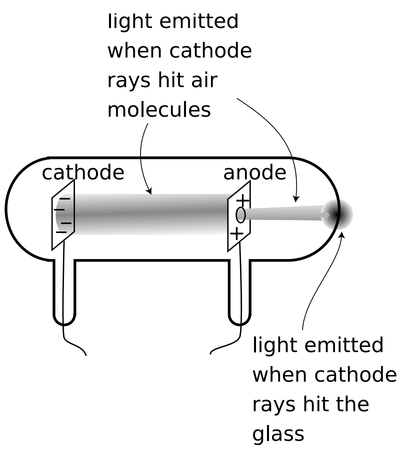
Is money quantized? What is the quantum of money? ▷ Answer, p. 708

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.

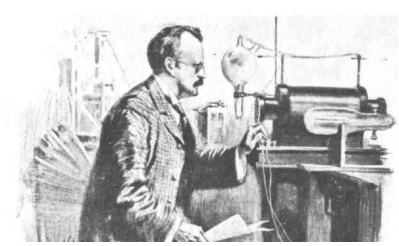
Other experiments, however, pointed to the contrary conclusion. Light is a wave phenomenon, and one distinguishing property of waves is demonstrated by speaking into one end of a paper towel roll. The sound waves do not emerge from the other end of the tube as a focused beam. Instead, they begin spreading 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.

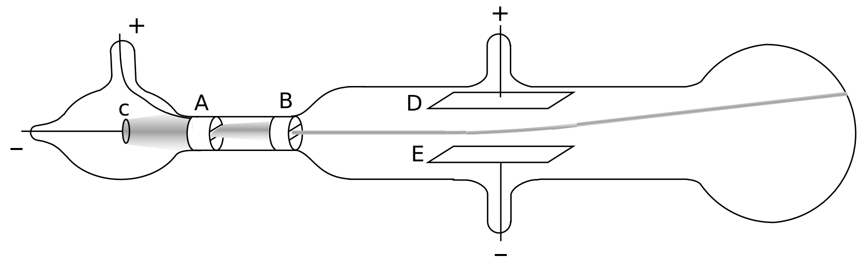
Nobody could think of a good way to weigh cathode rays, so the next most obvious way of settling the light/matter debate was to check whether the cathode rays possessed electrical charge. Light was known to be uncharged. If the cathode rays carried charge, they were definitely matter and not light, and 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

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.



j / J.J. Thomson in the lab.



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,

$$F_{elec} = (\text{known constant}) \cdot q \qquad .$$

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,

$$v = \frac{(\text{known constant})}{(\text{known constant \#2})} \qquad .$$

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 electron

What was significant about Thomson's experiment was not the actual numerical value of m/q, however, so much as the fact that, combined with Millikan's value of the fundamental charge, it gave a mass for the cathode ray particles that was thousands of times smaller than the mass of even the lightest atoms. Even without Millikan's results, which were 14 years in the future, Thomson recognized that the cathode rays' m/q was thousands of times smaller than the m/q ratios that had been measured for electrically charged atoms in chemical solutions. He correctly interpreted this as evidence that the cathode rays were smaller building blocks — he called them *electrons* — out of which atoms themselves were formed. This was an extremely radical claim, coming at a time when atoms had not yet been proven to exist! Even those who used the word "atom" often considered them no more than mathematical abstractions, not literal objects. The idea of searching for structure inside of "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

A Thomson started to become convinced during his experiments that the "cathode rays" observed coming from the cathodes of vacuum tubes were building blocks of atoms — what we now call electrons. He then carried out observations with cathodes made of a variety of metals, and found that m/q was roughly the same in every case, considering his limited accuracy. Given his suspicion, why did it make sense to try different metals? How would the consistent values of m/q serve to test his hypothesis?

B My students have frequently asked whether the m/q that Thomson measured was the value for a single electron, or for the whole beam. Can you answer this question?

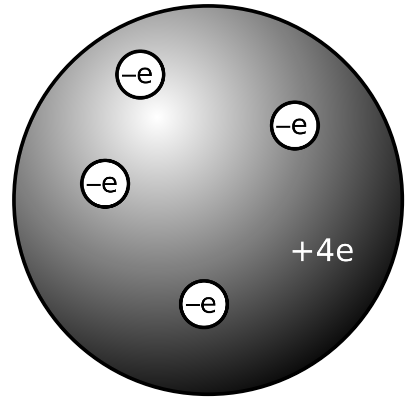
C Thomson found that the m/q of an electron was thousands of times smaller than that of charged atoms in chemical solutions. Would this imply that the electrons had more charge? Less mass? Would there be no way to tell? Explain. Remember that Millikan's results were still many years in the future, so q was unknown.

D Can you guess any practical reason why Thomson couldn't just let one electron fly across the gap before disconnecting the battery and turning off the beam, and then measure the amount of charge deposited on the anode, thus allowing him to measure the charge of a single electron directly?

E Why is it not possible to determine m and q themselves, rather than just their ratio, by observing electrons' motion in electric and magnetic fields?

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.

Flow of electrical charge in wires — *example 3*

One of my former students was the son of an electrician, and had become an electrician himself. He related to me how his father had remained refused to believe all his life that electrons really flowed through wires. If they had, he reasoned, the metal would have gradually become more and more damaged, eventually crumbling to dust.

His opinion is not at all unreasonable based on the fact that electrons are material particles, and that matter cannot normally pass through matter without making a hole through it. Nineteenth-century physicists would have shared his objection to a charged-particle model of the flow of electrical charge. In the raisin-cookie model, however, the electrons are very low in mass, and therefore presumably very small in size as well. It is not surprising that they can slip between the atoms without damaging them.

Flow of electrical charge across cell membranes *example 4*
Your nervous system is based on signals carried by charge moving from nerve cell to nerve cell. Your body is essentially all liquid, and atoms in a liquid are mobile. This means that, unlike the case of charge flowing in a solid wire, entire charged atoms can flow in your nervous system

Emission of electrons in a cathode ray tube *example 5*
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.)

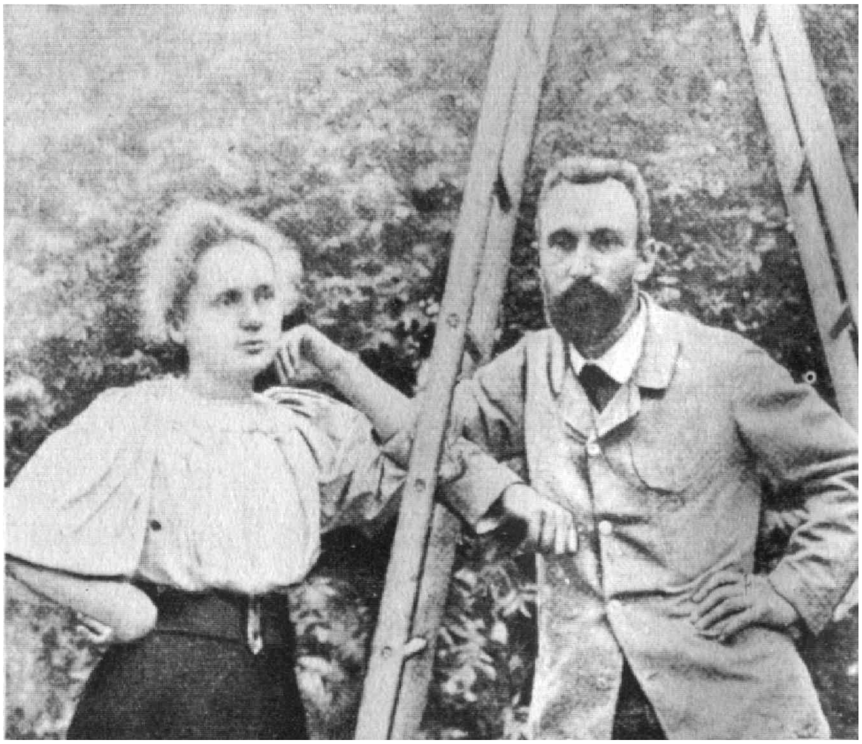
This interpretation explains the warm and friendly yellow glow of the vacuum tubes in an antique radio. To encourage the emission of electrons from the vacuum tubes' cathodes, the cathodes are intentionally warmed up with little heater coils.

Discussion Questions

A Today many people would define an ion as an atom (or molecule) with missing electrons or extra electrons added on. How would people have defined the word "ion" before the discovery of the electron?

B Since electrically neutral atoms were known to exist, there had to be positively charged subatomic stuff to cancel out the negatively charged electrons in an atom. Based on the state of knowledge immediately after the Millikan and Thomson experiments, was it possible that the positively charged stuff had an unquantized amount of charge? Could it be quantized in units of +e? In units of +2e? In units of +5/7e?

8.2 The Nucleus



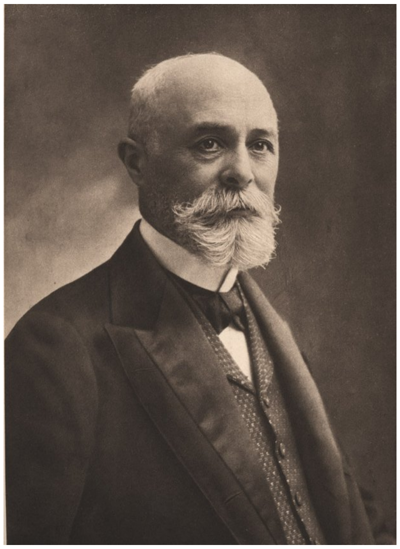
a / Marie and Pierre Curie were the first to purify radium in significant quantities. Radium's intense radioactivity made possible the experiments that led to the modern planetary model of the atom, in which electrons orbit a nucleus made of protons and neutrons.

8.2.1 Radioactivity

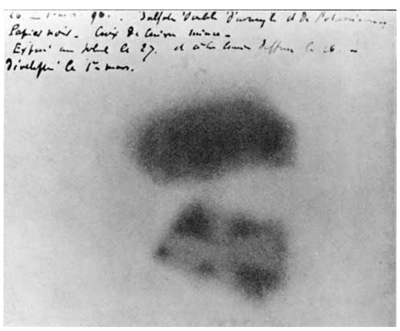
Becquerel's discovery of radioactivity

How did physicists figure out that the raisin cookie model was incorrect, and that the atom's positive charge was concentrated in a tiny, central nucleus? The story begins with the discovery of radioactivity by the French chemist Becquerel. Up until radioactivity was discovered, all the processes of nature were thought to be based on chemical reactions, which were rearrangements of combinations of atoms. Atoms exert forces on each other when they are close together, so sticking or unsticking them would either release or store electrical energy. That energy could be converted to and from other forms, as when a plant uses the energy in sunlight to make sugars and carbohydrates, or when a child eats sugar, releasing the energy in the form of kinetic energy.

Becquerel discovered a process that seemed to release energy from an unknown new source that was not chemical. Becquerel, whose father and grandfather had also been physicists, spent the first twenty years of his professional life as a successful civil engineer, teaching physics on a part-time basis. He was awarded the chair of physics at the Musée d'Histoire Naturelle in Paris after the death of his father, who had previously occupied it. Having now a significant amount of time to devote to physics, he began studying the interaction of light and matter. He became interested in the phenomenon of phosphorescence, in which a substance absorbs energy from light, then releases the energy via a glow that only gradually goes away. One of the substances he investigated was a uranium compound, the salt $UKSO_5$. One day in 1896, cloudy weather interfered with his plan to expose this substance to sunlight in order to observe its fluorescence. He stuck it in a drawer, coincidentally on top of a blank photographic plate — the old-fashioned glass-backed counterpart of the modern plastic roll of film. The plate had been carefully wrapped, but several days later when Becquerel checked it in the darkroom before using it, he found that it was ruined, as if it had been completely exposed to light.



b / Henri Becquerel (1852-1908).



c / Becquerel's photographic plate. In the exposure at the bottom of the image, he has found that he could absorb the radiations, casting the shadow of a Maltese cross that was placed between the plate and the uranium salts.

History provides many examples of scientific discoveries that happened this way: an alert and inquisitive mind decides to investigate a phenomenon that most people would not have worried about explaining. Becquerel first determined by further experiments that the effect was produced by the uranium salt, despite a thick wrapping of paper around the plate that blocked out all light. He tried a variety of compounds, and found that it was the uranium that did it: the effect was produced by any uranium compound, but not by any compound that didn't include uranium atoms. The effect could be at least partially blocked by a sufficient thickness of metal, and he was able to produce silhouettes of coins by interposing them between the uranium and the plate. This indicated that the effect traveled in a straight line., so that it must have been some kind of ray rather than, e.g., the seepage of chemicals through the paper. He used the word "radiations," since the effect radiated out from the uranium salt.

At this point Becquerel still believed that the uranium atoms were absorbing energy from light and then gradually releasing the energy in the form of the mysterious rays, and this was how he presented it in his first published lecture describing his experiments. Interesting, but not earth-shattering. But he then tried to determine how long it took for the uranium to use up all the energy that had supposedly been stored in it by light, and he found that it never seemed to become inactive, no matter how long he waited. Not only that, but a sample that had been exposed to intense sunlight for a whole afternoon was no more or less effective than a sample that had always been kept inside. Was this a violation of conservation of energy? If the energy didn't come from exposure to light, where did it come from?

Three kinds of "radiations"

Unable to determine the source of the energy directly, turn-of-the-century physicists instead studied the behavior of the "radiations" once they had been emitted. Becquerel had already shown that the radioactivity could penetrate through cloth and paper, so the first obvious thing to do was to investigate in more detail what thickness of material the radioactivity could get through. They soon learned that a certain fraction of the radioactivity's intensity would be eliminated by even a few inches of air, but the remainder was not eliminated by passing through more air. Apparently, then, the radioactivity was a mixture of more than one type, of which one was blocked by air. They then found that of the part that could penetrate air, a further fraction could be eliminated by a piece of paper or a very thin metal foil. What was left after that, however, was a third, extremely penetrating type, some of whose intensity would still remain even after passing through a brick wall. They decided that this showed there were three types of radioactivity, and without having the faintest idea of what they really were, they made up names for them. The least penetrating type was arbitrarily labeled α (alpha), the first letter of the Greek alphabet, and so on through β (beta) and finally γ (gamma) for the most penetrating type.

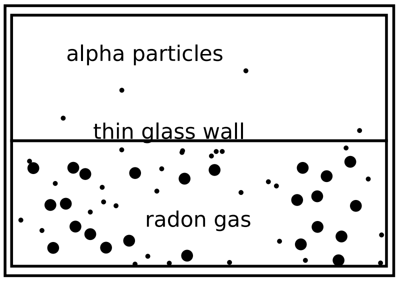
Radium: a more intense source of radioactivity

The measuring devices used to detect radioactivity were crude: photographic plates or even human eyeballs (radioactivity makes flashes of light in the jelly-like fluid inside the eye, which can be seen by the eyeball's owner if it is otherwise very dark). Because the ways of detecting radioactivity were so crude and insensitive, further progress was hindered by the fact that the amount of radioactivity emitted by uranium was not really very great. The vital contribution of physicist/chemist Marie Curie and her husband Pierre was to discover the element radium, and to purify and isolate significant quantities it. Radium emits about a million times more radioactivity per unit mass than uranium, making it possible to do the experiments that were needed to learn the true nature of radioactivity. The dangers of radioactivity to human health were then unknown, and Marie died of leukemia thirty years later. (Pierre was run over and killed by a horsecart.)

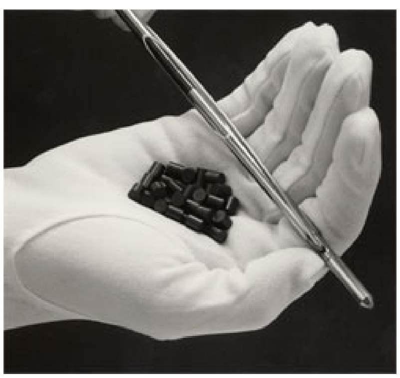
Tracking down the nature of alphas, betas, and gammas

As radium was becoming available, an apprentice scientist named Ernest Rutherford arrived in England from his native New Zealand and began studying radioactivity at the Cavendish Laboratory. The young colonial's first success was to measure the mass-to-charge ratio of beta rays. The technique was essentially the same as the one Thomson had used to measure the mass-to-charge ratio of cathode rays by measuring their deflections in electric and magnetic fields. The only difference was that instead of the cathode of a vacuum tube, a nugget of radium was used to supply the beta rays. Not only was the technique the same, but so was the result. Beta rays had the same m/q ratio as cathode rays, which suggested they were one and the same. Nowadays, it would make sense simply to use the term "electron," and avoid the archaic "cathode ray" and "beta particle," but the old labels are still widely used, and it is unfortunately necessary for physics students to memorize all three names for the same thing.

At first, it seemed that neither alphas or gammas could be deflected in electric or magnetic fields, making it appear that neither was electrically charged. But soon Rutherford obtained a much more powerful magnet, and was able to use it to deflect the alphas but not the gammas. The alphas had a much larger value of m/q than the betas (about 4000 times greater), which was why they had been so hard to deflect. Gammas are uncharged, and were later found to be a form of light.



d / A simplified version of Rutherford's 1908 experiment, showing that alpha particles were doubly ionized helium atoms.



e / These pellets of uranium fuel will be inserted into the metal fuel rod and used in a nuclear reactor. The pellets emit alpha and beta radiation, which the gloves are thick enough to stop.

The m/q ratio of alpha particles turned out to be the same as those of two different types of ions, He^{++} (a helium atom with two missing electrons) and H_2^+ (two hydrogen atoms bonded into a molecule, with one electron missing) , so it seemed likely that they were one or the other of those. The diagram shows a simplified version of Rutherford's ingenious experiment proving that they were He^{++} ions. The gaseous element radon, an alpha emitter, was introduced into one half of a double glass chamber. The glass wall dividing the chamber was made extremely thin, so that some of the rapidly moving alpha particles were able to penetrate it. The other chamber, which was initially evacuated, gradually began to accumulate a population of alpha particles (which would quickly pick up electrons from their surroundings and become electrically neutral). Rutherford then determined that it was helium gas that had appeared in the second chamber. Thus alpha particles were proved to be He^{++} ions. The nucleus was yet to be discovered, but in modern terms, we would describe a He^{++} ion as the nucleus of a He atom.

To summarize, here are the three types of radiation emitted by radioactive elements, and their descriptions in modern terms:

α particle	stopped by a few inches of air	He nucleus
β particle	stopped by a piece of paper	electron
γ ray	penetrates thick shielding	a type of light

Discussion Questions

A Most sources of radioactivity emit alphas, betas, and gammas, not just one of the three. In the radon experiment, how did Rutherford know that he was studying the alphas?

8.2.2 The planetary model of the atom

The stage was now set for the unexpected discovery that the positively charged part of the atom was a tiny, dense lump at the atom's center rather than the "cookie dough" of the raisin cookie model. By 1909, Rutherford was an established professor, and had students working under him. For a raw undergraduate named Marsden, he picked a research project he thought would be tedious but straightforward.

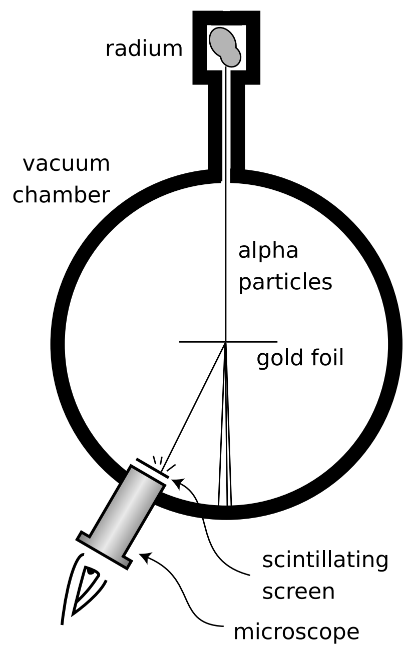
It was already known that although alpha particles would be stopped completely by a sheet of paper, they could pass through a sufficiently thin metal foil. Marsden was to work with a gold foil only 1000 atoms thick. (The foil was probably made by evaporating a little gold in a vacuum chamber so that a thin layer would be deposited on a glass microscope slide. The foil would then be lifted off the slide by submerging the slide in water.)

Rutherford had already determined in his previous experiments the speed of the alpha particles emitted by radium, a fantastic 1.5×10^7 m/s. The experimenters in Rutherford's group visualized them as very small, very fast cannonballs penetrating the "cookie dough" part of the big gold atoms. A piece of paper has a thickness of a hundred thousand atoms or so, which would be sufficient to stop them completely, but crashing through a thousand would only slow them a little and turn them slightly off of their original paths.

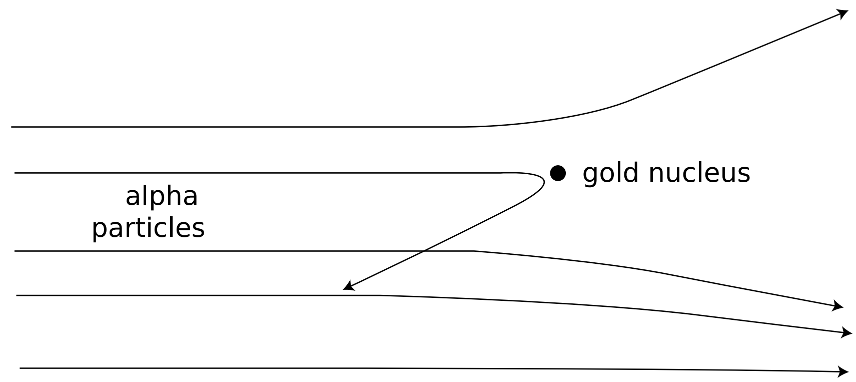
Marsden's supposedly ho-hum assignment was to use the apparatus shown in figure g to measure how often alpha particles were deflected at various angles. A tiny lump of radium in a box emitted alpha particles, and a thin beam was created by blocking all the alphas except those that happened to pass out through a tube. Typically deflected in the gold by only a small amount, they would reach a screen very much like the screen of a TV's picture tube, which would make a flash of light when it was hit. Here is the first example we have encountered of an experiment in which a beam of particles is detected one at a time. This was possible because each alpha particle carried so much kinetic energy; they were moving at about the same speed as the electrons in the Thomson experiment, but had ten thousand times more mass.

Marsden sat in a dark room, watching the apparatus hour after hour and recording the number of flashes with the screen moved to various angles. The rate of the flashes was highest when he set the screen at an angle close to the line of the alphas' original path, but if he watched an area farther off to the side, he would also occasionally see an alpha that had been deflected through a larger angle. After seeing a few of these, he got the crazy idea of moving the screen to see if even larger angles ever occurred, perhaps even angles larger than 90 degrees.

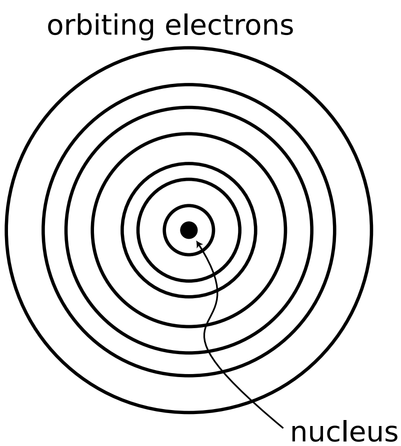
f / Ernest Rutherford (1871-1937).



g / Rutherford and Marsden's apparatus.



h / Alpha particles being scattered by a gold nucleus. On this scale, the gold atom is the size of a car, so all the alpha particles shown here are ones that just happened to come unusually close to the nucleus. For these exceptional alpha particles, the forces from the electrons are unimportant, because they are so much more distant than the nucleus.



i / The planetary model of the atom.

The crazy idea worked: a few alpha particles were deflected through angles of up to 180 degrees, and the routine experiment had become an epoch-making one. Rutherford said, "We have been able to get some of the alpha particles coming backwards. It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you." Explanations were hard to come by in the raisin cookie model. What intense electrical forces could have caused some of the alpha particles, moving at such astronomical speeds, to change direction so drastically? Since each gold atom was electrically neutral, it would not exert much force on an alpha particle outside it. True, if the alpha particle was very near to or inside of a particular atom, then the forces would not necessarily cancel out perfectly; if the alpha particle happened to come very close to a particular electron, the $1/r^2$ form of the Coulomb force law would make for a very strong force. But Marsden and Rutherford knew that an alpha particle was 8000 times more massive than an electron, and it is simply not possible for a more massive object to rebound backwards from a collision with a less massive object while conserving momentum and energy. It might be possible in principle for a particular alpha to follow a path that took it very close to one electron, and then very close to another electron, and so on, with the net result of a large deflection, but careful calculations showed that such multiple "close encounters" with electrons would be millions of times too rare to explain what was actually observed.

At this point, Rutherford and Marsden dusted off an unpopular and neglected model of the atom, in which all the electrons orbited around a small, positively charged core or "nucleus," just like the planets orbiting around the sun. All the positive charge and nearly all the mass of the atom would be concentrated in the nucleus, rather than spread throughout the atom as in the raisin cookie model. The positively charged alpha particles would be repelled by the gold atom's nucleus, but most of the alphas would not come close enough to any nucleus to have their paths drastically altered. The few that did come close to a nucleus, however, could rebound backwards from a single such encounter, since the nucleus of a heavy gold atom would be fifty times more massive than an alpha particle. It turned out that it was not even too difficult to derive a formula giving the relative frequency of deflections through various angles, and this calculation agreed with the data well enough (to within 15%), considering the difficulty in getting good experimental statistics on the rare, very large angles.

What had started out as a tedious exercise to get a student started in science had ended as a revolution in our understanding of nature. Indeed, the whole thing may sound a little too much like a moralistic fable of the scientific method with overtones of the Horatio Alger genre. The skeptical reader may wonder why the planetary model was ignored so thoroughly until Marsden and Rutherford's discovery. Is science really more of a sociological enterprise, in which certain ideas become accepted by the establishment, and other, equally plausible explanations are arbitrarily discarded? Some social scientists are currently ruffling a lot of scientists' feathers with critiques very much like this, but in this particular case, there were very sound reasons for rejecting the planetary model. As you'll learn in more detail later in this course, any charged particle that undergoes an acceleration dissipate energy in the form of light. In the planetary model, the electrons were orbiting the nucleus in circles or ellipses, which meant they were undergoing acceleration, just like the acceleration you feel in a car going around a curve. They should have dissipated energy as light, and eventually they should have lost all their energy. Atoms don't spontaneously collapse like that, which was why the raisin cookie model, with its stationary electrons, was originally preferred. There were other problems as well. In the planetary model, the one-electron atom would have to be flat, which would be inconsistent with the success of molecular modeling with spherical balls representing hydrogen and atoms. These molecular models also seemed to work best if specific sizes were used for different atoms, but there is no obvious reason in the planetary model why the radius of an electron's orbit should be a fixed number. In view of the conclusive Marsden-Rutherford results, however, these became fresh puzzles in atomic physics, not reasons for disbelieving the planetary model.

j / The planetary model applied to a nonmetal, 1, an unmagnetized metal, 2, and a magnetized metal, 3. Note that these figures are all simplified in several ways. For one thing, the electrons of an individual atom do not all revolve around the nucleus in the same plane. It is also very unusual for a metal to become so strongly magnetized that 100% of its atoms have their rotations aligned as shown in this figure.

Some phenomena explained with the planetary model

The planetary model may not be the ultimate, perfect model of the atom, but don't underestimate its power. It already allows us to visualize correctly a great many phenomena.

As an example, let's consider the distinctions among nonmetals, metals that are magnetic, and metals that are nonmagnetic. As shown in figure j, a metal differs from a nonmetal because its outermost electrons are free to wander rather than owing their allegiance to a particular atom. A metal that can be magnetized is one that is willing to line up the rotations of some of its electrons so that their axes are parallel. Recall that magnetic forces are forces made by moving charges; we have not yet discussed the mathematics and geometry of magnetic forces, but it is easy to see how random orientations of the atoms in the nonmagnetic substance would lead to cancellation of the forces.

Even if the planetary model does not immediately answer such questions as why one element would be a metal and another a nonmetal, these ideas would be difficult or impossible to conceptualize in the raisin cookie model.

Discussion Questions

A In reality, charges of the same type repel one another and charges of different types are attracted. Suppose the rules were the other way around, giving repulsion between opposite charges and attraction between similar ones. What would the universe be like?

8.2.3 Atomic number

As alluded to in a discussion question in the previous section, scientists of this period had only a very approximate idea of how many units of charge resided in the nuclei of the various chemical elements. Although we now associate the number of units of nuclear charge with the element's position on the periodic table, and call it the atomic number, they had no idea that such a relationship existed. Mendeleev's table just seemed like an organizational tool, not something with any necessary physical significance. And everything Mendeleev had done seemed equally valid if you turned the table upside-down or reversed its left and right sides, so even if you wanted to number the elements sequentially with integers, there was an ambiguity as to how to do it. Mendeleev's original table was in fact upside-down compared to the modern one.

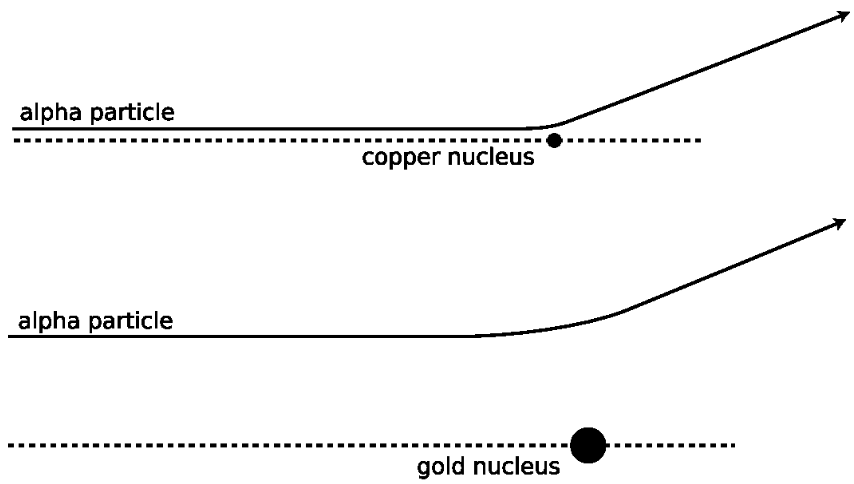
1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	57 La *	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	89 Ac **	104 Rf	105 Ha	106	107	108	109	110	111	112	113	114	115	116	117	118

*	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
**	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

k / A modern periodic table, labeled with atomic numbers. Mendeleev's original table was upside-down compared to this one.

In the period immediately following the discovery of the nucleus, physicists only had rough estimates of the charges of the various nuclei. In the case of the very lightest nuclei, they simply found the maximum number of electrons they could strip off by various methods: chemical reactions, electric sparks, ultraviolet light, and so on. For example they could easily strip of one or two electrons from helium, making He^{+} or He^{++}, but nobody could make He^{+++}, presumably because the nuclear charge of helium was only $+2e$. Unfortunately only a few of the lightest elements could be stripped completely, because the more electrons were stripped off, the greater the positive net charge remaining, and the more strongly the rest of the negatively charged electrons would be held on. The heavy elements' atomic numbers could only be roughly extrapolated from the light elements, where the atomic number was about half the atom's mass expressed in units of the mass of a hydrogen atom. Gold, for example, had a mass about 197 times that of hydrogen, so its atomic number was estimated to be about half that, or somewhere around 100. We now know it to be 79.

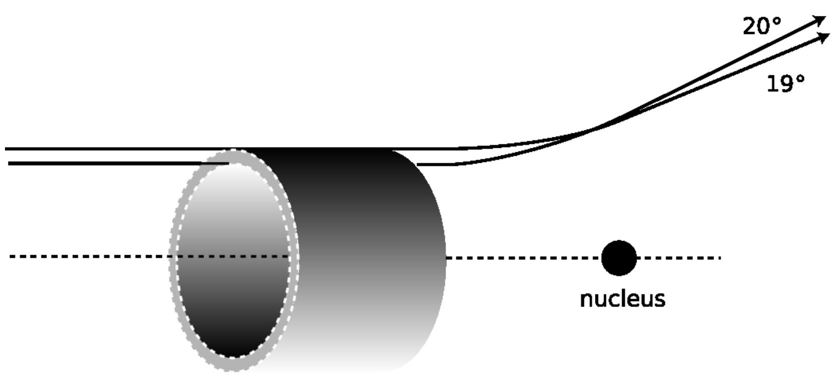
How did we finally find out? The riddle of the nuclear charges was at last successfully attacked using two different techniques, which gave consistent results. One set of experiments, involving x-rays, was performed by the young Henry Mosely, whose scientific brilliance was soon to be sacrificed in a battle between European imperialists over who would own the Dardanelles, during that pointless conflict then known as the War to End All Wars, and now referred to as World War I.



l / An alpha particle has to come much closer to the low-charged copper nucleus in order to be deflected through the same angle.

Since Mosely's analysis requires several concepts with which you are not yet familiar, we will instead describe the technique used by James Chadwick at around the same time. An added bonus of describing Chadwick's experiments is that they presaged the important modern technique of studying *collisions* of subatomic particles. In grad school, I worked with a professor whose thesis adviser's thesis adviser was Chadwick, and he related some interesting stories about the man. Chadwick was apparently a little nutty and a complete fanatic about science, to the extent that when he was held in a German prison camp during World War II, he managed to cajole his captors into allowing him to scrounge up parts from broken radios so that he could attempt to do physics experiments.

Chadwick's experiment worked like this. Suppose you perform two Rutherford-type alpha scattering measurements, first one with a gold foil as a target as in Rutherford's original experiment, and then one with a copper foil. It is possible to get large angles of deflection in both cases, but as shown in figure m, the alpha particle must be heading almost straight for the copper nucleus to get the same angle of deflection that would have occurred with an alpha that was much farther off the mark; the gold nucleus' charge is so much greater than the copper's that it exerts a strong force on the alpha particle even from far off. The situation is very much like that of a blindfolded person playing darts. Just as it is impossible to aim an alpha particle at an individual nucleus in the target, the blindfolded person cannot really aim the darts. Achieving a very close encounter with the copper atom would be akin to hitting an inner circle on the dartboard. It's much more likely that one would have the luck to hit the outer circle, which covers a greater number of square inches. By analogy, if you measure the frequency with which alphas are scattered by copper at some particular angle, say between 19 and 20 degrees, and then perform the same measurement at the same angle with gold, you get a much higher percentage for gold than for copper.



m / An alpha particle must be headed for the ring on the front of the imaginary cylindrical pipe in order to produce scattering at an angle between 19 and 20 degrees. The area of this ring is called the "cross-section" for scattering at 19-20 °because it is the cross-sectional area of a cut through the pipe.

In fact, the numerical ratio of the two nuclei's charges can be derived from this same experimentally determined ratio. Using the standard notation Z for the atomic number (charge of the nucleus divided by e), the following equation can be proved (example 6):

$$\frac{Z_{gold}^2}{Z_{copper}^2} = \frac{\text{number of alphas scattered by gold at 19-20}^\circ}{\text{number of alphas scattered by copper at 19-20}^\circ}$$

By making such measurements for targets constructed from all the elements, one can infer the ratios of all the atomic numbers, and since the atomic numbers of the light elements were already known, atomic numbers could be assigned to the entire periodic table. According to Mosely, the atomic numbers of copper, silver and platinum were 29, 47, and 78, which corresponded well with their positions on the periodic table. Chadwick's figures for the same elements were 29.3, 46.3, and 77.4, with error bars of about 1.5 times the fundamental charge, so the two experiments were in good agreement.

The point here is absolutely not that you should be ready to plug numbers into the above equation for a homework or exam question! My overall goal in this chapter is to explain how we know what we know about atoms. An added bonus of describing Chadwick's experiment is that the approach is very similar to that used in modern particle physics experiments, and the ideas used in the analysis are closely related to the now-ubiquitous concept of a "cross-section."

In the dartboard analogy, the cross-section would be the area of the circular ring you have to hit. The reasoning behind the invention of the term "cross-section" can be visualized as shown in figure m. In this language, Rutherford's invention of the planetary model came from his unexpected discovery that there was a nonzero cross-section for alpha scattering from gold at large angles, and Chadwick confirmed Mosely's determinations of the atomic numbers by measuring cross-sections for alpha scattering.

Proof of the relationship between Z and scattering *example 6*

The equation above can be derived by the following not very rigorous proof. To deflect the alpha particle by a certain angle requires that it acquire a certain momentum component in the direction perpendicular to its original momentum. Although the nucleus's force on the alpha particle is not constant, we can pretend that it is approximately constant during the time when the alpha is within a distance equal to, say, 150% of its distance of closest approach, and that the force is zero before and after that part of the motion. (If we chose 120% or 200%, it shouldn't make any difference in the final result, because the final result is a ratio, and the effects on the numerator and denominator should cancel each other.) In the approximation of constant force, the change in the alpha's perpendicular momentum component is then equal to $F\Delta t$. The Coulomb force law says the force is proportional to Z/r^2. Although r does change somewhat during the time interval of interest, it's good enough to treat it as a constant number, since we're only computing the ratio between the two experiments' results. Since we are approximating the force as acting over the time during which the distance is not too much greater than the distance of closest approach, the time interval Δt must be proportional to r, and the sideways momentum imparted to the alpha, $F\Delta t$, is proportional to $(Z/r^2)r$, or Z/r. If we're comparing alphas scattered at the same angle from gold and from copper, then Δp is the same in both cases, and the proportionality $\Delta p \propto Z/r$ tells us that the ones scattered from copper at that angle had to be headed in along a line closer to the central axis by a factor equaling $Z_{\text{gold}}/Z_{\text{copper}}$. If you imagine a "dartboard ring" that the alphas have to hit, then the ring for the gold experiment has the same proportions as the one for copper, but it is enlarged by a factor equal to $Z_{\text{gold}}/Z_{\text{copper}}$. That is, not only is the radius of the ring greater by that factor, but unlike the rings on a normal dartboard, the thickness of the outer ring is also greater in proportion to its radius. When you take a geometric shape and scale it up in size like a photographic enlargement, its area is increased in proportion to the square of the enlargement factor, so the area of the dartboard ring in the gold experiment is greater by a factor equal to $(Z_{\text{gold}}/Z_{\text{copper}})^2$. Since the alphas are aimed entirely randomly, the chances of an alpha hitting the ring are in proportion to the area of the ring, which proves the equation given above.

As an example of the modern use of scattering experiments and cross-section measurements, you may have heard of the recent experimental evidence for the existence of a particle called the top quark. Of the twelve subatomic particles currently believed to be the smallest constituents of matter, six form a family called the quarks, distinguished from the other six by the intense attractive forces that make the quarks stick to each other. (The other six consist of the electron plus five other, more exotic particles.) The only two types of quarks found in naturally occurring matter are the "up quark" and "down quark," which are what protons and neutrons are made of, but four other types were theoretically predicted to exist, for a total of six. (The whimsical term "quark" comes from a line by James Joyce reading "Three quarks for master Mark.") Until recently, only five types of quarks had been proven to exist via experiments, and the sixth, the top quark, was only theorized. There was no hope of ever detecting a top quark directly, since it is radioactive, and only exists for a zillionth of a second before evaporating. Instead, the researchers searching for it at the Fermi National Accelerator Laboratory near Chicago measured cross-sections for scattering of nuclei off of other nuclei. The experiment was much like those of Rutherford and Chadwick, except that the incoming nuclei had to be boosted to much higher speeds in a particle accelerator. The resulting encounter with a target nucleus was so violent that both nuclei were completely demolished, but, as Einstein proved, energy can be converted into matter, and the energy of the collision creates a spray of exotic, radioactive particles, like the deadly shower of wood fragments produced by a cannon ball in an old naval battle. Among those particles were some top quarks. The cross-sections being measured were the cross-sections for the production of certain combinations of these secondary particles. However different the details, the principle was the same as that employed at the turn of the century: you smash things together and look at the fragments that fly off to see what was inside them. The approach has been compared to shooting a clock with a rifle and then studying the pieces that fly off to figure out how the clock worked.

Discussion Questions

A The diagram, showing alpha particles being deflected by a gold nucleus, was drawn with the assumption that alpha particles came in on lines at many different distances from the nucleus. Why wouldn't they all come in along the same line, since they all came out through the same tube?

B Why does it make sense that, as shown in the figure, the trajectories that result in 19° and 20° scattering cross each other?

C Rutherford knew the velocity of the alpha particles emitted by radium, and guessed that the positively charged part of a gold atom had a charge of about $+100e$ (we now know it is $+79e$). Considering the fact that some alpha particles were deflected by 180°, how could he then use conservation of energy to derive an upper limit on the size of a gold nucleus? (For simplicity, assume the size of the alpha particle is negligible compared to that of the gold nucleus, and ignore the fact that the gold nucleus recoils a little from the collision, picking up a little kinetic energy.)

8.2.4 The structure of nuclei

The proton

The fact that the nuclear charges were all integer multiples of e suggested to many physicists that rather than being a pointlike object, the nucleus might contain smaller particles having individual charges of $+e$. Evidence in favor of this idea was not long in arriving. Rutherford reasoned that if he bombarded the atoms of a very light element with alpha particles, the small charge of the target nuclei would give a very weak repulsion. Perhaps those few alpha particles that happened to arrive on head-on collision courses would get so close that they would physically crash into some of the target nuclei. An alpha particle is itself a nucleus, so this would be a collision between two nuclei, and a violent one due to the high speeds involved. Rutherford hit pay dirt in an experiment with alpha particles striking a target containing nitrogen atoms. Charged particles were detected flying out of the target like parts flying off of cars in a high-speed crash. Measurements of the deflection of these particles in electric and magnetic fields showed that they had the same charge-to-mass ratio as singly-ionized hydrogen atoms. Rutherford concluded that these were the conjectured singly-charged particles that held the charge of the nucleus, and they were later named protons. The hydrogen nucleus consists of a single proton, and in general, an element's atomic number gives the number of protons contained in each of its nuclei. The mass of the proton is about 1800 times greater than the mass of the electron.

The neutron

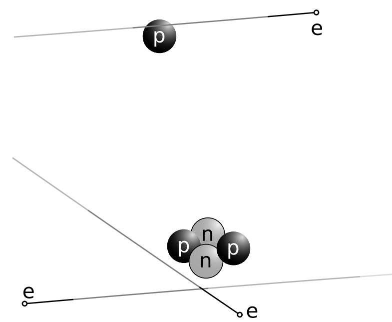
It would have been nice and simple if all the nuclei could have been built only from protons, but that couldn't be the case. If you spend a little time looking at a periodic table, you will soon notice that although some of the atomic masses are very nearly integer multiples of hydrogen's mass, many others are not. Even where the masses are close whole numbers, the masses of an element other than hydrogen is always greater than its atomic number, not equal to it. Helium, for instance, has two protons, but its mass is four times greater than that of hydrogen.

Chadwick cleared up the confusion by proving the existence of a new subatomic particle. Unlike the electron and proton, which are electrically charged, this particle is electrically neutral, and he named it the neutron. Chadwick's experiment has been described in detail in chapter 4 of book 2 of this series, but briefly the method was to expose a sample of the light element beryllium to a stream of alpha particles from a lump of radium. Beryllium has only four protons, so an alpha that happens to be aimed directly at a beryllium nucleus can actually hit it rather than being stopped short of a collision by electrical repulsion. Neutrons were observed as a new form of radiation emerging from the collisions, and Chadwick correctly inferred that they were previously unsuspected components of the nucleus that had been knocked out. As described in 3, Chadwick also determined the mass of the neutron; it is very nearly the same as that of the proton.

To summarize, atoms are made of three types of particles:

	charge	*mass in units of the proton's mass*	*location in atom*
proton	$+e$	1	in nucleus
neutron	0	1.001	in nucleus
electron	$-e$	1/1836	orbiting nucleus

The existence of neutrons explained the mysterious masses of the elements. Helium, for instance, has a mass very close to four times greater than that of hydrogen. This is because it contains two neutrons in addition to its two protons. The mass of an atom is essentially determined by the total number of neutrons and protons. The total number of neutrons plus protons is therefore referred to as the atom's *mass number*.



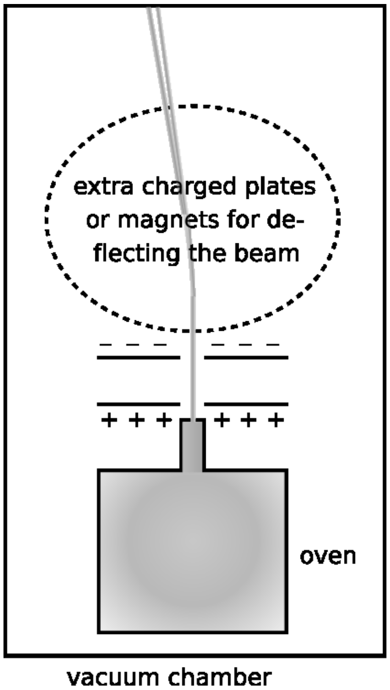
n / Examples of the construction of atoms: hydrogen (top) and helium (bottom). On this scale, the electrons' orbits would be the size of a college campus.

Isotopes

We now have a clear interpretation of the fact that helium is close to four times more massive than hydrogen, and similarly for all the atomic masses that are close to an integer multiple of the mass of hydrogen. But what about copper, for instance, which had an atomic mass 63.5 times that of hydrogen? It didn't seem reasonable to think that it possessed an extra half of a neutron! The solution was found by measuring the mass-to-charge ratios of singly-ionized atoms (atoms with one electron removed). The technique is essentially that same as the one used by Thomson for cathode rays, except that whole atoms do not spontaneously leap out of the surface of an object as electrons sometimes do. Figure o shows an example of how the ions can be created and injected between the charged plates for acceleration.

Injecting a stream of copper ions into the device, we find a surprise — the beam splits into two parts! Chemists had elevated to dogma the assumption that all the atoms of a given element were identical, but we find that 69% of copper atoms have one mass, and 31% have another. Not only that, but both masses are very nearly integer multiples of the mass of hydrogen (63 and 65, respectively). Copper gets its chemical identity from the number of protons in its nucleus, 29, since chemical reactions work by electric forces. But apparently some copper atoms have $63 - 29 = 34$ neutrons while others have $65 - 29 = 36$. The atomic mass of copper, 63.5, reflects the proportions of the mixture of the mass-63 and mass-65 varieties. The different mass varieties of a given element are called $isotopesof$ that element.

Isotopes can be named by giving the mass number as a subscript



o / A version of the Thomson apparatus modified for measuring the mass-to-charge ratios of ions rather than electrons. A small sample of the element in question, copper in our example, is boiled in the oven to create a thin vapor. (A vacuum pump is continuously sucking on the main chamber to keep it from accumulating enough gas to stop the beam of ions.) Some of the atoms of the vapor are ionized by a spark or by ultraviolet light. Ions that wander out of the nozzle and into the region between the charged plates are then accelerated toward the top of the figure. As in the Thomson experiment, mass-to-charge ratios are inferred from the deflection of the beam.

to the left of the chemical symbol, e.g. ^{65}Cu. Examples:

	protons	*neutrons*	*mass number*
^{1}H	1	0	$0+1=1$
^{4}He	2	2	$2+2=4$
^{12}C	6	6	$6+6=12$
^{14}C	6	8	$6+8=14$
^{262}Ha	105	157	$105+157=262$

Self-Check

Why are the positive and negative charges of the accelerating plates reversed in the isotope-separating apparatus compared to the Thomson apparatus? ▷ Answer, p. 708

Chemical reactions are all about the exchange and sharing of electrons: the nuclei have to sit out this dance because the forces of electrical repulsion prevent them from ever getting close enough to make contact with each other. Although the protons do have a vitally important effect on chemical processes because of their electrical forces, the neutrons can have no effect on the atom's chemical reactions. It is not possible, for instance, to separate ^{63}Cu from ^{65}Cu by chemical reactions. This is why chemists had never realized that different isotopes existed. (To be perfectly accurate, different isotopes do behave slightly differently because the more massive atoms move more sluggishly and therefore react with a tiny bit less intensity. This tiny difference is used, for instance, to separate out the isotopes of uranium needed to build a nuclear bomb. The smallness of this effect makes the separation process a slow and difficult one, which is what we have to thank for the fact that nuclear weapons have not been built by every terrorist cabal on the planet.)

Sizes and shapes of nuclei

Matter is nearly all nuclei if you count by weight, but in terms of volume nuclei don't amount to much. The radius of an individual neutron or proton is very close to 1 fm (1 fm=10^{-15} m), so even a big lead nucleus with a mass number of 208 still has a diameter of only about 13 fm, which is ten thousand times smaller than the diameter of a typical atom. Contrary to the usual imagery of the nucleus as a small sphere, it turns out that many nuclei are somewhat elongated, like an American football, and a few have exotic asymmetric shapes like pears or kiwi fruits.

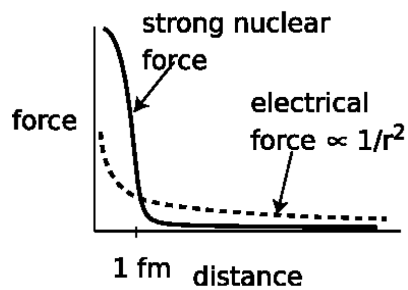
Discussion Questions

A Suppose the entire universe was in a (very large) cereal box, and the nutritional labeling was supposed to tell a godlike consumer what percentage of the contents was nuclei. Roughly what would the percentage be like if the labeling was according to mass? What if it was by volume?

p / A nuclear power plant at Cattenom, France. Unlike the coal and oil plants that supply most of the U.S.'s electrical power, a nuclear power plant like this one releases no pollution or greenhouse gases into the Earth's atmosphere, and therefore doesn't contribute to global warming. The white stuff puffing out of this plant is non-radioactive water vapor. Although nuclear power plants generate long-lived nuclear waste, this waste arguably poses much less of a threat to the biosphere than greenhouse gases would.

8.2.5 The strong nuclear force, alpha decay and fission

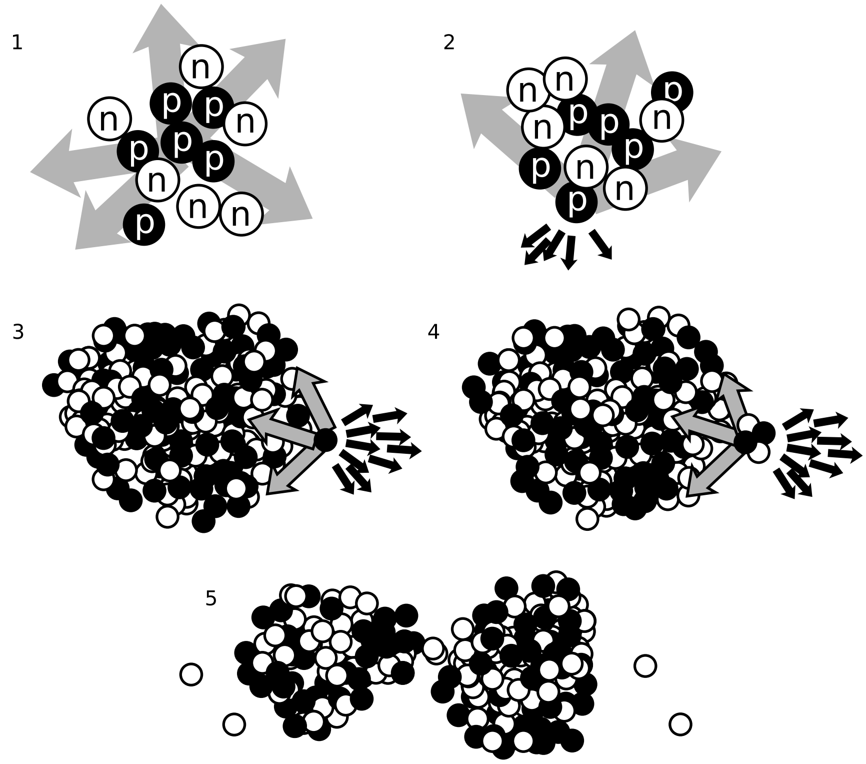
Once physicists realized that nuclei consisted of positively charged protons and uncharged neutrons, they had a problem on their hands. The electrical forces among the protons are all repulsive, so the nucleus should simply fly apart! The reason all the nuclei in your body are not spontaneously exploding at this moment is that there is another force acting. This force, called the strong nuclear force, is always attractive, and acts between neutrons and neutrons, neutrons and protons, and protons and protons with roughly equal strength. The strong nuclear force does not have any effect on electrons, which is why it does not influence chemical reactions.



q / The strong nuclear force cuts off very sharply at a range of about 1 fm.

Unlike electric forces, whose strengths are given by the simple Coulomb force law, there is no simple formula for how the strong nuclear force depends on distance. Roughly speaking, it is effective over ranges of ~ 1 fm, but falls off extremely quickly at larger distances (much faster than $1/r^2$). Since the radius of a neutron or proton is about 1 fm, that means that when a bunch of neutrons and protons are packed together to form a nucleus, the strong nuclear force is effective only between neighbors.

Figure r illustrates how the strong nuclear force acts to keep ordinary nuclei together, but is not able to keep very heavy nuclei from breaking apart. In r/1, a proton in the middle of a carbon nucleus feels an attractive strong nuclear force (arrows) from each of its nearest neighbors. The forces are all in different directions, and tend to cancel out. The same is true for the repulsive electrical



r / 1. The forces on the proton cancel. 2. The forces don't cancel. 3. In a heavy nucleus, the large number of electrical repulsions can add up to a force that is comparable to the strong nuclear attraction. 4. Alpha emission. 5. Fission.

forces (not shown). In figure r/2, a proton at the edge of the nucleus has neighbors only on one side, and therefore all the strong nuclear forces acting on it are tending to pull it back in. Although all the electrical forces from the other five protons (dark arrows) are all pushing it out of the nucleus, they are not sufficient to overcome the strong nuclear forces.

In a very heavy nucleus, r/3, a proton that finds itself near the edge has only a few neighbors close enough to attract it significantly via the strong nuclear force, but every other proton in the nucleus exerts a repulsive electrical force on it. If the nucleus is large enough, the total electrical repulsion may be sufficient to overcome the attraction of the strong force, and the nucleus may spit out a proton. Proton emission is fairly rare, however; a more common type of radioactive decay[1] in heavy nuclei is alpha decay, shown in r/4. The imbalance of the forces is similar, but the chunk that is ejected is an alpha particle (two protons and two neutrons) rather than a single proton.

It is also possible for the nucleus to split into two pieces of roughly equal size, r/5, a process known as fission. Note that in

[1]Alpha decay is more common for statistical reasons. An alpha particle happens to be a very stable arrangement, with an unusually low interaction energy due to the strong nuclear interactions within it. This leaves a lot of spare energy for its center of mass motion, which gives it a greater selection of possible momentum vectors. The idea of a "greater selection" can be made more precise using the concept of phase space, discussed in section 5.4.

addition to the two large fragments, there is a spray of individual neutrons. In a nuclear fission bomb or a nuclear fission reactor, some of these neutrons fly off and hit other nuclei, causing them to undergo fission as well. The result is a chain reaction.

When a nucleus is able to undergo one of these processes, it is said to be radioactive, and to undergo radioactive decay. Some of the naturally occurring nuclei on earth are radioactive. The term "radioactive" comes from Becquerel's image of rays radiating out from something, not from radio waves, which are a whole different phenomenon. The term "decay" can also be a little misleading, since it implies that the nucleus turns to dust or simply disappears – actually it is splitting into two new nuclei with an the same total number of neutrons and protons, so the term "radioactive transformation" would have been more appropriate. Although the original atom's electrons are mere spectators in the process of weak radioactive decay, we often speak loosely of "radioactive atoms" rather than "radioactive nuclei."

Randomness in physics

How does an atom decide when to decay? We might imagine that it is like a termite-infested house that gets weaker and weaker, until finally it reaches the day on which it is destined to fall apart. Experiments, however, have not succeeded in detecting such "ticking clock" hidden below the surface; the evidence is that all atoms of a given isotope are absolutely identical. Why, then, would one uranium atom decay today while another lives for another million years? The answer appears to be that it is entirely random. We can make general statements about the average time required for a certain isotope to decay, or how long it will take for half the atoms in a sample to decay (its half-life), but we can never predict the behavior of a particular atom.

This is the first example we have encountered of an inescapable randomness in the laws of physics. If this kind of randomness makes you uneasy, you're in good company. Einstein's famous quote is "...I am convinced that He [God] does not play dice." Einstein's distaste for randomness, and his association of determinism with divinity, goes back to the Enlightenment conception of the universe as a gigantic piece of clockwork that only had to be set in motion initially by the Builder. Physics had to be entirely rebuilt in the 20th century to incorporate the fundamental randomness of physics, and this modern revolution is the topic of chapter 12. In particular, we will delay the mathematical development of the half-life concept until then.

8.2.6 The weak nuclear force; beta decay

All the nuclear processes we've discussed so far have involved rearrangements of neutrons and protons, with no change in the total number of neutrons or the total number of protons. Now consider the proportions of neutrons and protons in your body and in the planet earth: neutrons and protons are roughly equally numerous in your body's carbon and oxygen nuclei, and also in the nickel and iron that make up most of the earth. The proportions are about 50-50. But the only chemical elements produced in any significant quantities by the big bang were hydrogen (about 90%) and helium (about 10%). If the early universe was almost nothing but hydrogen atoms, whose nuclei are protons, where did all those neutrons come from?

The answer is that there is another nuclear force, the weak nuclear force, that is capable of transforming neutrons into protons and vice-versa. Two possible reactions are

$$\mathrm{n} \rightarrow \mathrm{p} + \mathrm{e}^- + \bar{\nu} \qquad \text{[electron decay]}$$

and

$$\mathrm{p} \rightarrow \mathrm{n} + \mathrm{e}^+ + \nu \qquad . \qquad \text{[positron decay]}$$

(There is also a third type called electron capture, in which a proton grabs one of the atom's electrons and they produce a neutron and a neutrino.)

Whereas alpha decay and fission are just a redivision of the previously existing particles, these reactions involve the destruction of one particle and the creation of three new particles that did not exist before.

There are three new particles here that you have never previously encountered. The symbol e^+ stands for an antielectron, which is a particle just like the electron in every way, except that its electric charge is positive rather than negative. Antielectrons are also known as positrons. Nobody knows why electrons are so common in the universe and antielectrons are scarce. When an antielectron encounters an electron, they annihilate each other, producing gamma rays, and this is the fate of all the antielectrons that are produced by natural radioactivity on earth. Antielectrons are an example of antimatter. A complete atom of antimatter would consist of antiprotons, antielectrons, and antineutrons. Although individual particles of antimatter occur commonly in nature due to natural radioactivity and cosmic rays, only a few complete atoms of antihydrogen have ever been produced artificially.

The notation ν stands for a particle called a neutrino, and $\bar{\nu}$ means an antineutrino. Neutrinos and antineutrinos have no electric charge (hence the name).

We can now list all four of the known fundamental forces of physics:

- gravity

- electromagnetism
- strong nuclear force
- weak nuclear force

The other forces we have learned about, such as friction and the normal force, all arise from electromagnetic interactions between atoms, and therefore are not considered to be fundamental forces of physics.

Decay of ^{212}Pb *example 7*

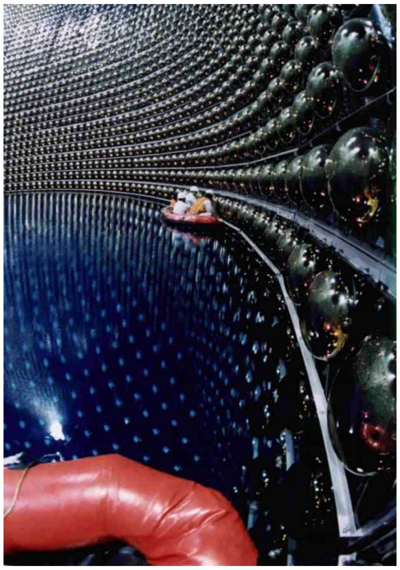
As an example, consider the radioactive isotope of lead ^{212}Pb. It contains 82 protons and 130 neutrons. It decays by the process $n \rightarrow p + e^{-} + \bar{\nu}$. The newly created proton is held inside the nucleus by the strong nuclear force, so the new nucleus contains 83 protons and 129 neutrons. Having 83 protons makes it the element bismuth, so it will be an atom of ^{212}Bi.

In a reaction like this one, the electron flies off at high speed (typically close to the speed of light), and the escaping electrons are the things that make large amounts of this type of radioactivity dangerous. The outgoing electron was the first thing that tipped off scientists in the early 1900*s* to the existence of this type of radioactivity. Since they didn't know that the outgoing particles were electrons, they called them beta particles, and this type of radioactive decay was therefore known as beta decay. A clearer but less common terminology is to call the two processes electron decay and positron decay.

The neutrino or antineutrino emitted in such a reaction pretty much ignores all matter, because its lack of charge makes it immune to electrical forces, and it also remains aloof from strong nuclear interactions. Even if it happens to fly off going straight down, it is almost certain to make it through the entire earth without interacting with any atoms in any way. It ends up flying through outer space forever. The neutrino's behavior makes it exceedingly difficult to detect, and when beta decay was first discovered nobody realized that neutrinos even existed. We now know that the neutrino carries off some of the energy produced in the reaction, but at the time it seemed that the total energy afterwards (not counting the unsuspected neutrino's energy) was greater than the total energy before the reaction, violating conservation of energy. Physicists were getting ready to throw conservation of energy out the window as a basic law of physics when indirect evidence led them to the conclusion that neutrinos existed.

The solar neutrino problem

What about these neutrinos? Why haven't you heard of them before? It's not because they're rare — a billion neutrinos pass through your body every microsecond, but until recently almost nothing was known about them. Produced as a side-effect of the nuclear reactions that power our sun and other stars, these ghostlike bits of matter are believed to be the most numerous particles in the universe. But they interact so weakly with ordinary matter that nearly all the neutrinos that enter the earth on one side will emerge from the other side of our planet without even slowing down.



s / This neutrino detector is in the process of being filled with ultrapure water.

Our first real peek at the properties of the elusive neutrino has come from a huge detector in a played-out Japanese zinc mine, s. An international team of physicists outfitted the mineshaft with wall-to-wall light sensors, and then filled the whole thing with water so pure that you can see through it for a hundred meters, compared to only a few meters for typical tap water. Neutrinos stream through the 50 million liters of water continually, just as they flood everything else around us, and the vast majority never interact with a water molecule. A very small percentage, however, do annihilate themselves in the water, and the tiny flashes of light they produce can be detected by the beachball-sized vacuum tubes that line the darkened mineshaft. Most of the neutrinos around us come from the sun, but for technical reasons this type of water-based detector is more sensitive to the less common but more energetic neutrinos produced when cosmic ray particles strike the earths atmosphere.

Neutrinos were already known to come in three "flavors," which can be distinguished from each other by the particles created when they collide with matter. An "electron-flavored neutrino" creates an ordinary electron when it is annihilated, while the two other types create more exotic particles called mu and tau particles. Think of the three types of neutrinos as chocolate, vanilla, and strawberry. When you buy a chocolate ice cream cone, you expect that it will keep being chocolate as you eat it. The unexpected finding from the Japanese experiment is that some of the neutrinos are changing flavor between the time when they are produced by a cosmic ray and the moment when they wink out of existence in the water. Its as though your chocolate ice cream cone transformed itself magically into strawberry while your back was turned.

How did the physicists figure out the change in flavor? The experiment detects some neutrinos originating in the atmosphere above Japan, and also many neutrinos coming from distant parts of the earth. A neutrino created above the Atlantic Ocean arrives in Japan from underneath, and the experiment can distinguish these upward-traveling neutrinos from the downward-moving local variety. They found that the mixture of neutrinos coming from below was different from the mixture arriving from above, with some of the electron-flavored and tau-flavored neutrinos having apparently changed into mu-flavored neutrinos during their voyage through the earth. The ones coming from above didn't have time to change flavors on their much shorter journey.

This is interpreted as evidence that the neutrinos are constantly changing back and forth among the three flavors. On theoretical grounds, it is believed that such a vibration can only occur if neutrinos have mass. Only a rough estimate of the mass is possible at this point: it appears that neutrinos have a mass somewhere in the neighborhood of one billionth of the mass of an electron, or about 10^{-39} kg.

If the neutrino's mass is so tiny, does it even matter? It matters to astronomers. Neutrinos are the only particles that can be used to probe certain phenomena. For example, they are the only direct probes we have for testing our models of the core of our own sun, which is the source of energy for all life on earth.

A In the reactions $n \rightarrow p + e^- + \bar{\nu}$ and $p \rightarrow n + e^+ + \nu$, verify that charge is conserved. In beta decay, when one of these reactions happens to a neutron or proton within a nucleus, one or more gamma rays may also be emitted. Does this affect conservation of charge? Would it be possible for some extra electrons to be released without violating charge conservation?

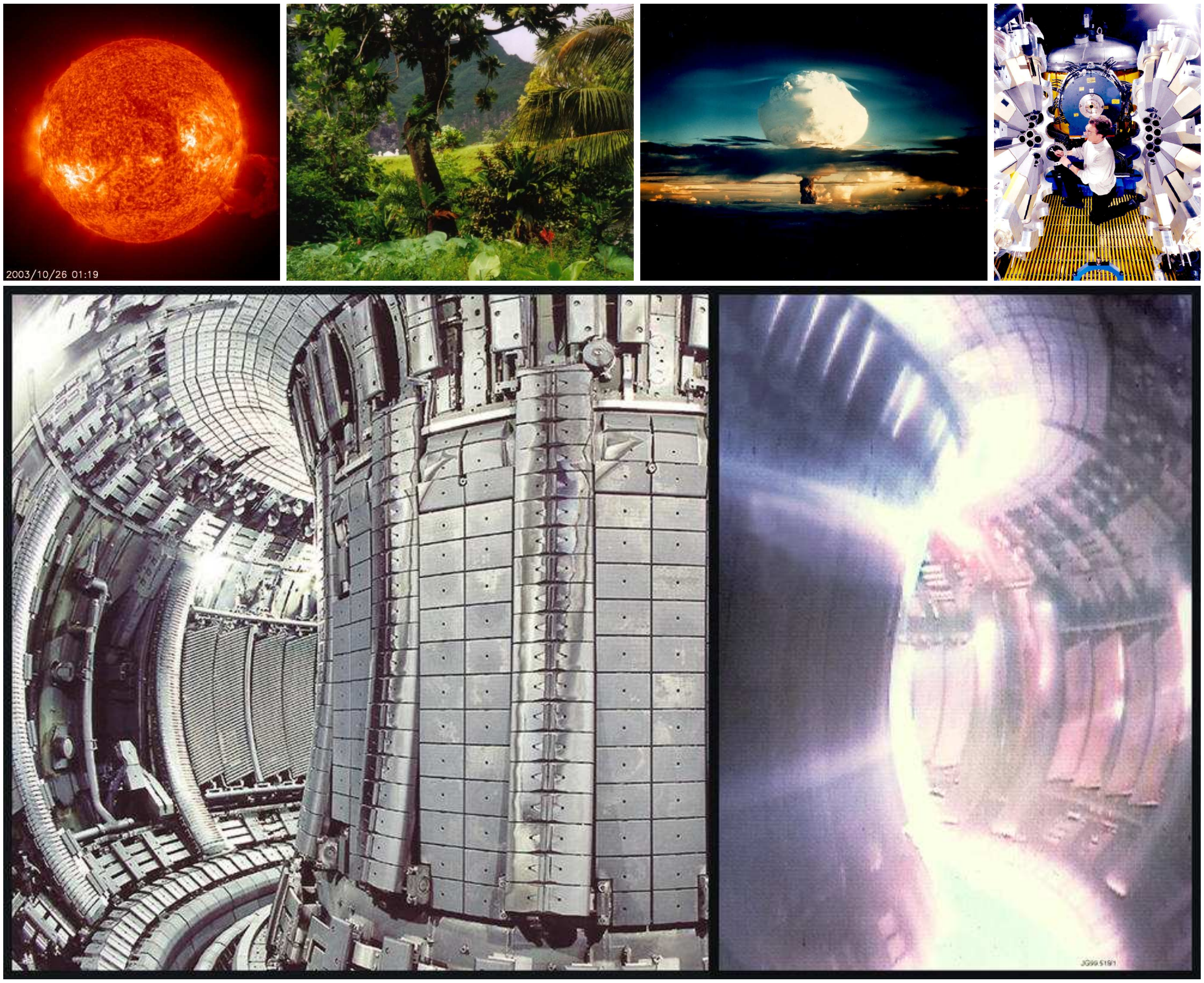
B When an antielectron and an electron annihilate each other, they produce two gamma rays. Is charge conserved in this reaction?

8.2.7 Fusion

As we have seen, heavy nuclei tend to fly apart because each proton is being repelled by every other proton in the nucleus, but is only attracted by its nearest neighbors. The nucleus splits up into two parts, and as soon as those two parts are more than about 1 fm apart, the strong nuclear force no longer causes the two fragments to attract each other. The electrical repulsion then accelerates them, causing them to gain a large amount of kinetic energy. This release of kinetic energy is what powers nuclear reactors and fission bombs.

It might seem, then, that the lightest nuclei would be the most stable, but that is not the case. Let's compare an extremely light nucleus like ^{4}He with a somewhat heavier one, ^{16}O. A neutron or proton in ^{4}He can be attracted by the three others, but in ^{16}O, it might have five or six neighbors attracting it. The ^{16}O nucleus is therefore more stable.

It turns out that the most stable nuclei of all are those around nickel and iron, having about 30 protons and 30 neutrons. Just as a nucleus that is too heavy to be stable can release energy by splitting apart into pieces that are closer to the most stable size, light nuclei can release energy if you stick them together to make bigger nuclei that are closer to the most stable size. Fusing one nucleus with another is called nuclear fusion. Nuclear fusion is what powers our sun and other stars.



t / 1. Our sun's source of energy is nuclear fusion, so nuclear fusion is also the source of power for all life on earth, including, 2, this rain forest in Fatu-Hiva. 3. The first release of energy by nuclear fusion through human technology was the 1952 Ivy Mike test at the Enewetak Atoll. 4. This array of gamma-ray detectors is called GAMMASPHERE. During operation, the array is closed up, and a beam of ions produced by a particle accelerator strikes a target at its center, producing nuclear fusion reactions. The gamma rays can be studied for information about the structure of the fused nuclei, which are typically varieties not found in nature. 5. Nuclear fusion promises to be a clean, inexhaustible source of energy. However, the goal of commercially viable nuclear fusion power has remained elusive, due to the engineering difficulties involved in magnetically containing a plasma (ionized gas) at a sufficiently high temperature and density. This photo shows the experimental JET reactor, with the device opened up on the left, and in action on the right.

8.2.8 Nuclear energy and binding energies

In the same way that chemical reactions can be classified as exothermic (releasing energy) or endothermic (requiring energy to react), so nuclear reactions may either release or use up energy. The energies involved in nuclear reactions are greater by a huge factor. Thousands of tons of coal would have to be burned to produce as much energy as would be produced in a nuclear power plant by one kg of fuel.

Although nuclear reactions that use up energy (endothermic reactions) can be initiated in accelerators, where one nucleus is rammed into another at high speed, they do not occur in nature, not even in the sun. The amount of kinetic energy required is simply not available.

To find the amount of energy consumed or released in a nuclear reaction, you need to know how much nuclear interaction energy, U_{nuc}, was stored or released. Experimentalists have determined the amount of nuclear energy stored in the nucleus of every stable element, as well as many unstable elements. This is the amount of mechanical work that would be required to pull the nucleus apart into its individual neutrons and protons, and is known as the nuclear binding energy.

A reaction occurring in the sun — *example 8*

The sun produces its energy through a series of nuclear fusion reactions. One of the reactions is

$$^1\text{H} + ^2\text{H} \rightarrow ^3\text{He} + \gamma$$

The excess energy is almost all carried off by the gamma ray (not by the kinetic energy of the helium-3 atom). The binding energies in units of pJ (picojoules) are:

^{1}H	0 J
^{2}H	0.35593 pJ
^{3}He	1.23489 pJ

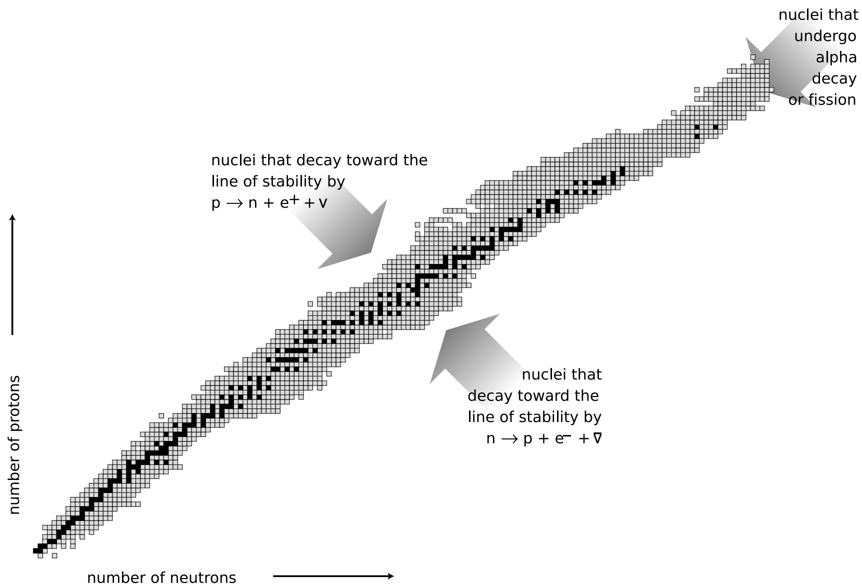
The total initial nuclear energy is 0 pJ+0.35593 pJ, and the final nuclear energy is 1.23489 pJ, so by conservation of energy, the gamma ray must carry off 0.87896 pJ of energy. The gamma ray is then absorbed by the sun and converted to heat.

Self-Check

Why is the binding energy of ^{1}H exactly equal to zero? ▷ Answer, p. 708

By the way, if you add up the masses of the three particles produced in the reaction $\text{n} \rightarrow \text{p} + \text{e}^- + \bar{\nu}$, you will find that they do not equal the mass of the neutron, so mass is not conserved. An even more blatant example is the annihilation of an electron with a positron, $\text{e}^- + \text{e}^+ \rightarrow \gamma$, in which the original mass is completely destroyed, since gamma rays have no mass. Nonconservation of mass is not just a property of nuclear reactions. It also occurs in chemical reactions, but the change in mass is too small to detect with ordinary laboratory balances. The reason why mass is not being conserved is that mass is being converted to energy, according to Einstein's celebrated equation $E = mc^2$. This was discussed briefly in section 1.4, and in more depth in section 7.3.

Figure u is a compact way of showing the vast variety of the nuclei. Each box represents a particular number of neutrons and protons. The black boxes are nuclei that are stable, i.e. that would



u / The known nuclei, represented on a chart of proton number versus neutron number. Note the two nuclei in the bottom row with zero protons. One is simply a single neutron. The other is a cluster of four neutrons. This “tetraneutron” was reported, unexpectedly, to be a bound system in results from a 2002 experiment. The result is controversial. If correct, it implies the existence of a heretofore unsuspected type of matter, the neutron droplet, which we can think of as an atom with no protons or electrons.

require an input of energy in order to change into another. The gray boxes show all the unstable nuclei that have been studied experimentally. Some of these last for billions of years on the average before decaying and are found in nature, but most have much shorter average lifetimes, and can only be created and studied in the laboratory.

The curve along which the stable nuclei lie is called the line of stability. Nuclei along this line have the most stable proportion of neutrons to protons. For light nuclei the most stable mixture is about 50-50, but we can see that stable heavy nuclei have two or three times more neutrons than protons. This is because the electrical repulsions of all the protons in a heavy nucleus add up to a powerful force that would tend to tear it apart. The presence of a large number of neutrons increases the distances among the protons, and also increases the number of attractions due to the strong nuclear force.

8.2.9 Biological effects of ionizing radiation

As a science educator, I find it frustrating that nowhere in the massive amount of journalism devoted to the Chernobyl disaster does one ever find any numerical statements about the amount of radiation to which people have been exposed. Anyone mentally capable of understanding sports statistics or weather reports ought to be able to understand such measurements, as long as something like the following explanatory text was inserted somewhere in the article:

Radiation exposure is measured in units of millirems. The average person is exposed to about 200 millirems each year from natural background sources.

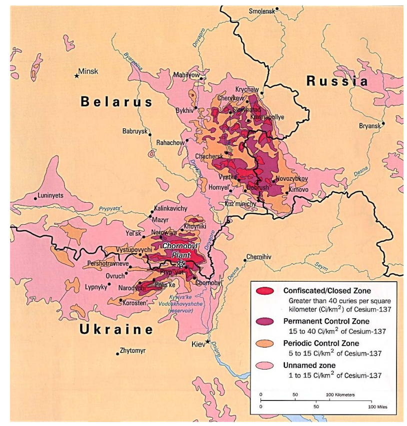
With this context, people would be able to come to informed conclusions based on statements such as, "Children in Finland received an average dose of ___________ millirems above natural background levels because of the Chernobyl disaster."

A millirem, or mrem, is, of course, a thousandth of a rem, but what is a rem? It measures the amount of energy per kilogram deposited in the body by ionizing radiation, multiplied by a "quality factor" to account for the different health hazards posed by alphas, betas, gammas, neutrons, and other types of radiation. Only ionizing radiation is counted, since nonionizing radiation simply heats one's body rather than killing cells or altering DNA. For instance, alpha particles are typically moving so fast that their kinetic energy is sufficient to ionize thousands of atoms, but it is possible for an alpha particle to be moving so slowly that it would not have enough kinetic energy to ionize even one atom.

Notwithstanding the pop culture images of the Incredible Hulk and Godzilla, it is not possible for a multicellular animal to become "mutated" as a whole. In most cases, a particle of ionizing radiation will not even hit the DNA, and even if it does, it will only affect the DNA of a single cell, not every cell in the animal's body. Typically, that cell is simply killed, because the DNA becomes unable to function properly. Once in a while, however, the DNA may be altered so as to make that cell cancerous. For instance, skin cancer can be caused by UV light hitting a single skin cell in the body of a sunbather. If that cell becomes cancerous and begins reproducing uncontrollably, she will end up with a tumor twenty years later.

Other than cancer, the only other dramatic effect that can result from altering a single cell's DNA is if that cell happens to be a sperm or ovum, which can result in nonviable or mutated offspring. Men are relatively immune to reproductive harm from radiation, because their sperm cells are replaced frequently. Women are more vulnerable because they keep the same set of ova as long as they live.

v / An abandoned village near Chernobyl.



w / A map showing levels of radiation near the site of the Chernobyl disaster. At the boundary of the most highly contaminated (bright red) areas, people would be exposed to about 1300 millirem per year, or about four times the natural background level. In the pink areas, which are still densely populated, the exposure is comparable to the natural level found in a high-altitude city such as Denver.

A whole-body exposure of 500,000 mrem will kill a person within a week or so. Luckily, only a small number of humans have ever been exposed to such levels: one scientist working on the Manhattan Project, some victims of the Nagasaki and Hiroshima explosions, and 31 workers at Chernobyl. Death occurs by massive killing of cells, especially in the blood-producing cells of the bone marrow.

Lower levels, on the order of 100,000 mrem, were inflicted on some people at Nagasaki and Hiroshima. No acute symptoms result from this level of exposure, but certain types of cancer are significantly more common among these people. It was originally expected that the radiation would cause many mutations resulting in birth defects, but very few such inherited effects have been observed.

A great deal of time has been spent debating the effects of very low levels of ionizing radiation. A medical x-ray, for instance, may result in a dose on the order of 100 mrem above background, i.e. a doubling of the normal background level. Similar doses in excess of the average background level may be received by people living at high altitudes or people with high concentrations of radon gas in their houses. Unfortunately (or fortunately, depending on how you look at it), the added risks of cancer or birth defects resulting from these levels of exposure are extremely small, and therefore nearly impossible to measure. As with many suspected carcinogenic chemicals, the only practical method of estimating risks is to give laboratory animals doses many orders of magnitude greater, and then assume that the health risk is directly proportional to the dose. Under these assumptions, the added risk posed by a dental x-ray or radon in one's basement is negligible on a personal level, and is only significant in terms of a slight increase in cancer throughout the population. As a matter of social policy, excess radiation exposure is not a significant public health problem compared to car accidents or tobacco smoking.

Discussion Questions

A Should the quality factor for neutrinos be very small, because they mostly don't interact with your body?

B Would an alpha source be likely to cause different types of cancer depending on whether the source was external to the body or swallowed in contaminated food? What about a gamma source?

8.2.10 The creation of the elements

Creation of hydrogen and helium in the Big Bang

We have discussed in chapter 6 the evidence for the Big Bang theory of the origin of the universe. Did all the chemical elements we're made of come into being in the Big Bang? Temperatures in the first microseconds after the Big Bang were so high that atoms and nuclei could not hold together at all. After things had cooled down enough for nuclei and atoms to exist, there was a period of about three minutes during which the temperature and density were high enough for fusion to occur, but not so high that atoms could hold together. We have a good, detailed understanding of the laws of physics that apply under these conditions, so theorists are able to say with confidence that the only element heavier than hydrogen that was created in significant quantities was helium.

We are stardust

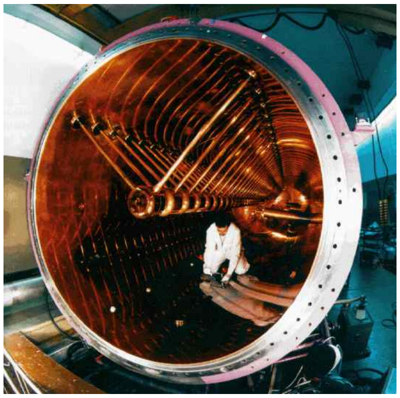
In that case, where did all the other elements come from? Astronomers came up with the answer. By studying the combinations of wavelengths of light, called spectra, emitted by various stars, they had been able to determine what kinds of atoms they contained. (We will have more to say about spectra at the end of this book.) They found that the stars fell into two groups. One type was nearly 100% hydrogen and helium, while the other contained 99% hydrogen and helium and 1% other elements. They interpreted these as two generations of stars. The first generation had formed out of clouds of gas that came fresh from the big bang, and their composition reflected that of the early universe. The nuclear fusion reactions by which they shine have mainly just increased the proportion of helium relative to hydrogen, without making any heavier elements.

The members of the first generation that we see today, however, are only those that lived a long time. Small stars are more miserly with their fuel than large stars, which have short lives. The large stars of the first generation have already finished their lives. Near the end of its lifetime, a star runs out of hydrogen fuel and undergoes a series of violent and spectacular reorganizations as it fuses heavier and heavier elements. Very large stars finish this sequence of events by undergoing supernova explosions, in which some of their material is flung off into the void while the rest collapses into an exotic object such as a black hole or neutron star.

x / The Crab Nebula is a remnant of a supernova explosion. Almost all the elements our planet is made of originated in such explosions.

The second generation of stars, of which our own sun is an example, condensed out of clouds of gas that had been enriched in heavy elements due to supernova explosions. It is those heavy elements that make up our planet and our bodies.

Artificial synthesis of heavy elements



y / Construction of the UNILAC accelerator in Germany, one of whose uses is for experiments to create very heavy artificial elements. In such an experiment, fusion products recoil through a device called SHIP (not shown) that separates them based on their charge-to-mass ratios — it is essentially just a scaled-up version of Thomson's apparatus. A typical experiment runs for several months, and out of the billions of fusion reactions induced during this time, only one or two may result in the production of superheavy atoms. In all the rest, the fused nucleus breaks up immediately. SHIP is used to identify the small number of "good" reactions and separate them from this intense background.

Elements up to uranium, atomic number 92, were created by these astronomical processes. Beyond that, the increasing electrical repulsion of the protons leads to shorter and shorter half-lives. Even if a supernova a billion years ago did create some quantity of an element such as Berkelium, number 97, there would be none left in the Earth's crust today. The heaviest elements have all been created by artificial fusion reactions in accelerators. As of 2006, the heaviest element that has been created is 116 – two atoms of it![2]

Although the creation of a new element, i.e. an atom with a novel number of protons, has historically been considered a glamorous accomplishment, to the nuclear physicist the creation of an atom with a hitherto unobserved number of neutrons is equally important. The greatest neutron number reached so far is 179. One tantalizing goal of this type of research is the theoretical prediction that there might be an island of stability beyond the previously explored tip of the chart of the nuclei shown in section 8.2.8. Just as certain numbers of electrons lead to the chemical stability of the noble gases (helium, neon, argon, ...), certain numbers of neutrons and protons lead to a particularly stable packing of orbits. Calculations dating back to the 1960's have hinted that there might be relatively stable nuclei having approximately 114 protons and 184 neutrons. The isotopes of elements 114 and 116 that have been produced so far have had half-lives in the second or millosecond range. This may not seem like very long, but lifetimes in the microsecond range are more typical for the superheavy elements that have previously been discovered. There is even speculation that certain superheavy isotopes would be stable enough to be produced in quantities that could for instance be weighed and used in chemical reactions.

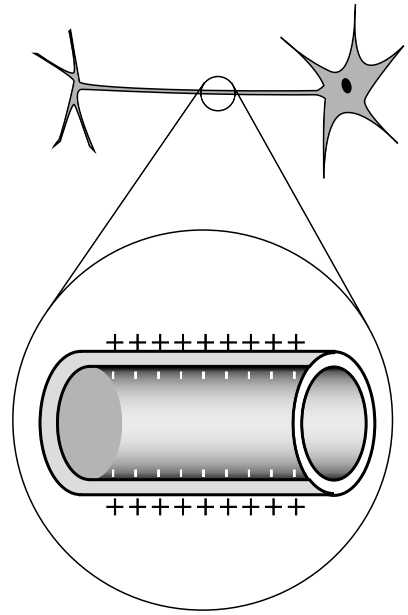
This chapter is summarized on page 732. Notation and terminology are tabulated on pages 718-719.

[2]An earlier claim of the creation of element 116 by a group at Berkeley turned out to be a case of scientific fraud, but the element was later produced by a different group, at Dubna, Russia.

Problems

The symbols √, , etc. are explained on page 405.

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 doesn't the positive charge wander away from the outside surface?

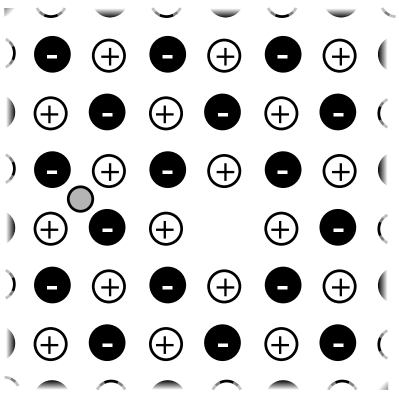


Problem 1. Top: A realistic picture of a neuron. Bottom: A simplified diagram of one segment of the tail (axon).

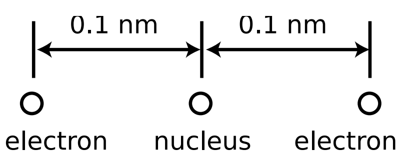
2 Use the nutritional information on some packaged food to make an order-of-magnitude estimate of the amount of chemical energy stored in one atom of food, in units of joules. Assume that a typical atom has a mass of 10^{-26} kg. This constitutes a rough estimate of the amounts of energy there are on the atomic scale. [See chapter 1 of my book Newtonian Physics, at http://www.lightandmatter.com, for help on how to do order-of-magnitude estimates. Note that a nutritional "calorie" is really a kilocalorie; see page 720.] √

3 (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. √
(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. √
(c) Compare this with the result of problem 2.

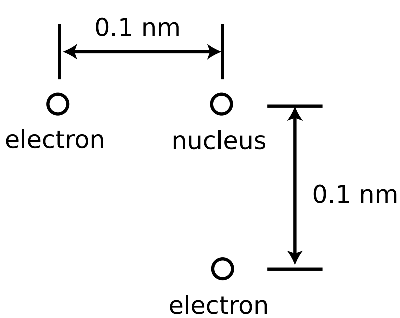
4 A neon light consists of a long glass tube full of neon, with metal caps on the ends. Positive charge is placed on one end of the tube, and negative charge on the other. The electric forces generated can be strong enough to strip electrons off of a certain number of neon atoms. Assume for simplicity 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.] √



Problem 6.



Problem 8.



Problem 9.

5 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 can't explain it using a simple model, you probably don't understand the fancy model as well as you thought you did!)

6 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.
▷ Hint, p. 705 √

7 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. √

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

9 The helium atom of problem 8 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.) √

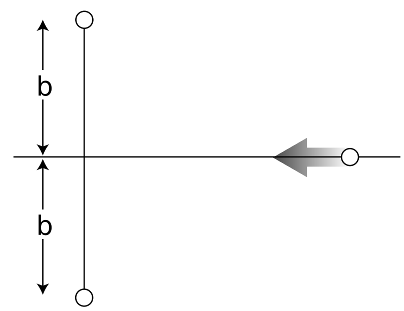
10 Suppose you are holding your hands in front of you, 10 cm apart.
(a) Estimate the total number of electrons in each hand. √
(b) Estimate the total repulsive force of all the electrons in one hand on all the electrons in the other. √
(c) Why don't you feel your hands repelling each other?
(d) Estimate 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.

11 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, and is about 6.5 fm in radius. √

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? √

13 The nuclear process of beta decay by electron capture is described parenthetically on page 392. 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 doesn't occur in hydrogen atoms. (If it did, matter wouldn't exist!)

14 In the semifinals of an electrostatic croquet tournament, Jessica hits her positively charged ball, sending it across the playing field, rolling to the left along the x axis. It is repelled by two other positive charges. These two 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.)



Problem 14.

15 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, $\mathrm{d}q/\mathrm{d}x$, is typically notated as λ (Greek letter lambda). Find the total charge on the wire. √

Key to symbols:
easy typical challenging difficult very difficult
√ An answer check is available at www.lightandmatter.com.

Chapter 9
DC 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.

9.1 Current and Voltage

9.1.1 Current

Unity of all types of electricity

We are surrounded by things we have been *told* are "electrical," but it's far from obvious what they have in common to justify being grouped together. What relationship is there between the way socks cling together and the way a battery lights a lightbulb? We have been told that both an electric eel and our own brains are somehow electrical in nature, but what do they have in common?

British physicist Michael Faraday (1791-1867) set out to address this problem. He investigated electricity from a variety of sources — including electric eels! — to see whether they could all produce the same effects, such as shocks and sparks, attraction and repulsion. "Heating" refers, for example, to the way a lightbulb filament gets hot enough to glow and emit light. Magnetic induction is an effect discovered by Faraday himself that connects electricity and magnetism. We will not study this effect, which is the basis for the electric generator, in detail until later in the book.

	shocks	sparks	attraction and repulsion	heating
rubbing	√	√	√	√
battery	√	√	√	√
animal	√	√	(√)	√
magnetically induced	√	√	√	√

a / *Gymnotus carapo*, a knifefish, uses electrical signals to sense its environment and to communicate with others of its species. (*Greg DeGreef*)

The table shows a summary of some of Faraday's results. Check marks indicate that Faraday or his close contemporaries were able to verify that a particular source of electricity was capable of producing a certain effect. (They evidently failed to demonstrate attraction and repulsion between objects charged by electric eels, although modern workers have studied these species in detail and been able to understand all their electrical characteristics on the same footing as other forms of electricity.)

Faraday's results indicate that there is nothing fundamentally different about the types of electricity supplied by the various sources. They are all able to produce a wide variety of identical effects. Wrote Faraday, "The general conclusion which must be drawn from this collection of facts is that electricity, whatever may be its source, is identical in its nature."

If the types of electricity are the same thing, what thing is that? The answer is provided by the fact that all the sources of electricity can cause objects to repel or attract each other. We use the word "charge" to describe the property of an object that allows it to participate in such electrical forces, and we have learned that charge is present in matter in the form of nuclei and electrons. Evidently all these electrical phenomena boil down to the motion of charged particles in matter.

Electric current

If the fundamental phenomenon is the motion of charged particles, then how can we define a useful numerical measurement of it? We might describe the flow of a river simply by the velocity of the water, but velocity will not be appropriate for electrical purposes because we need to take into account how much charge the moving particles have, and in any case there are no practical devices sold at Radio Shack that can tell us the velocity of charged particles. Experiments show that the intensity of various electrical effects is related to a different quantity: the number of coulombs of charge that pass by a certain point per second. By analogy with the flow of water, this quantity is called the electric *current* , I. Its units of coulombs/second are more conveniently abbreviated as amperes, 1 A=1 C/s. (In informal speech, one usually says "amps.")

The main subtlety involved in this definition is how to account for the two types of charge. The stream of water coming from a hose is made of atoms containing charged particles, but it produces none of the effects we associate with electric currents. For example, you do not get an electrical shock when you are sprayed by a hose. This type of experiment shows that the effect created by the motion of one type of charged particle can be canceled out by the motion of the opposite type of charge in the same direction. In water, every oxygen atom with a charge of $+8e$ is surrounded by eight electrons with charges of $-e$, and likewise for the hydrogen atoms.

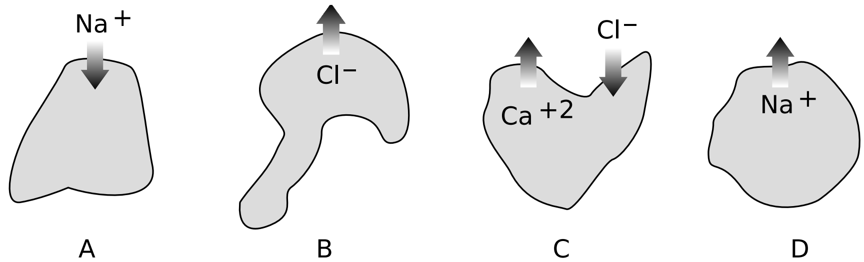
We therefore refine our definition of current as follows:

When charged particles are exchanged between regions of space A and B, the electric current flowing from A to B is defined as

$$I = \frac{\mathrm{d}q}{\mathrm{d}t} \qquad ,$$

where $\mathrm{d}q$ is the change in region B's total charge occurring over a period of time $\mathrm{d}t$.

In the garden hose example, your body picks up equal amounts of positive and negative charge, resulting in no change in your total charge, so the electrical current flowing into you is zero.



b / Example 1

Ions moving across a cell membrane *example 1*

▷ Figure b shows ions, labeled with their charges, moving in or out through the membranes of three cells. If the ions all cross the membranes during the same interval of time, how would the currents into the cells compare with each other?

▷ We're just assuming the rate of flow is constant, so we can talk about Δq instead of dq.

Cell A has positive current going into it because its charge is increased, i.e. has a positive value of Δq.

Cell B has the same current as cell A, because by losing one unit of negative charge it also ends up increasing its own total charge by one unit.

Cell C's total charge is reduced by three units, so it has a large negative current going into it.

Cell D loses one unit of charge, so it has a small negative current into it.

Finding current given charge *example 2*

▷ A charged balloon falls to the ground, and its charge begins leaking off to the Earth. Suppose that the charge on the balloon is given by $q = ae^{-bt}$. Find the current as a function of time, and interpret the answer.

▷ Taking the derivative, we have

$$\begin{aligned} I &= \frac{\mathrm{d}q}{\mathrm{d}t} \\ &= -abe^{-bt} \end{aligned}$$

An exponential function approaches zero as the exponent gets more and more negative. This means that both the charge and the current are decreasing in magnitude with time. It makes sense that the charge approaches zero, since the balloon is losing its charge. It also makes sense that the current is decreasing in magnitude, since charge cannot flow at the same rate forever without overshooting zero.

The reverse of differentiation is integration, so if we know the current as a function of time, we can find the charge by integrating. Example 8 on page 420 shows such a calculation.

It may seem strange to say that a negatively charged particle going one way creates a current going the other way, but this is quite ordinary. As we will see, currents flow through metal wires via the motion of electrons, which are negatively charged, so the direction of motion of the electrons in a circuit is always opposite to the direction of the current. Of course it would have been convenient of Benjamin Franklin had defined the positive and negative signs of charge the opposite way, since so many electrical devices are based on metal wires.

Number of electrons flowing through a lightbulb *example 3*

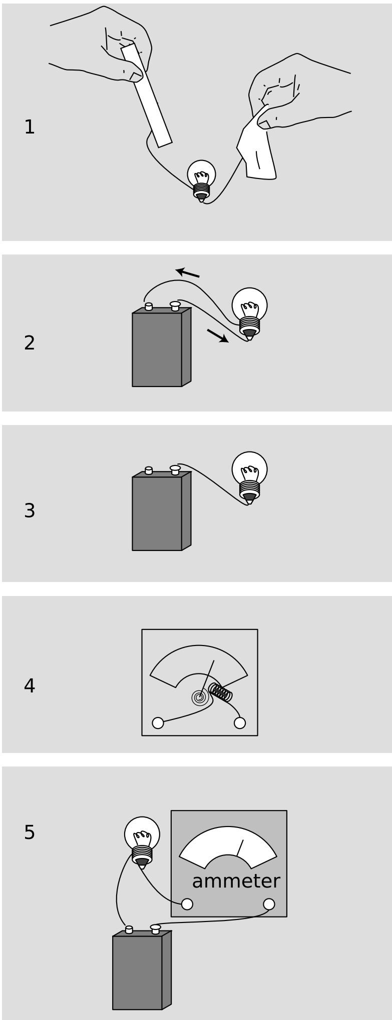
▷ If a lightbulb has 1.0 A flowing through it, how many electrons will pass through the filament in 1.0 s?

▷ We are only calculating the number of electrons that flow, so we can ignore the positive and negative signs. Also, since the rate of flow is constant, we don't really need to think in terms of calculus; the derivative dq/dt that defines current is the same as $\Delta q/\Delta t$ in this situation. Solving for $\Delta q = I\Delta t$ gives a charge of 1.0 C flowing in this time interval. The number of electrons is

$$\begin{aligned}\text{number of electrons} &= \text{coulombs} \times \frac{\text{electrons}}{\text{coulomb}} \\ &= \text{coulombs}/\frac{\text{coulombs}}{\text{electron}} \\ &= 1.0\ \text{C}/e \\ &= 6.2 \times 10^{18}\end{aligned}$$

9.1.2 Circuits

How can we put electric currents to work? The only method of controlling electric charge we have studied so far is to charge different substances, e.g. rubber and fur, by rubbing them against each other. Figure c/1 shows an attempt to use this technique to light a lightbulb. This method is unsatisfactory. True, current will flow through the bulb, since electrons can move through metal wires, and the excess electrons on the rubber rod will therefore come through the wires and bulb due to the attraction of the positively charged fur and the repulsion of the other electrons. The problem is that after a zillionth of a second of current, the rod and fur will both have run out of charge. No more current will flow, and the lightbulb will go out.



c / 1. Static electricity runs out quickly. 2. A practical circuit. 3. An open circuit. 4. How an ammeter works. 5. Measuring the current with an ammeter.

Figure c/2 shows a setup that works. The battery pushes charge through the circuit, and recycles it over and over again. (We will have more to say later in this chapter about how batteries work.) This is called a *complete circuit*. Today, the electrical use of the word "circuit" is the only one that springs to mind for most people, but the original meaning was to travel around and make a round trip, as when a circuit court judge would ride around the boondocks, dispensing justice in each town on a certain date.

Note that an example like c/3 does not work. The wire will quickly begin acquiring a net charge, because it has no way to get rid of the charge flowing into it. The repulsion of this charge will make it more and more difficult to send any more charge in, and soon the electrical forces exerted by the battery will be canceled out completely. The whole process would be over so quickly that the filament would not even have enough time to get hot and glow. This is known as an *open circuit*. Exactly the same thing would happen if the complete circuit of figure c/2 was cut somewhere with a pair of scissors, and in fact that is essentially how an ordinary light switch works: by opening up a gap in the circuit.

The definition of electric current we have developed has the great virtue that it is easy to measure. In practical electrical work, one almost always measures current, not charge. The instrument used to measure current is called an *ammeter*. A simplified ammeter, c/4, simply consists of a coiled-wire magnet whose force twists an iron needle against the resistance of a spring. The greater the current, the greater the force. Although the construction of ammeters may differ, their use is always the same. We break into the path of the electric current and interpose the meter like a tollbooth on a road, c/5. There is still a complete circuit, and as far as the battery and bulb are concerned, the ammeter is just another segment of wire.

Does it matter where in the circuit we place the ammeter? Could we, for instance, have put it in the left side of the circuit instead of the right? Conservation of charge tells us that this can make no difference. Charge is not destroyed or "used up" by the lightbulb, so we will get the same current reading on either side of it. What is "used up" is energy stored in the battery, which is being converted into heat and light energy.

9.1.3 Voltage

The volt unit

Electrical circuits can be used for sending signals, storing information, or doing calculations, but their most common purpose by far is to manipulate energy, as in the battery-and-bulb example of the previous section. We know that lightbulbs are rated in units of watts, i.e. how many joules per second of energy they can convert into heat and light, but how would this relate to the flow of charge as measured in amperes? By way of analogy, suppose your friend, who didn't take physics, can't find any job better than pitching bales of hay. The number of calories he burns per hour will certainly depend on how many bales he pitches per minute, but it will also be proportional to how much mechanical work he has to do on each bale. If his job is to toss them up into a hayloft, he will got tired a lot more quickly than someone who merely tips bales off a loading dock into trucks. In metric units,

$$\frac{\text{joules}}{\text{second}} = \frac{\text{haybales}}{\text{second}} \times \frac{\text{joules}}{\text{haybale}} \qquad .$$

Similarly, the rate of energy transformation by a battery will not just depend on how many coulombs per second it pushes through a circuit but also on how much mechanical work it has to do on each coulomb of charge:

$$\frac{\text{joules}}{\text{second}} = \frac{\text{coulombs}}{\text{second}} \times \frac{\text{joules}}{\text{coulomb}}$$

or

$$\text{power} = \text{current} \times \text{work per unit charge} \qquad .$$

Units of joules per coulomb are abbreviated as *volts*, 1 V=1 J/C, named after the Italian physicist Alessandro Volta. Everyone knows that batteries are rated in units of volts, but the voltage concept is more general than that; it turns out that voltage is a property of every point in space. To gain more insight, let's think more carefully about what goes on in the battery and bulb circuit.

The voltage concept in general

To do work on a charged particle, the battery apparently must be exerting forces on it. How does it do this? Well, the only thing that can exert an electrical force on a charged particle is another charged particle. It's as though the haybales were pushing and pulling each other into the hayloft! This is potentially a horribly complicated situation. Even if we knew how much excess positive or negative charge there was at every point in the circuit (which realistically we don't) we would have to calculate zillions of forces using Coulomb's law, perform all the vector additions, and finally calculate how much work was being done on the charges as they moved along. To make things even more scary, there is more than one type of charged particle that moves: electrons are what move in the wires and the bulb's filament, but ions are the moving charge carriers inside the battery. Luckily, there are two ways in which we can simplify things:

> **The situation is unchanging.** Unlike the imaginary setup in which we attempted to light a bulb using a rubber rod and a piece of fur, this circuit maintains itself in a steady state (after perhaps a microsecond-long period of settling down after the circuit is first assembled). The current is steady, and as charge flows out of any area of the circuit it is replaced by the same amount of charge flowing in. The amount of excess positive or negative charge in any part of the circuit therefore stays constant. Similarly, when we watch a river flowing, the water goes by but the river doesn't disappear.
>
> **Force depends only on position.** Since the charge distribution is not changing, the total electrical force on a charged particle depends only on its own charge and on its location. If another charged particle of the same type visits the same location later on, it will feel exactly the same force.

The second observation tells us that there is nothing all that different about the experience of one charged particle as compared to another's. If we single out one particle to pay attention to, and figure out the amount of work done on it by electrical forces as it goes from point A to point B along a certain path, then this is the same amount of work that will be done on any other charged particles of the same type as it follows the same path. For the sake of visualization, let's think about the path that starts at one terminal of the battery, goes through the light bulb's filament, and ends at the other terminal. When an object experiences a force that depends only on its position (and when certain other, technical conditions are satisfied), we can define an electrical energy associated with the position of that object. The amount of work done on the particle by electrical forces as it moves from A to B equals the drop in electrical energy between A and B. This electrical energy is what is being converted into other forms of energy such as heat and light. We therefore define voltage in general as electrical energy per unit charge:

The difference in voltage between two points in space is defined as

$$\Delta V = \Delta U_{elec}/q \qquad ,$$

where ΔU_{elec} is the change in the electrical energy of a particle with charge q as it moves from the initial point to the final point.

The amount of power dissipated (i.e., rate at which energy is transformed by the flow of electricity) is then given by the equation

$$P = I\Delta V \qquad .$$

Energy stored in a battery *example 4*

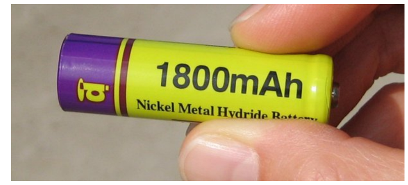
▷ The 1.2 V rechargeable battery in figure d is labeled 1800 milliamp-hours. What is the maximum amount of energy the battery can store?

▷ An ampere-hour is a unit of current multiplied by a unit of time. Current is charge per unit time, so an ampere-hour is in fact a funny unit of *charge*:

$$\begin{aligned}(1\text{ A})(1\text{ hour}) &= (1\text{ C/s})(3600\text{ s})\\ &= 3600\text{ C}\end{aligned}$$

1800 milliamp-hours is therefore $1800 \times 10^{-3} \times 3600$ C = 6.5×10^3 C. That's a huge number of charged particles, but the total loss of electrical energy will just be their total charge multiplied by the voltage difference across which they move:

$$\begin{aligned}\Delta U_{elec} &= q\Delta V\\ &= (6.5 \times 10^3\text{ C})(1.2\text{ V})\\ &= 7.8\text{ kJ}\end{aligned}$$



d / Example 4.

Units of volt-amps *example 5*

▷ Doorbells are often rated in volt-amps. What does this combination of units mean?

▷ Current times voltage gives units of power, $P = I\Delta V$, so volt-amps are really just a nonstandard way of writing watts. They are telling you how much power the doorbell requires.

Power dissipated by a battery and bulb *example 6*

▷ If a 9.0-volt battery causes 1.0 A to flow through a lightbulb, how much power is dissipated?

▷ The voltage rating of a battery tells us what voltage difference ΔV it is designed to maintain between its terminals.

$$\begin{aligned}P &= I\,\Delta\text{V}\\ &= 9.0\text{ A}\cdot\text{V}\\ &= 9.0\,\frac{\text{C}}{\text{s}}\cdot\frac{\text{J}}{\text{C}}\\ &= 9.0\text{ J/s}\\ &= 9.0\text{ W}\end{aligned}$$

The only nontrivial thing in this problem was dealing with the units. One quickly gets used to translating common combinations like A · V into simpler terms.

Here are a few questions and answers about the voltage concept.

Question: OK, so what *is* voltage, really?
Answer: A device like a battery has positive and negative charges inside it that push other charges around the outside circuit. A higher-voltage battery has denser charges in it, which will do more work on each charged particle that moves through the outside circuit.

To use a gravitational analogy, we can put a paddlewheel at the bottom of either a tall waterfall or a short one, but a kg of water that falls through the greater gravitational energy difference will have more energy to give up to the paddlewheel at the bottom.

Question: Why do we define voltage as electrical energy divided by charge, instead of just defining it as electrical energy?
Answer: One answer is that it's the only definition that makes the equation $P = I\Delta V$ work. A more general answer is that we want to be able to define a voltage difference between any two points in space without having to know in advance how much charge the particles moving between them will have. If you put a nine-volt battery on your tongue, then the charged particles that move across your tongue and give you that tingly sensation are not electrons but ions, which may have charges of $+e$, $-2e$, or practically anything. The manufacturer probably expected the battery to be used mostly in circuits with metal wires, where the charged particles that flowed would be electrons with charges of $-e$. If the ones flowing across your tongue happen to have charges of $-2e$, the electrical energy difference for them will be twice as much, but dividing by their charge of $-2e$ in the definition of voltage will still give a result of 9 V.

Question: Are there two separate roles for the charged particles in the circuit, a type that sits still and exerts the forces, and another that moves under the influence of those forces?
Answer: No. Every charged particle simultaneously plays both roles. Newton's third law says that any particle that has an electrical forces acting on it must also be exerting an electrical force back on the other particle. There are no "designated movers" or "designated force-makers."

Question: Why does the definition of voltage only refer to voltage *differences*?
Answer: It's perfectly OK to define voltage as $V = U_{elec}/q$. But recall that it is only *differences* in interaction energy, U, that have direct physical meaning in physics. Similarly, voltage differences are really more useful than absolute voltages. A voltmeter measures voltage differences, not absolute voltages.

Discussion Questions

A A roller coaster is sort of like an electric circuit, but it uses gravitational forces on the cars instead of electric ones. What would a high-voltage roller coaster be like? What would a high-current roller coaster be like?

B Criticize the following statements:

> "He touched the wire, and 10000 volts went through him."
>
> "That battery has a charge of 9 volts."
>
> "You used up the charge of the battery."

C When you touch a 9-volt battery to your tongue, both positive and negative ions move through your saliva. Which ions go which way?

D I once touched a piece of physics apparatus that had been wired incorrectly, and got a several-thousand-volt voltage difference across my hand. I was not injured. For what possible reason would the shock have had insufficient power to hurt me?

9.1.4 Resistance

Resistance

So far we have simply presented it as an observed fact that a battery-and-bulb circuit quickly settles down to a steady flow, but why should it? Newton's second law, $a = F/m$, would seem to predict that the steady forces on the charged particles should make them whip around the circuit faster and faster. The answer is that as charged particles move through matter, there are always forces, analogous to frictional forces, that resist the motion. These forces need to be included in Newton's second law, which is really $a = F_{total}/m$, not $a = F/m$. If, by analogy, you push a crate across the floor at constant speed, i.e. with zero acceleration, the total force on it must be zero. After you get the crate going, the floor's frictional force is exactly canceling out your force. The chemical energy stored in your body is being transformed into heat in the crate and the floor, and no longer into an increase in the crate's kinetic energy. Similarly, the battery's internal chemical energy is converted into heat, not into perpetually increasing the charged particles' kinetic energy. Changing energy into heat may be a nuisance in some circuits, such as a computer chip, but it is vital in a lightbulb, which must get hot enough to glow. Whether we like it or not, this kind of heating effect is going to occur any time charged particles move through matter.

What determines the amount of heating? One flashlight bulb designed to work with a 9-volt battery might be labeled 1.0 watts, another 5.0. How does this work? Even without knowing the details of this type of friction at the atomic level, we can relate the heat dissipation to the amount of current that flows via the equation $P = I\Delta\text{V}$. If the two flashlight bulbs can have two different values of P when used with a battery that maintains the same ΔV, it must be that the 5.0-watt bulb allows five times more current to flow through it.

e / Georg Simon Ohm (1787-1854).

For many substances, including the tungsten from which lightbulb filaments are made, experiments show that the amount of current that will flow through it is directly proportional to the voltage difference placed across it. For an object made of such a substance, we define its electrical *resistance* as follows:

If an object inserted in a circuit displays a current flow which is proportional to the voltage difference across it, then we define its resistance as the constant ratio

$$R = \Delta V/I$$

The units of resistance are volts/ampere, usually abbreviated as ohms, symbolized with the capital Greek letter omega, Ω.

Resistance of a lightbulb — *example 7*

▷ A flashlight bulb powered by a 9-volt battery has a resistance of 10 Ω. How much current will it draw?

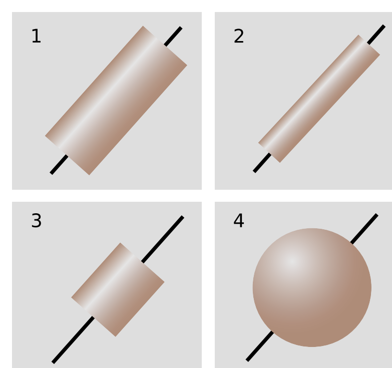
▷ Solving the definition of resistance for I, we find

$$\begin{aligned} I &= \Delta V/R \\ &= 0.9\ \text{V}/\Omega \\ &= 0.9\ \text{V}/(\text{V}/\text{A}) \\ &= 0.9\ \text{A} \end{aligned}$$

Ohm's law states that many substances, including many solids and some liquids, display this kind of behavior, at least for voltages that are not too large. The fact that Ohm's law is called a "law" should not be taken to mean that all materials obey it, or that it has the same fundamental importance as Newton's laws, for example. Materials are called *ohmic* or *nonohmic*, depending on whether they obey Ohm's law.

If objects of the same size and shape made from two different ohmic materials have different resistances, we can say that one material is more resistive than the other, or equivalently that it is less conductive. Materials, such as metals, that are very conductive are said to be good *conductors*. Those that are extremely poor conductors, for example wood or rubber, are classified as *insulators*. There is no sharp distinction between the two classes of materials. Some, such as silicon, lie midway between the two extremes, and are called semiconductors.

On an intuitive level, we can understand the idea of resistance by making the sounds "hhhhhh" and "ffffff." To make air flow out of your mouth, you use your diaphragm to compress the air in your chest. The pressure difference between your chest and the air outside your mouth is analogous to a voltage difference. When you make the "h" sound, you form your mouth and throat in a way that allows air to flow easily. The large flow of air is like a large current. Dividing by a large current in the definition of resistance means that we get a small resistance. We say that the small resistance of your mouth and throat allows a large current to flow. When you make the "f" sound, you increase the resistance and cause a smaller current to flow.



f / Four objects made of the same substance have different resistances.

Note that although the resistance of an object depends on the substance it is made of, we cannot speak simply of the "resistance of gold" or the "resistance of wood." Figure f shows four examples of objects that have had wires attached at the ends as electrical connections. If they were made of the same substance, they would all nevertheless have different resistances because of their different sizes and shapes. A more detailed discussion will be more natural in the context of the following chapter, but it should not be too surprising that the resistance of f/2 will be greater than that of

g / A superconducting segment of the ATLAS accelerator at Argonne National Laboratory near Chicago. It is used to accelerate beams of ions to a few percent of the speed of light for nuclear physics research. The shiny silver-colored surfaces are made of the element niobium, which is a superconductor at relatively high temperatures compared to other metals — relatively high meaning the temperature of liquid helium! The beam of ions passes through the holes in the two small cylinders on the ends of the curved rods. Charge is shuffled back and forth between them at a frequency of 12 million cycles per second, so that they take turns being positive and negative. The positively charged beam consists of short spurts, each timed so that when it is in one of the segments it will be pulled forward by negative charge on the cylinder in front of it and pushed forward by the positively charged one behind. The huge currents involved would quickly melt any metal that was not superconducting, but in a superconductor they produce no heat at all.

f/1 — the image of water flowing through a pipe, however incorrect, gives us the right intuition. Object f/3 will have a smaller resistance than f/1 because the charged particles have less of it to get through.

Superconductors

All materials display some variation in resistance according to temperature (a fact that is used in thermostats to make a thermometer that can be easily interfaced to an electric circuit). More spectacularly, most metals have been found to exhibit a sudden change to *zero* resistance when cooled to a certain critical temperature. They are then said to be superconductors. Theoretically, superconductors should make a great many exciting devices possible, for example coiled-wire magnets that could be used to levitate trains. In practice, the critical temperatures of all metals are very low, and the resulting need for extreme refrigeration has made their use uneconomical except for such specialized applications as particle accelerators for physics research.

But scientists have recently made the surprising discovery that certain ceramics are superconductors at less extreme temperatures. The technological barrier is now in finding practical methods for making wire out of these brittle materials. Wall Street is currently investing billions of dollars in developing superconducting devices for cellular phone relay stations based on these materials. In 2001, the city of Copenhagen replaced a short section of its electrical power trunks with superconducing cables, and they are now in operation and supplying power to customers.

There is currently no satisfactory theory of superconductivity in general, although superconductivity in metals is understood fairly well. Unfortunately I have yet to find a fundamental explanation of superconductivity in metals that works at the introductory level.

Finding charge given current *example 8*

▷ In the segment of the ATLAS accelerator shown in figure g, the current flowing back and forth between the two cylinders is given by $I = a\cos bt$. What is the charge on one of the cylinders as a function of time? ▷ We are given the current and want to find the charge, i.e. we are given the derivative and we want to find the original function that would give that derivative. This means we need to integrate:

$$\begin{aligned} q &= \int I\mathrm{d}t \\ &= \int a\cos bt\,\mathrm{d}t \\ &= \frac{a}{b}\sin bt + q_\mathrm{o} \quad , \end{aligned}$$

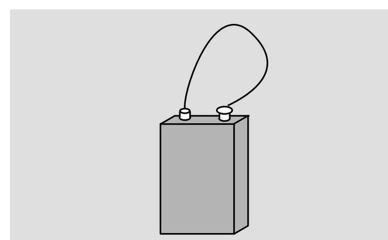
where q_o is a constant of integration.

We can interpret this in order to explain why a superconductor needs to be used. The constant b must be very large, since the current is supposed to oscillate back and forth millions of times a second. Looking at the final result, we see that if b is a very large number, and q is to be a significant amount of charge, then *a* must be a very large number as well. If *a* is numerically large, then the current must be very large, so it would heat the accelerator too much if it was flowing through an ordinary conductor.

Constant voltage throughout a conductor

The idea of a superconductor leads us to the question of how we should expect an object to behave if it is made of a very good conductor. Superconductors are an extreme case, but often a metal wire can be thought of as a perfect conductor, for example if the parts of the circuit other than the wire are made of much less conductive materials. What happens if R equals zero in the equation $R = \Delta V/I$? The result of dividing two numbers can only be zero if the number on top equals zero. This tells us that if we pick any two points in a perfect conductor, the voltage difference between them must be zero. In other words, the entire conductor must be at the same voltage.

Constant voltage means that no work would be done on a charge as it moved from one point in the conductor to another. If zero work was done only along a certain path between two specific points, it might mean that positive work was done along part of the path and negative work along the rest, resulting in a cancellation. But there is no way that the work could come out to be zero for all possible paths unless the electrical force on a charge was in fact zero at every point. Suppose, for example, that you build up a static charge by scuffing your feet on a carpet, and then you deposit some of that charge onto a doorknob, which is a good conductor. How can all that charge be in the doorknob without creating any electrical force at any point inside it? The only possible answer is that the charge moves around until it has spread itself into just the right configuration so that the forces exerted by all the little bits of excess surface charge on any charged particle within the doorknob exactly canceled out.



h / Short-circuiting a battery. Warning: you can burn yourself this way or start a fire! If you want to try this, try making the connection only very briefly, use a low-voltage battery, and avoid touching the battery or the wire, both of which will get hot.

We can explain this behavior if we assume that the charge placed on the doorknob eventually settles down into a stable equilibrium. Since the doorknob is a conductor, the charge is free to move through it. If it was free to move and any part of it did experience a nonzero total force from the rest of the charge, then it would move, and we would not have an equilibrium.

It also turns out that charge placed on a conductor, once it reaches its equilibrium configuration, is entirely on the surface, not on the interior. We will not prove this fact formally, but it is intuitively reasonable. Suppose, for instance, that the net charge on the conductor is negative, i.e. it has an excess of electrons. These electrons all repel each other, and this repulsion will tend to push them onto the surface, since being on the surface allows them to be as far apart as possible.

Short circuits

So far we have been assuming a perfect conductor. What if it is a good conductor, but not a perfect one? Then we can solve for $\Delta V = IR$. An ordinary-sized current will make a very small result when we multiply it by the resistance of a good conductor such as a metal wire. The voltage throughout the wire will then be nearly constant. If, on the other hand, the current is extremely large, we can have a significant voltage difference. This is what happens in a *short-circuit:* a circuit in which a low-resistance pathway connects the two sides of a voltage source. Note that this is much more specific than the popular use of the term to indicate any electrical malfunction at all. If, for example, you short-circuit a 9-volt battery as shown in figure h, you will produce perhaps a thousand amperes of current, leading to a very large value of $P = I\Delta V$. The wire gets hot!

color	meaning
black	0
brown	1
red	2
orange	3
yellow	4
green	5
blue	6
violet	7
gray	8
white	9
silver	$\pm 10\%$
gold	$\pm 5\%$

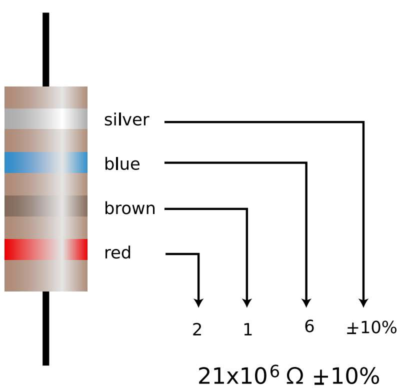
i / Color codes used on resistors.

Self-Check

What would happen to the battery in this kind of short circuit? ▷ Answer, p. 708

Resistors

Inside any electronic gadget you will see quite a few little circuit elements like the one shown below. These *resistors* are simply a cylinder of ohmic material with wires attached to the end.



j / An example of a resistor with a color code.

k / The symbol used in schematics to represent a resistor.

At this stage, most students have a hard time understanding why resistors would be used inside a radio or a computer. We obviously want a lightbulb or an electric stove to have a circuit element that resists the flow of electricity and heats up, but heating is undesirable in radios and computers. Without going too far afield, let's use a mechanical analogy to get a general idea of why a resistor would be used in a radio.

The main parts of a radio receiver are an antenna, a tuner for selecting the frequency, and an amplifier to strengthen the signal sufficiently to drive a speaker. The tuner resonates at the selected frequency, just as in the examples of mechanical resonance discussed in 3. The behavior of a mechanical resonator depends on three things: its inertia, its stiffness, and the amount of friction or damping. The first two parameters locate the peak of the resonance curve, while the damping determines the width of the resonance. In the radio tuner we have an electrically vibrating system that resonates at a particular frequency. Instead of a physical object moving back and forth, these vibrations consist of electrical currents that flow first in one direction and then in the other. In a mechanical system, damping means taking energy out of the vibration in the form of heat, and exactly the same idea applies to an electrical system: the resistor supplies the damping, and therefore controls the width of the resonance. If we set out to eliminate all resistance in the tuner circuit, by not building in a resistor and by somehow getting rid of all the inherent electrical resistance of the wires, we would have a useless radio. The tuner's resonance would be so narrow that we could never get close enough to the right frequency to bring in the station. The roles of inertia and stiffness are played by other circuit elements we have not discusses (a capacitor and a coil).

Many electrical devices are based on electrical resistance and Ohm's law, even if they do not have little components in them that look like the usual resistor. The following are some examples.

Lightbulb

There is nothing special about a lightbulb filament — you can easily make a lightbulb by cutting a narrow waist into a metallic gum wrapper and connecting the wrapper across the terminals of a 9-volt battery. The trouble is that it will instantly burn out. Edison solved this technical challenge by encasing the filament in an evacuated bulb, which prevented burning, since burning requires oxygen.

Polygraph

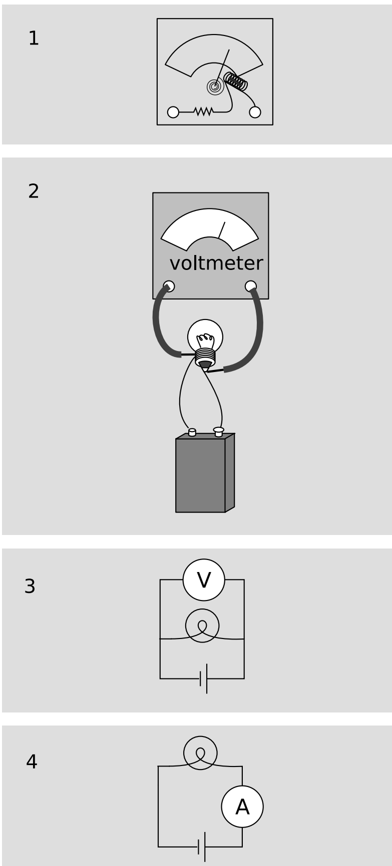
The polygraph, or "lie detector," is really just a set of meters for recording physical measures of the subject's psychological stress, such as sweating and quickened heartbeat. The real-time sweat measurement works on the principle that dry skin is a good insulator, but sweaty skin is a conductor. Of course a truthful subject may become nervous simply because of the situation, and a practiced liar may not even break a sweat. The method's practitioners claim that they can tell the difference, but you should think twice before allowing yourself to be polygraph tested. Most U.S. courts exclude all polygraph evidence, but some employers attempt to screen out dishonest employees by polygraph testing job applicants, an abuse that ranks with such pseudoscience as handwriting analysis.

Fuse

A fuse is a device inserted in a circuit tollbooth-style in the same manner as an ammeter. It is simply a piece of wire made of metals having a relatively low melting point. If too much current passes through the fuse, it melts, opening the circuit. The purpose is to make sure that the building's wires do not carry so much current that they themselves will get hot enough to start a fire. Most modern houses use circuit breakers instead of fuses, although fuses are still common in cars and small devices. A circuit breaker is a switch operated by a coiled-wire magnet, which opens the circuit when enough current flows. The advantage is that once you turn off some of the appliances that were sucking up too much current, you can immediately flip the switch closed. In the days of fuses, one might get caught without a replacement fuse, or even be tempted to stuff aluminum foil in as a replacement, defeating the safety feature.

Voltmeter

A voltmeter is nothing more than an ammeter with an additional high-value resistor through which the current is also forced to flow. Ohm's law relates the current through the resistor is related directly to the voltage difference across it, so the meter can be calibrated in units of volts based on the known value of the resistor. The voltmeter's two probes are touched to the two locations in a circuit between which we wish to measure the voltage difference, l/2. Note how cumbersome this type of drawing is, and how difficult it can be to tell what is connected to what. This is why electrical drawing are usually shown in schematic form. Figure l/3 is a schematic representation of figure l/2.



l / 1. A simplified diagram of how a voltmeter works. 2. Measuring the voltage difference across a lightbulb. 3. The same setup drawn in schematic form. 4. The setup for measuring current is different.

The setups for measuring current and voltage are different. When we are measuring current, we are finding "how much stuff goes through," so we place the ammeter where all the current is forced to go through it. Voltage, however, is not "stuff that goes through," it is a measure of electrical energy. If an ammeter is like the meter that measures your water use, a voltmeter is like a measuring stick that tells you how high a waterfall is, so that you can determine how much energy will be released by each kilogram of falling water. We do not want to force the water to go through the measuring stick! The arrangement in figure l/3 is a *parallel* circuit: one in there are "forks in the road" where some of the current will flow one way and some will flow the other. Figure l/4 is said to be wired in *series:* all the current will visit all the circuit elements one after the other. We will deal with series and parallel circuits in more detail in the following chapter.

If you inserted a voltmeter incorrectly, in series with the bulb and battery, its large internal resistance would cut the current down so low that the bulb would go out. You would have severely disturbed the behavior of the circuit by trying to measure something about it.

Incorrectly placing an ammeter in parallel is likely to be even more disconcerting. The ammeter has nothing but wire inside it to provide resistance, so given the choice, most of the current will flow through it rather than through the bulb. So much current will flow through the ammeter, in fact, that there is a danger of burning out the battery or the meter or both! For this reason, most ammeters have fuses or circuit breakers inside. Some models will trip their circuit breakers and make an audible alarm in this situation, while others will simply blow a fuse and stop working until you replace it.

Discussion Questions

A In figure l/1, would it make any difference in the voltage measurement if we touched the voltmeter's probes to different points along the same segments of wire?

B Explain why it would be incorrect to define resistance as the amount of charge the resistor allows to flow.

9.1.5 Current-conducting properties of materials

Ohm's law has a remarkable property, which is that current will flow even in response to a voltage difference that is as small as we care to make it. In the analogy of pushing a crate across a floor, it is as though even a flea could slide the crate across the floor, albeit at some very low speed. The flea cannot do this because of static friction, which we can think of as an effect arising from the tendency of the microscopic bumps and valleys in the crate and floor to lock together. The fact that Ohm's law holds for nearly all solids has an interesting interpretation: at least some of the electrons are not "locked down" at all to any specific atom.

More generally we can ask how charge actually flows in various solids, liquids, and gases. This will lead us to the explanations of many interesting phenomena, including lightning, the bluish crust that builds up on the terminals of car batteries, and the need for electrolytes in sports drinks.

Solids

In atomic terms, the defining characteristic of a solid is that its atoms are packed together, and the nuclei cannot move very far from their equilibrium positions. It makes sense, then, that electrons, not ions, would be the charge carriers when currents flow in solids. This fact was established experimentally by Tolman and Stewart, in an experiment in which they spun a large coil of wire and then abruptly stopped it. They observed a current in the wire immediately after the coil was stopped, which indicated that charged particles that were not permanently locked to a specific atom had continued to move because of their own inertia, even after the material of the wire in general stopped. The direction of the current showed that it was negatively charged particles that kept moving. The current only lasted for an instant, however; as the negatively charged particles collected at the downstream end of the wire, farther particles were prevented joining them due to their electrical repulsion, as well as the attraction from the upstream end, which was left with a net positive charge. Tolman and Stewart were even able to determine the mass-to-charge ratio of the particles. We need not go into the details of the analysis here, but particles with high mass would be difficult to decelerate, leading to a stronger and longer pulse of current, while particles with high charge would feel stronger electrical forces decelerating them, which would cause a weaker and shorter pulse. The mass-to-charge ratio thus determined was consistent with the m/q of the electron to within the accuracy of the experiment, which essentially established that the particles were electrons.

The fact that only electrons carry current in solids, not ions, has many important implications. For one thing, it explains why wires don't fray or turn to dust after carrying current for a long time. Electrons are very small (perhaps even pointlike), and it is easy to imagine them passing between the cracks among the atoms without creating holes or fractures in the atomic framework. For those who know a little chemistry, it also explains why all the best conductors are on the left side of the periodic table. The elements in that area are the ones that have only a very loose hold on their outermost electrons.

Gases

The molecules in a gas spend most of their time separated from each other by significant distances, so it is not possible for them to conduct electricity the way solids do, by handing off electrons from atom to atom. It is therefore not surprising that gases are good insulators.

Gases are also usually nonohmic. As opposite charges build up on a stormcloud and the ground below, the voltage difference becomes greater and greater. Zero current flows, however, until finally the voltage reaches a certain threshold and we have an impressive example of what is known as a spark or electrical discharge. If air was ohmic, the current between the cloud and the ground would simply increase steadily as the voltage difference increased, rather than being zero until a threshold was reached. This behavior can be explained as follows. At some point, the electrical forces on the air electrons and nuclei of the air molecules become so strong that electrons are ripped right off of some of the molecules. The electrons then accelerate toward either the cloud or the ground, whichever is positively charged, and the positive ions accelerate the opposite way. As these charge carriers accelerate, they strike and ionize other molecules, which produces a rapidly growing cascade.

Liquids

Molecules in a liquid are able to slide past each other, so ions as well as electrons can carry currents. Pure water is a poor conductor because the water molecules tend to hold onto their electrons strongly, and there are therefore not many electrons or ions available to move. Water can become quite a good conductor, however, with the addition of even a small amount of certain substances called electrolytes, which are typically salts. For example, if we add table salt, NaCl, to water, the NaCl molecules dissolve into Na^+ and Cl^- ions, which can then move and create currents. This is why electric currents can flow among the cells in our bodies: cellular fluid is quite salty. When we sweat, we lose not just water but electrolytes, so dehydration plays havoc with our cells' electrical systems. It is for this reason that electrolytes are included in sports drinks and formulas for rehydrating infants who have diarrhea.

Since current flow in liquids involves entire ions, it is not surprising that we can see physical evidence when it has occurred. For example, after a car battery has been in use for a while, the H_2SO_4

battery acid becomes depleted of hydrogen ions, which are the main charge carriers that complete the circuit on the inside of the battery. The leftover SO_4 then forms a visible blue crust on the battery posts.

Speed of currents and electrical signals

When I talk on the phone to my mother in law two thousand miles away, I do not notice any delay while the signal makes its way back and forth. Electrical signals therefore must travel very quickly, but how fast exactly? The answer is rather subtle. For the sake of concreteness, let's restrict ourselves to currents in metals, which consist of electrons.

The electrons themselves are only moving at speeds of perhaps a few thousand miles per hour, and their motion is mostly random thermal motion. This shows that the electrons in my phone cannot possibly be zipping back and forth between California and New York fast enough to carry the signals. Even if their thousand-mile-an-hour motion was organized rather than random, it would still take them many minutes to get there. Realistically, it will take the average electron even longer than that to make the trip. The current in the wire consists only of a slow overall drift, at a speed on the order of a few centimeters per second, superimposed on the more rapid random motion. We can compare this with the slow westward drift in the population of the U.S. If we could make a movie of the motion of all the people in the U.S. from outer space, and could watch it at high speed so that the people appeared to be scurrying around like ants, we would think that the motion was fairly random, and we would not immediately notice the westward drift. Only after many years would we realize that the number of people heading west over the Sierras had exceeded the number going east, so that California increased its share of the country's population.

So why are electrical signals so fast if the average drift speed of electrons is so slow? The answer is that a disturbance in an electrical system can move much more quickly than the charges themselves. It is as though we filled a pipe with golf balls and then inserted an extra ball at one end, causing a ball to fall out at the other end. The force propagated to the other end in a fraction of a second, but the balls themselves only traveled a few centimeters in that time.

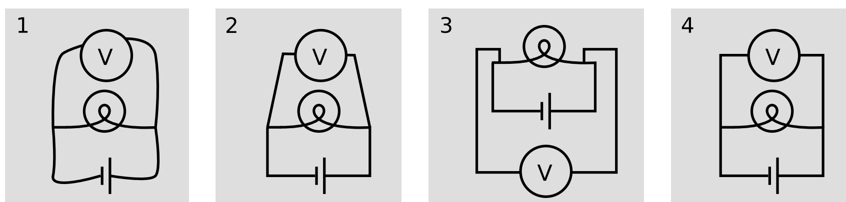
Because the reality of current conduction is so complex, we often describe things using mental shortcuts that are technically incorrect. This is OK as long as we know that they are just shortcuts. For example, suppose the presidents of France and Russia shake hands, and the French politician has inadvertently picked up a positive electrical charge, which shocks the Russian. We may say that the excess positively charged particles in the French leader's body, which all repel each other, take the handshake as an opportunity to get farther apart by spreading out into two bodies rather than one. In reality, it would be a matter of minutes before the ions in one person's body could actually drift deep into the other's. What really happens is that throughout the body of the recipient of the shock there are already various positive and negative ions which are free to move. Even before the perpetrator's charged hand touches the victim's sweaty palm, the charges in the shocker's body begin to repel the positive ions and attract the negative ions in the other person. The split-second sensation of shock is caused by the sudden jumping of the victim's ions by distances of perhaps a micrometer, this effect occurring simultaneously throughout the whole body, although more violently in the hand and arm, which are closer to the other person.

9.2 Parallel and Series Circuits

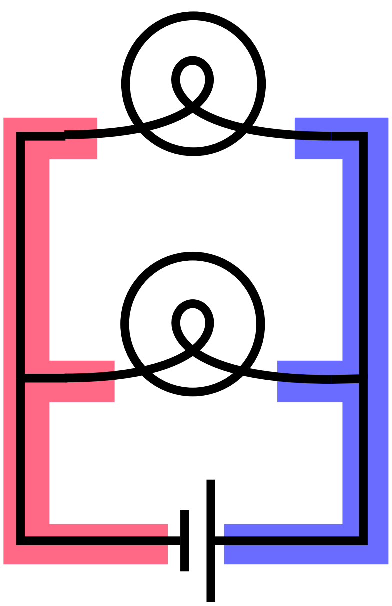
In section 9.1, we limited ourselves to relatively simple circuits, essentially nothing more than a battery and a single lightbulb. The purpose of this chapter is to introduce you to more complex circuits, containing multiple resistors or voltage sources in series, in parallel, or both.

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.



b / The two shaded areas shaped like the letter "E" are both regions of constant voltage.

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.

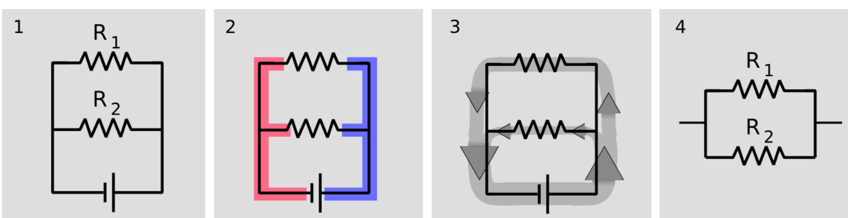
The most important first step in learning to read schematics is to learn to recognize contiguous pieces of wire which must have constant voltage throughout. In figure b, for example, the two shaded E-shaped pieces of wire must each have constant voltage. This focuses our attention on two of the main 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:

$$\Delta V_1 = \Delta V_2 = \Delta V_{battery} \qquad .$$

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

$$I_{total} = I_1 + I_2 \qquad .$$

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

$$\begin{aligned} I_{total} &= \Delta V/R_1 + \Delta V/R_2 \\ &= \Delta V \left(\frac{1}{R_1} + \frac{1}{R_2}\right) \qquad . \end{aligned}$$

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

$$I_{total} = \Delta V/R \qquad ,$$

from which we can determine the equivalent resistance by comparison with the previous expression:

$$1/R = \frac{1}{R_1} + \frac{1}{R_2}$$
$$R = \left(\frac{1}{R_1} + \frac{1}{R_2}\right)^{-1}$$

[equivalent resistance of two resistors in parallel]

Two resistors in parallel, c/4, are equivalent to a single resistor with a value given by the above equation.

Two lamps on the same household circuit *example 9*

▷ 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.

Three resistors in parallel *example 10*

▷ What happens if we have three or more resistors in parallel?

▷ This is an important example, because the solution involves an important technique for understanding circuits: breaking them down into smaller parts and 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

$$R_{12} = \left(\frac{1}{R_1} + \frac{1}{R_2}\right)^{-1} \quad .$$

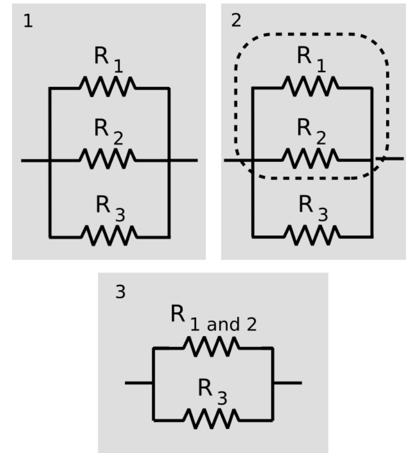
We can then simplify the circuit as shown in d/3, so that it contains only two resistances. The equivalent resistance of the whole circuit is then given by

$$R_{123} = \left(\frac{1}{R_{12}} + \frac{1}{R_3}\right)^{-1} \quad .$$

Substituting for R_{12} and simplifying, we find the result

$$R_{123} = \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right)^{-1} \quad ,$$

which you probably could have guessed. The interesting point here is the divide-and-conquer concept, not the mathematical result.



d / Three resistors in parallel.

An arbitrary number of identical resistors in parallel *example 11*

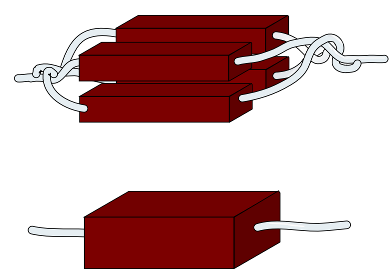
▷ What is the resistance of N identical resistors in parallel?

▷ Generalizing the results for two and three resistors, we have

$$R_N = \left(\frac{1}{R_1} + \frac{1}{R_2} + \ldots\right)^{-1} \quad ,$$

where "..." means that the sum includes all the resistors. If all the resistors are identical, this becomes

$$\begin{aligned} R_N &= \left(\frac{N}{R}\right)^{-1} \\ &= \frac{R}{N} \end{aligned}$$


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.

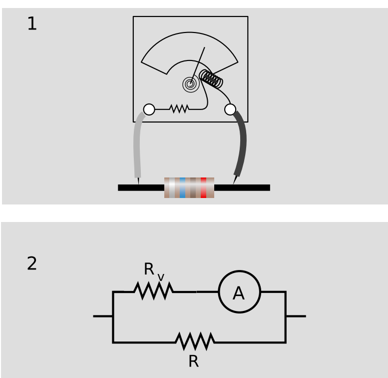
Dependence of resistance on cross-sectional area *example 12*

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

f / A fat pipe has less resistance than a skinny pipe.

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.

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.



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.

Incorrect readings from a voltmeter *example 13*

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

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):

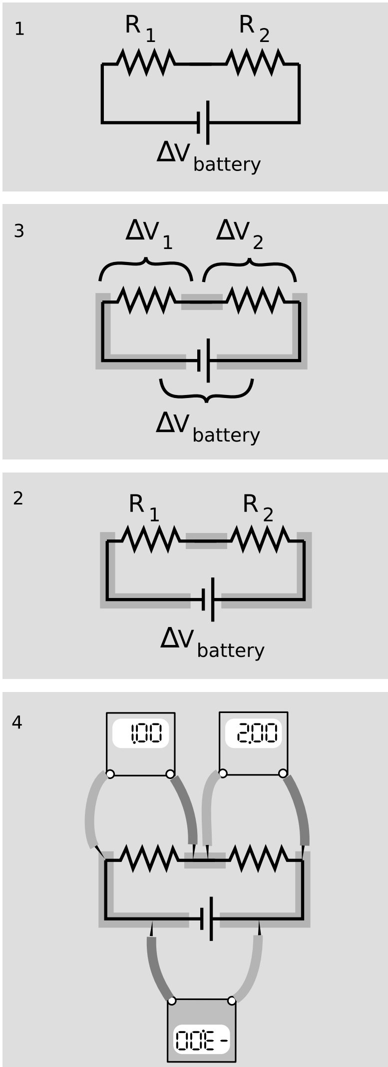
$$I_1 = I_2 \qquad .$$

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

$$|\Delta V_1| + |\Delta V_2| = |\Delta V_{battery}| \qquad .$$

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.



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.

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

$$\Delta V_1 + \Delta V_2 = -\Delta V_{battery} \qquad ,$$

or, in more symmetrical form,

$$\Delta V_1 + \Delta V_2 + \Delta V_{battery} = 0 \qquad .$$

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

$$\Delta V_1 + \Delta V_2 + \Delta V_{battery} = 0$$

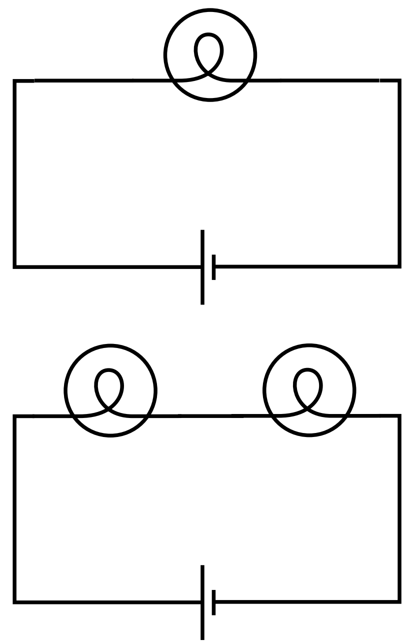
can now be rewritten by applying Ohm's law to each resistor:

$$I_1 R_1 + I_2 R_2 + \Delta V_{battery} = 0 \qquad .$$

The currents are the same, so we can factor them out:

$$I\left(R_1 + R_2\right) + \Delta V_{battery} = 0 \qquad ,$$

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.



i / Example 14.

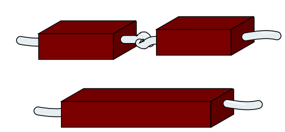
Two lightbulbs in series — *example 14*

▷ If two identical lightbulbs are placed in series, how do their brightnesses compare with the brightness of a single bulb?

▷ Taken as a whole, the pair of bulbs act like a doubled resistance, so they will draw half as much current from the wall. Each bulb will be dimmer than a single bulb would have been.

The total power dissipated by the circuit is $I\Delta V$. The voltage drop across the whole circuit is the same as before, but the current is halved, so the two-bulb circuit draws half as much total power as the one-bulb circuit. Each bulb draws one-quarter of the normal power.

Roughly speaking, we might expect this to result in one quarter the light being produced by each bulb, but in reality lightbulbs waste quite a high percentage of their power in the form of heat and wavelengths of light that are not visible (infrared and ultraviolet). Less light will be produced, but it's hard to predict exactly how much less, since the efficiency of the bulbs will be changed by operating them under different conditions.



j / Doubling the length of a resistor is like putting two resistors in series. The resistance is doubled.

More than two equal resistances in series — *example 15*

By straightforward application of the divide-and-conquer technique discussed in the previous section, we find that the equivalent resistance of N identical resistances R in series will be NR.

Dependence of resistance on length — *example 16*

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.

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.

Choice of high voltage for power lines — *example 17*

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.

Getting killed by your ammeter *example 18*

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.

Discussion Questions

A We have stated the loop rule in a symmetric form where a series of voltage drops adds up to zero. To do this, we had to define a standard way of connecting the voltmeter to the circuit so that the plus and minus signs would come out right. Suppose we wish to restate the junction rule in a similar symmetric way, so that instead of equating the current coming in to the current going out, it simply states that a certain sum of currents at a junction adds up to zero. What standard way of inserting the ammeter would we have to use to make this work?

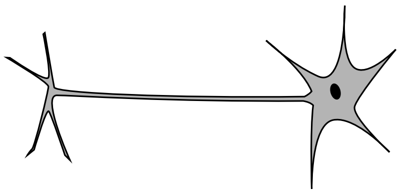
This chapter is summarized on page 733. Notation and terminology are tabulated on pages 718-719.

Problems

The symbols $\surd$, , etc. are explained on page 446.

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, p. 716



Problem 2.

2 Referring back to our old friend the neuron from problem 1 on page 403, 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? $\surd$

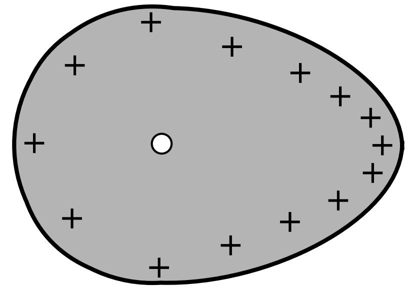
3 If a typical light bulb draws about 900 mA from a 110-V household circuit, what is its resistance? (Don't worry about the fact that it's alternating current.) $\surd$

4 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. $\surd$
(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. $\surd$

5 (a) Express the power dissipated by a resistor in terms of R and ΔV only, eliminating I. $\surd$
(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? $\surd$

6 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. $\surd$

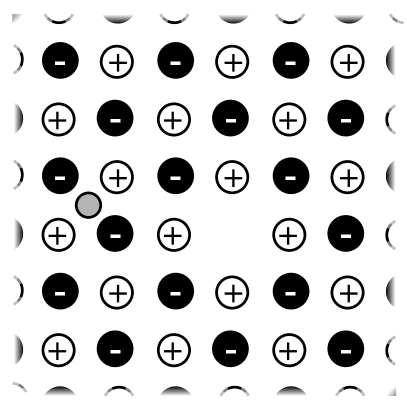
7 In AM (amplitude-modulated) radio, an audio signal $f(t)$ is multiplied by a sine wave $\sin \omega t$ in the megahertz frequency range. For simplicity, let's imagine that the transmitting antenna is a whip, and that charge goes back and forth between the top and bottom. Suppose that, during a certain 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. √



Problem 8.

8 As discussed in the text, when a conductor reaches an equilibrium where its charge is at rest, there is always zero electric force on a charge in its interior, and any excess charge concentrates itself on the surface. The surface layer of charge arranges itself so as to produce zero total force at any point in the interior. (Otherwise the free charge in the interior could not be at rest.) Suppose you have a teardrop-shaped conductor like the one shown in the figure. Since the teardrop is a conductor, there are free charges everywhere inside it, but consider a free charged particle at the location shown with a white circle. Explain why, in order to produce zero force on this particle, the surface layer of charge must be denser in the pointed part of the teardrop. (Similar reasoning shows why lightning rods are made with points. The charged stormclouds induce positive and negative charges to move to opposite ends of the rod. At the pointed upper end of the rod, the charge tends to concentrate at the point, and this charge attracts the lightning.)

9 Use the result of problem 3 on page 403 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.) √



Problem 10.

10 Referring back to problem 6 on page 404 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.] √

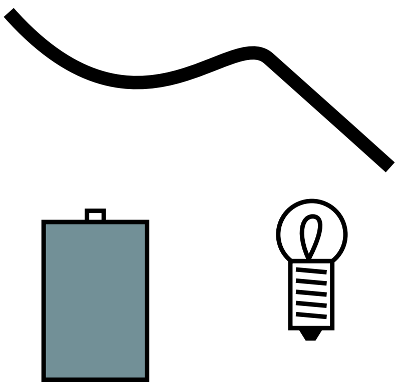
11 Today, even a big luxury car like a Cadillac can have an electrical system that is relatively low in power, since it doesn't need to do much more than run headlights, power windows, etc. In the near future, however, manufacturers plan to start making cars with electrical systems about five times more 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? ✓

12 (a) You take an LP record out of its sleeve, and it acquires a static charge of 1 nC. You play it at the normal speed of $33\frac{1}{3}$ r.p.m., and the charge moving in a circle creates an electric current. What is the current, in amperes? ✓

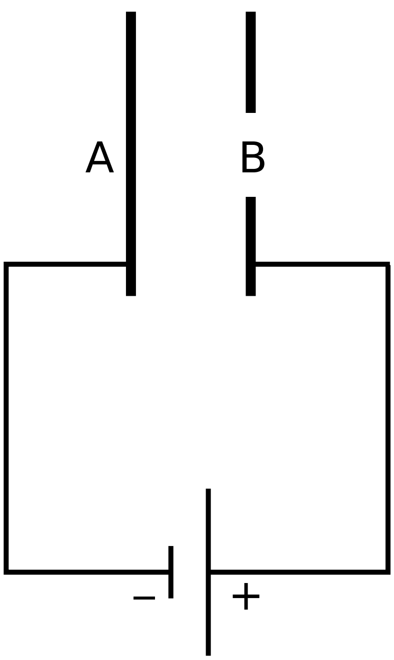
(b) Although the planetary model of the atom can be made to work with any value for the radius of the electrons' orbits, more advanced models that we will study later in this course predict definite radii. If the electron is imagined as circling around the proton at a speed of 2.2×10^6 m/s, in an orbit with a radius of 0.05 nm, what electric current is created? ✓

13 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?



Problem 14.

14 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. [Problem by Arnold Arons.]



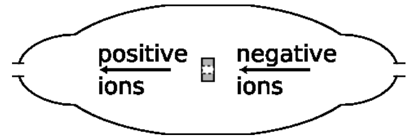
Problem 15.

15 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.) ✓

(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, p. 716 ✓

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? ✓
(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. ✓



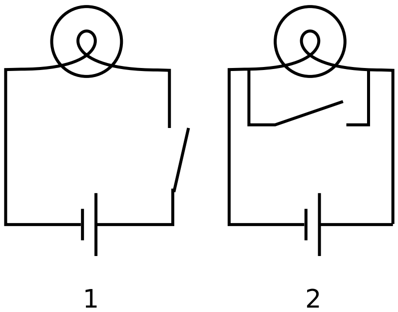
Problem 16.

17 What resistance values can be created by combining a 1 kΩ resistor and a 10 kΩ resistor? ▷ Solution, p. 716

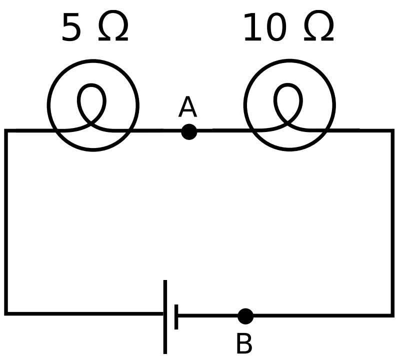
18 (a) Many battery-operated devices take more than one battery. If you look closely in the battery compartment, you will see that the batteries are wired in series. Consider a flashlight circuit. What does the loop rule tell you about the effect of putting several batteries in series in this way?
(b) 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?

19 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.
(c) Suppose they were instead hooked up in series. Draw a schematic and calculate the same things.



Problem 22.



Problem 23.

20 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?

21 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? √

22 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?

23 In the figure, the battery is 9 V.
(a) What are the voltage differences across each light bulb?
(b) What current flows through each of the three components of the circuit?
(c) If a new wire is added to connect points A and B, how will the appearances of the bulbs change? What will be the new voltages and currents?
(d) Suppose no wire is connected from A to B, but the two bulbs are switched. How will the results compare with the results from the original setup as drawn?

24 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?

25 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.

26 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.)

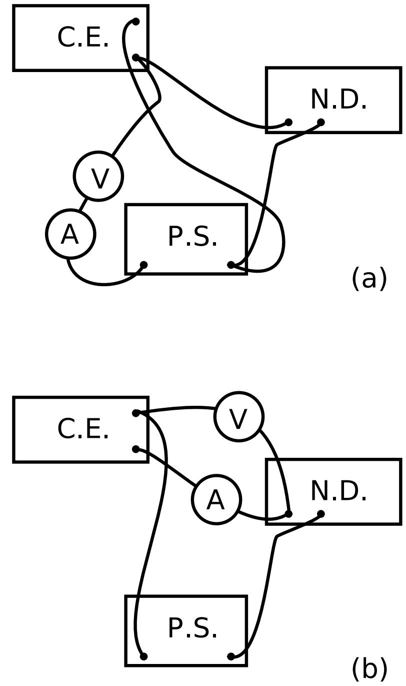
27 Suppose six identical resistors, each with resistance R, are connected so that they form the edges of a tetrahedron (a pyramid with three sides in addition to the base, i.e. one less side than an Egyptian pyramid). What resistance value or values can be obtained by making connections onto any two points on this arrangement?
▷ Solution, p. 716

28 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.

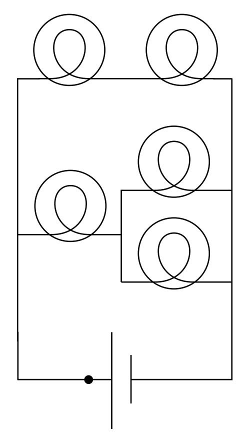
29 The figure shows a circuit containing five lightbulbs connected to a battery. Suppose you're going to connect one probe of a voltmeter to the circuit at the point marked with a dot. How many unique, nonzero voltage differences could you measure by connecting the other probe to other wires in the circuit?

30 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.

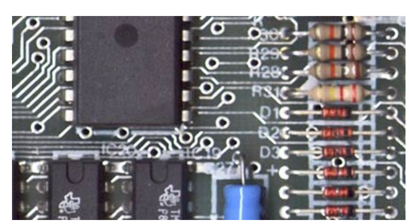
31 You have to do different things with a circuit to measure current than to measure a voltage difference. Which would be more practical for a printed circuit board, in which the wires are actually strips of metal embedded inside the board?
▷ Solution, p. **??**



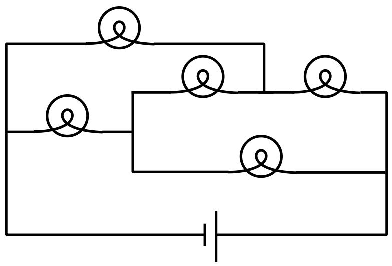
Problem 25.



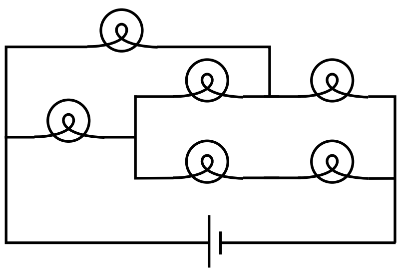
Problems 29 and 30.



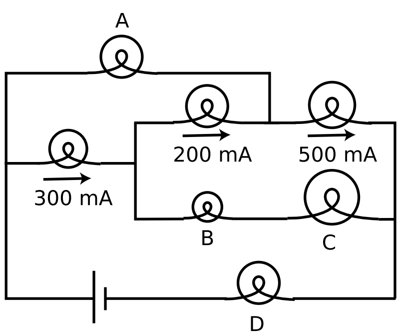
A printed circuit board, like the kind referred to in problem 31.



Problem 32.



Problem 33.



Problem 34.

32 The bulbs are all identical. Which one doesn't light up?

33 Each bulb has a resistance of one ohm. How much power is drawn from the one-volt battery?

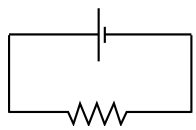
34 The bulbs all have unequal resistances. Given the three currents shown in the figure, find the currents through bulbs A, B, C, and D.

Key to symbols:
easy typical challenging difficult very difficult
√ An answer check is available at www.lightandmatter.com.

Exercises

Exercise 9A: Voltage and Current

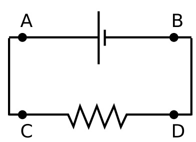
1. How many different currents could you measure in this circuit? Make a prediction, and then try it.



What do you notice? How does this make sense in terms of the roller coaster metaphor introduced in discussion question 9.1.3A on page 417?

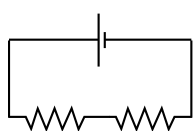
What is being *used up* in the resistor?

2. By connecting probes to these points, how many ways could you measure a voltage? How many of them would be different numbers? Make a prediction, and then do it.



What do you notice? Interpret this using the roller coaster metaphor, and color in parts of the circuit that represent constant voltages.

3. The resistors are unequal. How many *different* voltages and currents can you measure? Make a prediction, and then try it.



What do you notice? Interpret this using the roller coaster metaphor, and color in parts of the circuit that represent constant voltages.

Exercise 9B: The Loop and Junction Rules

Apparatus:

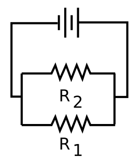
DC power supply

multimeter

resistors

1. The junction rule

Construct a circuit like this one, using the power supply as your voltage source. To make things more interesting, don't use equal resistors. Use nice big resistors (say 100 $k\Omega$ to 1 $M\Omega$) — this will ensure that you don't burn up the resistors, and that the multimeter's small internal resistance when used as an ammeter is negligible in comparison.



Insert your multimeter in the circuit to measure all three currents that you need in order to test the junction rule.

2. The loop rule

Now come up with a circuit to test the loop rule. Since the loop rule is always supposed to be true, it's hard to go wrong here! Make sure you have at least three resistors in a loop, and make sure you hook in the power supply in a way that creates non-zero voltage differences across all the resistors. Measure the voltage differences you need to measure to test the loop rule. Here it is best to use fairly small resistances, so that the multimeter's large internal resistance when used in parallel as a voltmeter will not significantly reduce the resistance of the circuit. Do not use resistances of less than about 100 Ω, however, or you may blow a fuse or burn up a resistor.

Exercise 9C: Reasoning About Circuits

The questions in this exercise can all be solved using some combination of the following approaches:

a) There is constant voltage throughout any conductor.

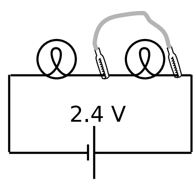
b) Ohm's law can be applied to any *part* of a circuit.

c) Apply the loop rule.

d) Apply the junction rule.

In each case, discuss the question, decide what you think is the right answer, and then try the experiment.

1. A wire is added in parallel with one bulb.

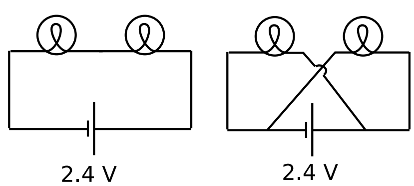


Which reasoning is correct?

- *Each bulb still has 1.2 V across it, so both bulbs are still lit up.*

- *All parts of a wire are at the same voltage, and there is now a wire connection from one side of the right-hand bulb to the other. The right-hand bulb has no voltage difference across it, so it goes out.*

2. The series circuit is changed as shown.

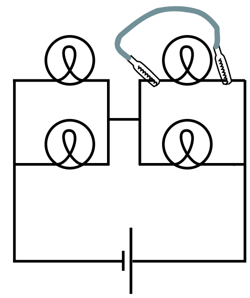


Which reasoning is correct?

- *Each bulb now has its sides connected to the two terminals of the battery, so each now has 2.4 V across it instead of 1.2 V. They get brighter.*

- *Just as in the original circuit, the current goes through one bulb, then the other. It's just that now the current goes in a figure-8 pattern. The bulbs glow the same as before.*

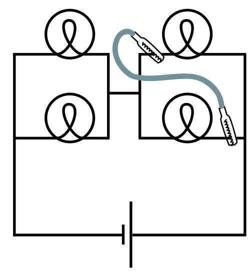
3. A wire is added as shown to the original circuit.



What is wrong with the following reasoning?

The top right bulb will go out, because its two sides are now connected with wire, so there will be no voltage difference across it. The other three bulbs will not be affected.

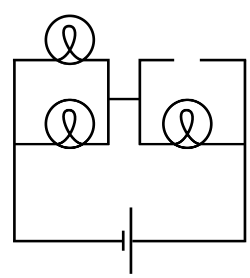
4. A wire is added as shown to the original circuit.



What is wrong with the following reasoning?

The current flows out of the right side of the battery. When it hits the first junction, some of it will go left and some will keep going up The part that goes up lights the top right bulb. The part that turns left then follows the path of least resistance, going through the new wire instead of the bottom bulb. The top bulb stays lit, the bottom one goes out, and others stay the same.

5. What happens when one bulb is unscrewed, leaving an air gap?



Chapter 10
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?"

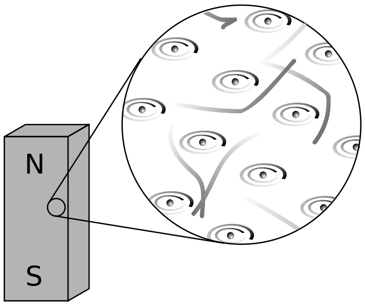
From THE REVERSIBLE REVOLUTIONS by Cecil Corwin, Cosmic Stories, March 1941. Art by Morey, Bok, Kyle, Hunt, Forte. Copyright expired.

10.1 Fields of Force

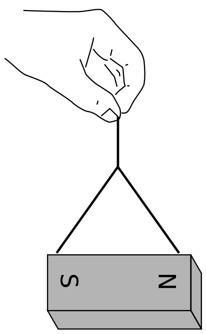
Cutting-edge science readily infiltrates popular culture, though sometimes in garbled form. The Newtonian imagination populated the universe mostly with that nice solid stuff called matter, which was made of little hard balls called atoms. In the early twentieth century, consumers of pulp fiction and 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?

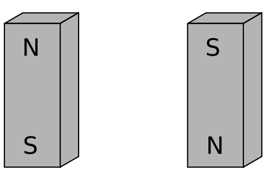
Time delays in forces exerted at a distance



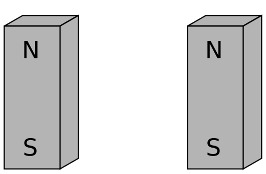
a / A bar magnet's atoms are (partially) aligned.



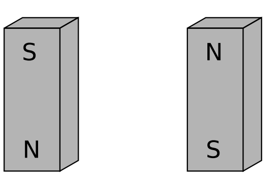
b / A bar magnet interacts with our magnetic planet.



c / Magnets aligned north-south.



d / The second magnet is reversed.



e / Both magnets are reversed.

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 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. 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). 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.[1] (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

[1] In chapter 7, we saw that as a consequence of Einstein's theory of relativity, material objects can never move faster than the speed of light. Although we haven't explicitly shown that signals or information are subject to the same limit, the reasoning of homework problem 5 on page 341 can be extended to show that causality would be violated if signals could exceed the speed of light.

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.

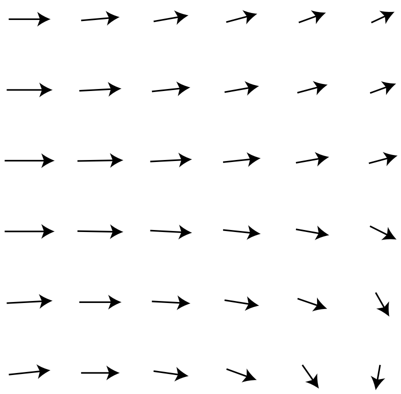
The smoking-gun argument for this strange notion of traveling force ripples comes from the fact that they carry energy.

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.

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").



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.

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.

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.

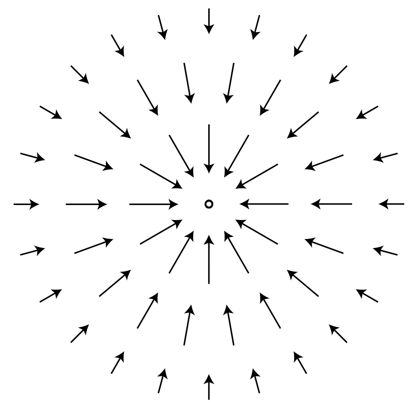
We now have three ways of representing a gravitational field. The magnitude of the gravitational field near the surface of the earth, for instance, could be written as 9.8 N/kg, 9.8 J/kg · m, or 9.8 m/s^2. If we already had two names for it, why invent a third? The main reason is that it prepares us 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.

Gravitational field of the earth *example 1*

▷ What is the magnitude of the earth's gravitational field, in terms of its mass, M, and the distance r from its center?

▷ 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.

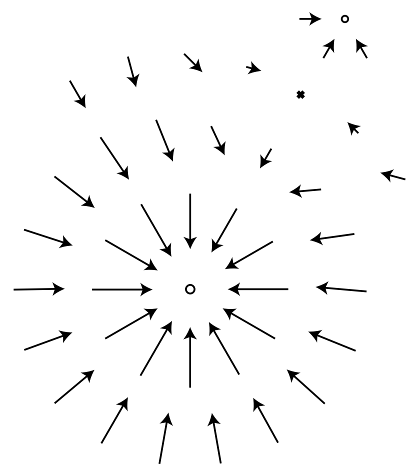


g / The gravitational field surrounding a clump of mass such as the earth.

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.



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.

Superposition of fields

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.

Reduction in gravity on Io due to Jupiter's gravity *example 2*

▷ 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:

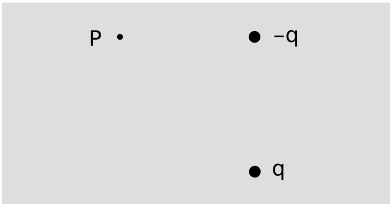
The electric field vector, $\mathbf{E}$, at any location in space is found by placing a test charge q_t at that point. The electric field vector is then given by $\mathbf{E} = \mathbf{F}/q_t$, where $\mathbf{F}$ is the electric force on the test charge.

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.

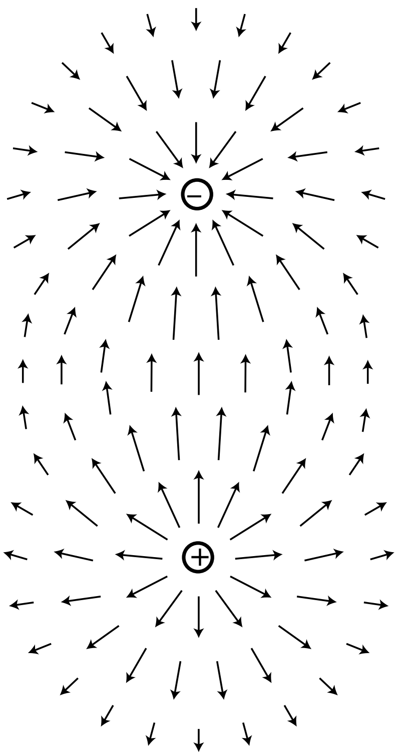
The most difficult point about the definition of the electric field is that the force on a negative charge is in the opposite direction compared to the field. This follows from the definition, since dividing a vector by a negative number reverses its direction. It's as though we had some objects that fell upward instead of down.

Self-Check

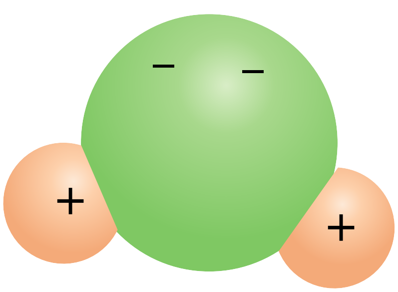
Find an equation for the magnitude of the field of a single point charge Q. ▷ Answer, p. 708



j / Example 3.



k / A dipole field. Electric fields diverge from a positive charge and converge on a negative charge.



l / A water molecule is a dipole.

Superposition of electric fields — *example 3*

▷ 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?

▷ 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

$$\begin{aligned} E_{-q,x} &= \frac{kq}{b^2} \\ E_{-q,y} &= 0 \qquad . \end{aligned}$$

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

$$\begin{aligned} E_x &= \left(1 - 2^{-3/2}\right)\frac{kq}{b^2} \\ E_y &= \frac{kq}{2^{3/2}b^2} \end{aligned}$$

Dipoles

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. A water molecule, l, is a dipole because the electrons tend to shift away from the hydrogen atoms and onto the oxygen atom.

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

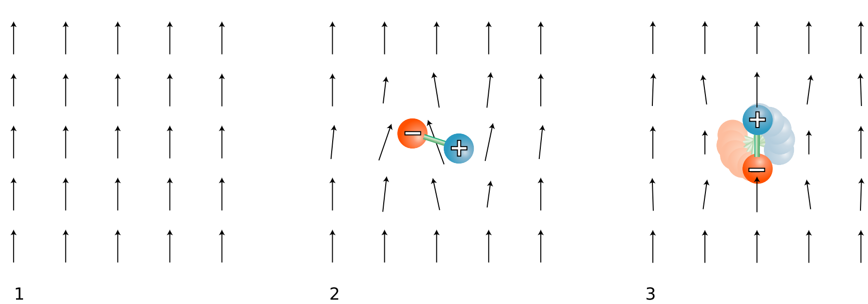
$$\begin{aligned}\tau &= \frac{\ell}{2}q|\mathbf{E}|\sin\theta + \frac{\ell}{2}q|\mathbf{E}|\sin\theta \\ &= \ell q|\mathbf{E}|\sin\theta \qquad .\end{aligned}$$

(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*,

$$\mathbf{D} = \sum q_i \mathbf{r}_i \qquad ,$$

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



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.

(page 213),

$$\boldsymbol{\tau} = \mathbf{D} \times \mathbf{E} \qquad .$$

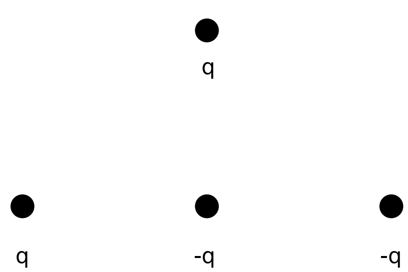
No matter how we notate it, the definition of the dipole moment requires that we choose point from which we measure all the position vectors of the charges. However, in the commonly encountered special case where the total charge of the object is zero, the dipole moment is the same regardless of this choice.

Dipole moment of a molecule of NaCl gas — *example 4*

▷ 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

$$\begin{aligned} D &= (6 \times 10^{-10}\ \text{m})(e) \\ &= (6 \times 10^{-10}\ \text{m})(1.6 \times 10^{-19}\ \text{C}) \\ &= 1 \times 10^{-28}\ \text{C} \cdot \text{m} \end{aligned}$$



n / Example 5.

Dipole moments as vectors — *example 5*

▷ The horizontal and vertical spacing between the charges in the figure is b. Find the dipole moment.

▷ 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

A In the definition of the electric field, does the test charge need to be 1 coulomb? Does it need to be positive?

B Does a charged particle such as an electron or proton feel a force from its own electric field?

C 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?

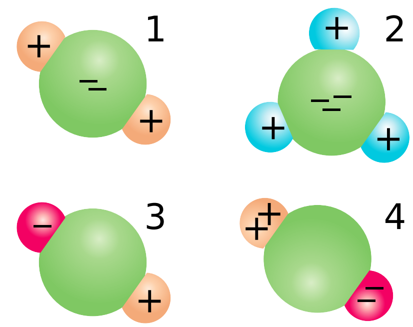
D In a flashlight powered by a battery, which way do the electric fields point? What would the fields be like inside the wires? Inside the filament of the bulb?

E Criticize the following statement: "An electric field can be represented by a sea of arrows showing how current is flowing."

F 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?

G 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.

H Compare the dipole moments of the molecules and molecular ions shown in the figure.



o / Discussion question H.

I 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?

10.2 Voltage Related to Field

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,

$$\mathrm{d}U = -F_x \mathrm{d}x \qquad ,$$

and divide by charge,

$$\frac{\mathrm{d}U}{q} = -\frac{F_x}{q}\mathrm{d}x \qquad ,$$

giving

$$\mathrm{d}V = -E_x \mathrm{d}x \qquad ,$$

or

$$\frac{\mathrm{d}V}{\mathrm{d}x} = -E_x \qquad .$$

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.

Field generated by an electric eel — *example 6*

▷ 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

$$\begin{aligned} |\mathbf{E}| &= \frac{\mathrm{d}V}{\mathrm{d}x} \\ &\approx \frac{\Delta V}{\Delta x} \qquad \text{[assumption of constant field]} \\ &= 1000\ \text{V/m} \qquad . \end{aligned}$$

Relating the units of electric field and voltage — *example 7*

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:

$$\begin{aligned} \frac{\text{V}}{\text{m}} &= \frac{\text{J/C}}{\text{m}} \\ &= \frac{\text{J}}{\text{C}\cdot\text{m}} \qquad . \end{aligned}$$

To connect joules to newtons, we recall that work equals force times distance, so $\mathrm{J} = \mathrm{N}\cdot\mathrm{m}$, so

$$\begin{aligned}\frac{\mathrm{V}}{\mathrm{m}} &= \frac{\mathrm{N}\cdot\mathrm{m}}{\mathrm{C}\cdot\mathrm{m}}\\ &= \frac{\mathrm{N}}{\mathrm{C}}\end{aligned}$$

As with other such difficulties with electrical units, one quickly begins to recognize frequently occurring combinations.

Voltage associated with a point charge — *example 8*

▷ What is the voltage associated with a point charge?

▷ As derived previously in self-check 10.1.3 on page 457, the field is

$$|\mathbf{E}| = \frac{kQ}{r^2}$$

The difference in voltage between two points on the same radius line is

$$\begin{aligned}\Delta V &= -\int \mathrm{d}V\\ &= -\int E_x \mathrm{d}x\end{aligned}$$

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.

$$\begin{aligned}\Delta V &= -\int_{r_1}^{r_2} E_r \mathrm{d}r\\ &= -\int_{r_1}^{r_2} \frac{kQ}{r^2}\mathrm{d}r\\ &= \frac{kQ}{r}\bigg]_{r_1}^{r_2}\\ &= \frac{kQ}{r_2} - \frac{kQ}{r_1} \qquad .\end{aligned}$$

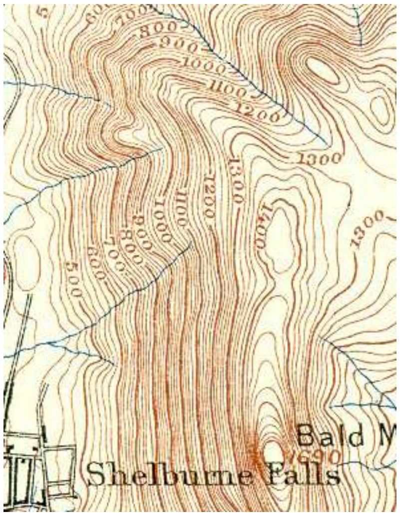
The standard convention is to use $r_1 = \infty$ as a reference point, so that the voltage at any distance r from the charge is

$$V = \frac{kQ}{r} \qquad .$$

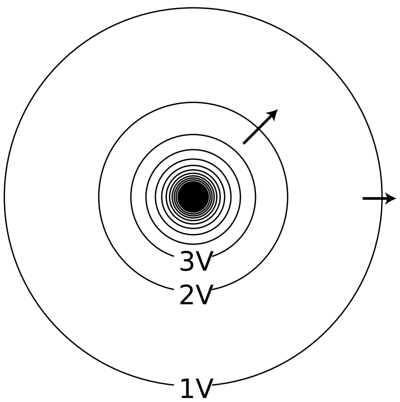
The interpretation is that if you bring a positive test charge closer to a positive charge, its electrical energy is increased; if it was released, it would spring away, releasing this as kinetic energy.

Self-Check

Show that you can recover the expression for the field of a point charge by evaluating the derivative $E_x = -\mathrm{d}V/\mathrm{d}x$. ▷ Answer, p. 708



a / A topographical map of Shelburne Falls, Mass. (*USGS*)



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.

10.2.2 Two or three dimensions

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 $\mathrm{d}V/\mathrm{d}x$ (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.

The one-dimensional relationship $E = -\mathrm{d}V/\mathrm{d}x$ generalizes to three dimensions as follows:

$$E_x = -\frac{\mathrm{d}V}{\mathrm{d}x}$$
$$E_y = -\frac{\mathrm{d}V}{\mathrm{d}y}$$
$$E_z = -\frac{\mathrm{d}V}{\mathrm{d}z}$$

This can be notated as a gradient (page 161),

$$\mathbf{E} = \nabla V \qquad ,$$

and if we know the field and want to find the voltage, we can use a line integral,

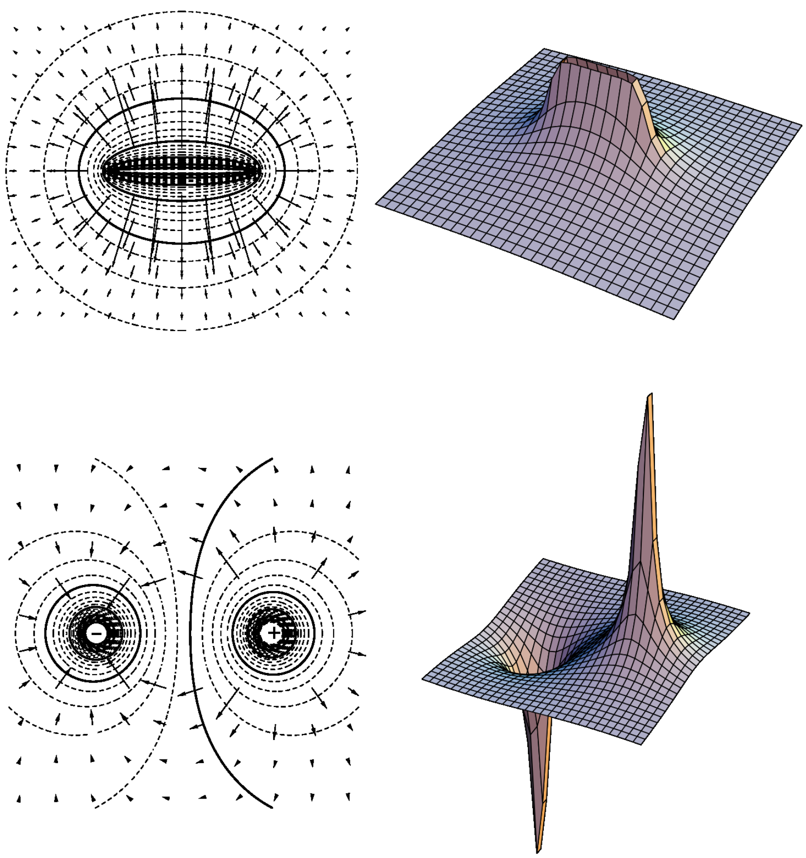
$$\Delta V = \int_C \mathbf{E} \cdot \mathrm{d}\mathbf{r} \qquad ,$$

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 $\mathrm{d}V/\mathrm{d}x$ and $\mathrm{d}V/\mathrm{d}y$, 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, p. 709

Figure c shows some examples of ways to visualize field and voltage patterns.

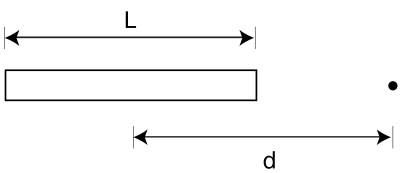


c / Two-dimensional field and voltage patterns. Top: A uniformly charged rod. Bottom: A dipole. In each case, the diagram on the left shows the field vectors and constant-voltage curves, while the one on the right shows the voltage (up-down coordinate) as a function of x and y. Interpreting the field 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.

10.3 Fields by Superposition

10.3.1 Electric field of a continuous charge distribution

Charge really comes in discrete chunks, but often it is mathematically convenient to treat a set of charges as if they were like a continuous fluid spread throughout a region of space. For example, a charged metal ball will have charge spread nearly uniformly all over its surface, and for most purposes it will make sense to ignore the fact that this uniformity is broken at the atomic level. The electric field made by such a continuous charge distribution is the sum of the fields created by every part of it. If we let the "parts" become infinitesimally small, we have a sum of an infinitely many infinitesimal numbers: an integral. If it was a discrete sum, as in example 3 on page 458, we would have a total electric field in the x direction that was the sum of all the x components of the individual fields, and similarly we'd have sums for the y and z components. In the continuous case, we have three integrals. Let's keep it simple by starting with a one-dimensional example.



a / Example 9.

Field of a uniformly charged rod *example 9*

▷ A rod of length L has charge Q spread uniformly along it. Find the electric field at a point a distance d from the center of the rod, along the rod's axis.

▷ This is a one-dimensional situation, so we really only need to do a single integral representing the total field along the axis. We imagine breaking the rod down into short pieces of length dz, each with charge dq. Since charge is uniformly spread along the rod, we have $\mathrm{d}q = \lambda \mathrm{d}z$, where $\lambda = Q/L$ (Greek lambda) is the charge per unit length, in units of coulombs per meter. Since the pieces are infinitesimally short, we can treat them as point charges and use the expression $k\mathrm{d}q/r^2$ for their contributions to the field, where $r = d - z$ is the distance from the charge at z to the point in which we are interested.

$$\begin{aligned} E_z &= \int \frac{k\mathrm{d}q}{r^2} \\ &= \int_{-L/2}^{+L/2} \frac{k\lambda \mathrm{d}z}{r^2} \\ &= k\lambda \int_{-L/2}^{+L/2} \frac{\mathrm{d}z}{(d-z)^2} \end{aligned}$$

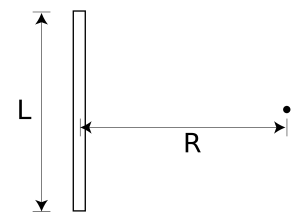
The integral can be looked up in a table, or reduced to an elementary form by substituting a new variable for $d - z$. The result is

$$\begin{aligned} E_z &= k\lambda \left(\frac{1}{d-z}\right)_{-L/2}^{+L/2} \\ &= \frac{kQ}{L}\left(\frac{1}{d-L/2} - \frac{1}{d+L/2}\right) \end{aligned} \quad .$$

For large values of d, this expression gets smaller for two reasons: (1) the denominators of the fractions become large, and (2) the two fractions become nearly the same, and tend to cancel out. This makes sense, since the field should get weaker as we get farther away from the charge. In fact, the field at large distances must approach kQ/d^2 (homework problem 17).

It's also interesting to note that the field becomes infinite at the ends of the rod, but is not infinite on the interior of the rod. Can you explain physically why this happens?

Example 9 was one-dimensional. In the general three-dimensional case, we might have to integrate all three components of the field. However, there is a trick that lets us avoid this much complication. The voltage is a scalar, so we can find the voltage by doing just a single integral, then use the voltage to find the field.



b / Example 10.

Voltage, then field — *example 10*

▷ A rod of length L is uniformly charged with charge Q. Find the field at a point lying in the midplane of the rod at a distance R.

▷ By symmetry, the field has only a radial component, E_R, pointing directly away from the rod (or toward it for $Q<0$). The brute-force approach, then, would be to evaluate the integral $E=\int|\mathrm{d}\mathbf{E}|\cos\theta$, where $\mathrm{d}\mathbf{E}$ is the contribution to the field from a charge $\mathrm{d}q$ at some point along the rod, and θ is the angle $\mathrm{d}\mathbf{E}$ makes with the radial line.

It's easier, however, to find the voltage first, and then find the field from the voltage. Since the voltage is a scalar, we simply integrate the contribution $\mathrm{d}V$ from each charge $\mathrm{d}q$, without even worrying about angles and directions. Let z be the coordinate that measures distance up and down along the rod, with $z=0$ at the center of the rod. Then the distance between a point z on the rod and the point of interest is $r=\sqrt{z^2+R^2}$, and we have

$$\begin{aligned}
V &= \int \frac{k\,\mathrm{d}q}{r} \\
&= k\lambda \int_{-L/2}^{+L/2} \frac{\mathrm{d}z}{r} \\
&= k\lambda \int_{-L/2}^{+L/2} \frac{\mathrm{d}z}{\sqrt{z^2+R^2}}
\end{aligned}$$

The integral can be looked up in a table, or evaluated using computer software:

$$\begin{aligned}
V &= k\lambda \ln\left(z+\sqrt{z^2+R^2}\right)\Big|_{-L/2}^{+L/2} \\
&= k\lambda \ln\left(\frac{L/2+\sqrt{L^2/4+R^2}}{-L/2+\sqrt{L^2/4+R^2}}\right)
\end{aligned}$$

The expression inside the parentheses can be simplified a little. Leaving out some tedious algebra, the result is

$$V = 2k\lambda \ln\left(\frac{L}{2R}+\sqrt{1+\frac{L^2}{4R^2}}\right)$$

This can readily be differentiated to find the field:

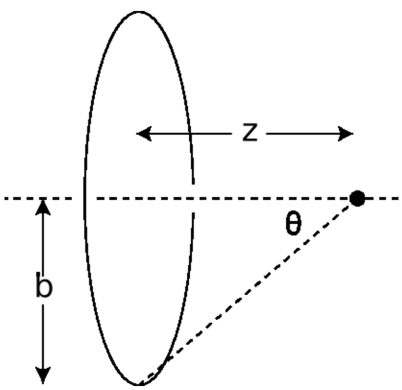
$$E_R = -\frac{dV}{dR}$$
$$= (-2k\lambda)\frac{-L/2R^2 + (1/2)(1 + L^2/4R^2)^{-1/2}(-L^2/2R^3)}{L/2R + (1 + L^2/4R^2)^{1/2}} \quad ,$$

or, after some simplification,

$$E_R = \frac{k\lambda L}{R^2\sqrt{1 + L^2/4R^2}}$$

For large values of R, the square root approaches one, and we have simply $E_R \approx k\lambda L/R^2 = kQ/R^2$. In other words, the field very far away is the same regardless of whether the charge is a point charge or some other shape like a rod. This is intuitively appealing, and doing this kind of check also helps to reassure one that the final result is correct.

The preceding example, although it involved some messy algebra, required only straightforward calculus, and no vector operations at all, because we only had to integrate a scalar function to find the voltage. The next example is one in which we can integrate either the field or the voltage without too much complication.



c / Example 11.

On-axis field of a ring of charge *example 11*

▷ Find the voltage and field along the axis of a uniformly charged ring.

▷ Integrating the voltage is straightforward.

$$V = \int \frac{k\,dq}{r}$$
$$= k\int \frac{dq}{\sqrt{b^2 + z^2}}$$
$$= \frac{k}{\sqrt{b^2 + z^2}}\int dq$$
$$= \frac{kQ}{\sqrt{b^2 + z^2}} \quad ,$$

where Q is the total charge of the ring. This result could have been derived without calculus, since the distance r is the same for every point around the ring, i.e. the integrand is a constant. It would also be straightforward to find the field by differentiating this expression with respect to z (homework problem 18).

Instead, let's see how to find the field by direct integration. By symmetry, the field at the point of interest can have only a component along the axis of symmetry, the z axis:

$$E_x = 0$$
$$E_y = 0$$

To find the field in the z direction, we integrate the z components contributed to the field by each infinitesimal part of the ring.

$$E_z = \int dE_z$$
$$= \int |d\mathbf{E}| \cos\theta \quad ,$$

where θ is the angle shown in the figure.

$$\begin{aligned} E_z &= \int \frac{k\mathrm{d}q}{r^2} \cos\theta \\ &= k \int \frac{\mathrm{d}q}{b^2+z^2} \cos\theta \end{aligned}$$

Everything inside the integral is a constant, so we have

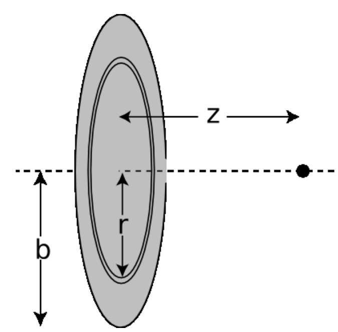
$$\begin{aligned} E_z &= \frac{k}{b^2+z^2} \cos\theta \int \mathrm{d}q \\ &= \frac{kQ}{b^2+z^2} \cos\theta \\ &= \frac{kQ}{b^2+z^2} \frac{z}{r} \\ &= \frac{kQz}{\left(b^2+z^2\right)^{3/2}} \end{aligned}$$

In all the examples presented so far, the charge has been confined to a one-dimensional line or curve. Although it is possible, for example, to put charge on a piece of wire, it is more common to encounter practical devices in which the charge is distributed over a two-dimensional surface, as in the flat metal plates used in Thomson's experiments. Mathematically, we can approach this type of calculation with the divide-and-conquer technique: slice the surface into lines or curves whose fields we know how to calculate, and then add up the contributions to the field from all these slices. In the limit where the slices are imagined to be infinitesimally thin, we have an integral.

Field of a uniformly charged disk — *example 12*

▷ A circular disk is uniformly charged. (The disk must be an insulator; if it was a conductor, then the repulsion of all the charge would cause it to collect more densely near the edge.) Find the field at a point on the axis, at a distance z from the plane of the disk.

▷ We're given that every part of the disk has the same charge per unit area, so rather than working with Q, the total charge, it will be easier to use the charge per unit area, conventionally notated σ (Greek sigma), $\sigma = Q/\pi b^2$.



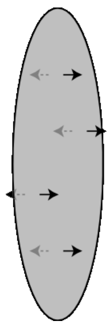
d / Example 12: geometry.

Since we already know the field due to a ring of charge, we can solve the problem by slicing the disk into rings, with each ring extending from r to $r+\mathrm{d}r$. The area of such a ring equals its circumference multiplied by its width, i.e. $2\pi r\mathrm{d}r$, so its charge is $\mathrm{d}q = 2\pi\sigma r\mathrm{d}r$, and from the result of example 11, its contribution to the field is

$$\begin{aligned} \mathrm{d}E_z &= \frac{kz\mathrm{d}q}{\left(r^2+z^2\right)^{3/2}} \\ &= \frac{2\pi\sigma kzr\mathrm{d}r}{\left(r^2+z^2\right)^{3/2}} \end{aligned}$$

The total field is

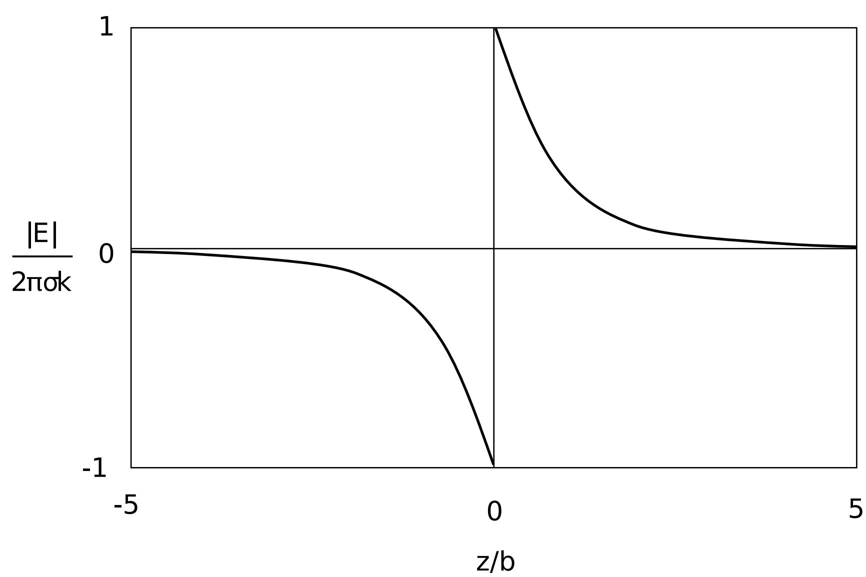
$$\begin{aligned}
E_z &= \int \mathrm{d}E_z \\
&= 2\pi\sigma kz \int_0^b \frac{r\,\mathrm{d}r}{\left(r^2+z^2\right)^{3/2}} \\
&= 2\pi\sigma kz \left. \frac{-1}{\sqrt{r^2+z^2}} \right|_{r=0}^{r=b} \\
&= 2\pi\sigma k \left(1 - \frac{z}{\sqrt{b^2+z^2}}\right)
\end{aligned}$$



e / Example 12: the field on both sides (for $\sigma > 0$).

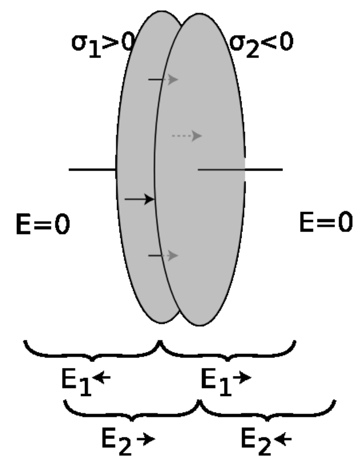
The result of example 12 has some interesting properties. First, we note that it was derived on the unspoken assumption of $z > 0$. By symmetry, the field on the other side of the disk must be equally strong, but in the opposite direction, as shown in figures e and f. Thus there is a discontinuity in the field at $z = 0$. In reality, the disk will have some finite thickness, and the switching over of the field will be rapid, but not discontinuous.

At large values of z, i.e. $z \gg b$, the field rapidly approaches the $1/r^2$ variation that we expect when we are so far from the disk that the disk's size and shape cannot matter (homework problem 17).



f / Example 12: variation of the field ($\sigma > 0$).

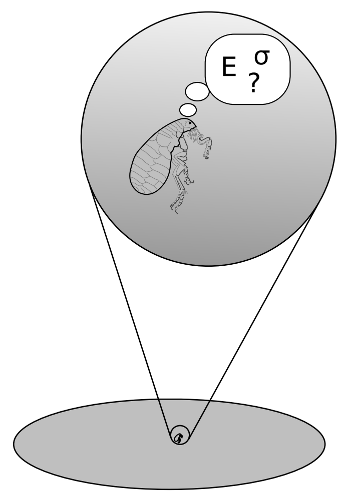
A practical application is the case of a capacitor, g, having two parallel circular plates very close together. In normal operation, the charges on the plates are opposite, so one plate has fields pointing into it and the other one has fields pointing out. In a real capacitor, the plates are a metal conductor, not an insulator, so the charge will tend to arrange itself more densely near the edges, rather than spreading itself uniformly on each plate. Furthermore, we have only calculated the *on-axis* field in example 12; in the off-axis region, each disk's contribution to the field will be weaker, and it will also point away from the axis a little. But if we are willing to ignore these complications for the sake of a rough analysis, then the fields superimpose as shown in figure g: the fields cancel the outside of the capacitor, but between the plates its value is double that contributed by a single plate. This cancellation on the outside is a very useful property for a practical capacitor. For instance, if you look at the printed circuit board in a typical piece of consumer electronics, there are many capacitors, often placed fairly close together. If their exterior fields didn't cancel out nicely, then each capacitor would interact with its neighbors in a complicated way, and the behavior of the circuit would depend on the exact physical layout, since the interaction would be stronger or weaker depending on distance. In reality, a capacitor does create weak external electric fields, but their effects are often negligible, and we can then use the *lumped-circuit approximation*, which states that each component's behavior depends only on the currents that flow in and out of it, not on the interaction of its fields with the other components.



g / A capacitor consisting of two disks with opposite charges.

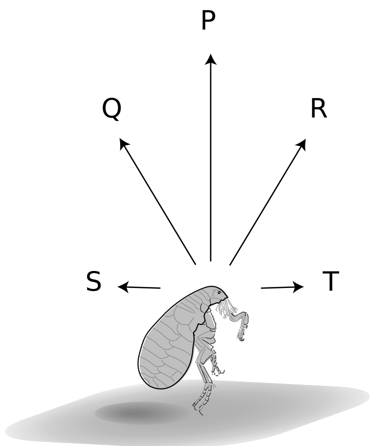
10.3.2 The field near a charged surface

From a theoretical point of view, there is something even more intriguing about example 12: the magnitude of the field for small values of z ($z \ll b$) is $E = 2\pi k\sigma$, which doesn't depend on b at all for a fixed value of σ. If we made a disk with twice the radius, and covered it with the same number of coulombs per square meter (resulting in a total charge four times as great), the field close to the disk would be unchanged! That is, a flea living near the center of the disk, h, would have no way of determining the size of her flat "planet" by measuring the local field and charge density. (Only by leaping off the surface into outer space would she be able to measure fields that were dependent on b. If she traveled very far, to $z \gg b$, she would be in the region where the field is well approximated by $|\mathbf{E}| \approx kQ/z^2 = k\pi b^2\sigma/z^2$, which she could solve for b.)



h / Close to the surface, the relationship between E and σ is a fixed one, regardless of the geometry. The flea can't determine the size or shape of her world by comparing E and σ.

What is the reason for this surprisingly simple behavior of the field? Is it a piece of mathematical trivia, true only in this particular case? What if the shape was a square rather than a circle? In other words, the flea gets no information about the *size* of the disk from measuring E, since $E = 2\pi k\sigma$, independent of b, but what if she didn't know the *shape*, either? If the result for a square had some other geometrical factor in front instead of 2π, then she could tell which shape it was by measuring E. The surprising mathematical fact, however, is that the result for a square, indeed for any shape whatsoever, is $E = 2\pi\sigma k$. It doesn't even matter whether the surface is flat or warped, or whether the density of charge is different at parts of the surface which are far away compared to the flea's distance above the surface.



i / Fields contributed by nearby parts of the surface, P, Q, and R, contribute to $E_\perp$. Fields due to distant charges, S, and T, have very small contributions to $E_\perp$ because of their shallow angles.

This universal $E_\perp = 2\pi k\sigma$ field perpendicular to a charged surface can be proved mathematically based on Gauss's law[2] (section 10.6), but we can understand what's happening on qualitative grounds. Suppose on night, while the flea is asleep, someone adds more surface area, also positively charged, around the outside edge of her disk-shaped world, doubling its radius. The added charge, however, has very little effect on the field in her environment, as long as she stays at low altitudes above the surface. As shown in figure i, the new charge to her west contributes a field, T, that is almost purely "horizontal" (i.e. parallel to the surface) and to the east. It has a negligible upward component, since the angle is so shallow. This new eastward contribution to the field is exactly canceled out by the westward field, S, created by the new charge to her east. There is likewise almost perfect cancellation between any other pair of opposite compass directions.

A similar argument can be made as to the shape-independence of the result, as long as the shape is symmetric. For example, suppose that the next night, the tricky real estate developers decide to add corners to the disk and transform it into a square. Each corner's contribution to the field measured at the center is canceled by the field due to the corner diagonally across from it.

What if the flea goes on a trip away from the center of the disk? The perfect cancellation of the "horizontal" fields contributed by distant charges will no longer occur, but the "vertical" field (i.e. the field perpendicular to the surface) will still be $E_\perp = 2\pi k\sigma$, where σ is the local charge density, since the distant charges can't contribute to the vertical field. The same result applies if the shape of the surface is asymmetric, and doesn't even have any well-defined geometric center: the component perpendicular to the surface is $E_\perp = 2\pi k\sigma$, but we may have $E_\parallel \neq 0$. All of the above arguments can be made more rigorous by discussing mathematical limits rather than using words like "very small." There is not much point in giving a rigorous proof here, however, since we will be able to demonstrate this fact as a corollary of Gauss' Law in section 10.6. The result is as follows:

At a point lying a distance z from a charged surface, the component of the electric field perpendicular to the surface obeys

$$\lim_{z\rightarrow 0} E_\perp = 2\pi k\sigma \qquad ,$$

where σ is the charge per unit area. This is true regardless of the shape or size of the surface.

[2] rhymes with "mouse"

The field near a point, line, or surface charge *example 13*

▷ Compare the variation of the electric field with distance, d, for small values of d in the case of a point charge, an infinite line of charge, and an infinite charged surface.

▷ For a point charge, we have already found $E \propto d^{-2}$ for the magnitude of the field, where we are now using d for the quantity we would ordinarily notate as r. This is true for all values of d, not just for small d — it has to be that way, because the point charge has no size, so if E behaved differently for small and large d, there would be no way to decide what d had to be small or large relative to.

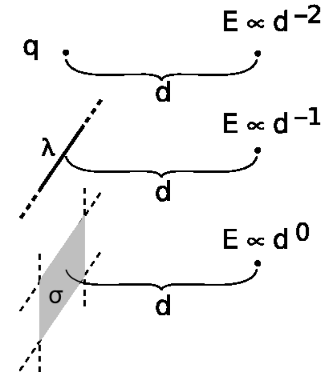
For a line of charge, the result of example 10 is

$$E = \frac{k\lambda L}{d^2\sqrt{1 + L^2/4d^2}} \qquad .$$

In the limit of $d \ll L$, the quantity inside the square root is dominated by the second term, and we have $E \propto d^{-1}$.

Finally, in the case of a charged surface, the result is simply $E = 2\pi\sigma k$, or $E \propto d^0$.

Notice the lovely simplicity of the pattern, as shown in figure j. A point is zero-dimensional: it has no length, width, or breadth. A line is one-dimensional, and a surface is two-dimensional. As the dimensionality of the charged object changes from 0 to 1, and then to 2, the exponent in the near-field expression goes from 2 to 1 to 0.

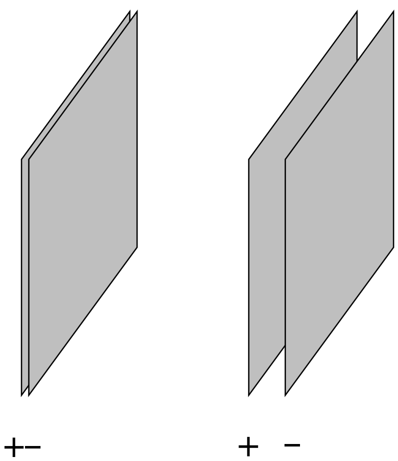


j / Example 13.

10.4 Energy in Fields

10.4.1 Electric field energy

Fields possess energy, as argued on page 453, but how much energy? The answer can be found using the following elegant approach. We assume that the electric energy contained in an infinitesimal volume of space $\mathrm{d}v$ is given by $\mathrm{d}U_e = f(\mathbf{E})\mathrm{d}v$, where f is some function, which we wish to determine, of the field $\mathbf{E}$. It might seem that we would have no easy way to determine the function f, but many of the functions we could cook up would violate the symmetry of space. For instance, we could imagine $f(\mathbf{E}) = aE_y$, where a is some constant with the appropriate units. However, this would violate the symmetry of space, because it would give the y axis a different status from x and z. As discussed on page 158, if we wish to calculate a scalar based on some vectors, the dot product is the only way to do it that has the correct symmetry properties. If all we have is one vector, $\mathbf{E}$, then the only scalar we can form is $\mathbf{E}\cdot\mathbf{E}$, which is the square of the magnitude of the electric field vector.



a / Two oppositely charged capacitor plates are pulled apart.

In principle, the energy function we are seeking could be proportional to $\mathbf{E}\cdot\mathbf{E}$, or to any function computed from it, such as $\sqrt{\mathbf{E}\cdot\mathbf{E}}$ or $(\mathbf{E}\cdot\mathbf{E})^7$. On physical grounds, however, the only possibility that works is $\mathbf{E}\cdot\mathbf{E}$. Suppose, for instance, that we pull apart two oppositely charged capacitor plates, as shown in figure a. We are doing work by pulling them apart against the force of their electrical attraction, and this quantity of mechanical work equals the increase in electrical energy, U_e. Using our previous approach to energy, we would have thought of U_e as a quantity which depended on the distance of the positive and negative charges from each other, but now we're going to imagine U_e as being stored within the electric field that exists in the space between and around the charges. When the plates are touching, their fields cancel everywhere, and there is zero electrical energy. When they are separated, there is still approximately zero field on the outside, but the field between the plates is nonzero, and holds some energy. Now suppose we carry out the whole process, but with the plates carrying double their previous charges. Since Coulomb's law involves the product q_1q_2 of two charges, we have quadrupled the force between any given pair of charged particles, and the total attractive force is therefore also four times greater than before. This means that the work done in separating the plates is four times greater, and so is the energy U_e stored in the field. The field, however, has merely been doubled at any given location: the electric field $\mathbf{E}_+$ due to the positively charged plate is doubled, and similarly for the contribution $\mathbf{E}_-$ from the negative one, so the total electric field $\mathbf{E}_+ + \mathbf{E}_-$ is also doubled. Thus doubling the field results in an electrical energy which is four times greater, i.e. the energy density must be proportional to the square of the field, $\mathrm{d}U_e \propto (\mathbf{E}\cdot\mathbf{E})\mathrm{d}v$. For ease of notation, we write this as $\mathrm{d}U_e \propto E^2\mathrm{d}v$, or $\mathrm{d}U_e = aE^2\mathrm{d}v$, where a is a constant of proportionality. Note that we never really made use of any of the details of the geometry of figure a, so the reasoning is of general validity. In other words, not only is $\mathrm{d}U_e = aE^2\mathrm{d}v$ the function that works in this particular case, but there is every reason to believe that it would work in other cases as well.

It now remains only to find a. Since the constant must be the same in all situations, we only need to find one example in which we can compute the field and the energy, and then we can determine a. The situation shown in figure a is just about the easiest example to analyze. We let the square capacitor plates be uniformly covered with charge densities $+\sigma$ and $-\sigma$, and we write b for the lengths of their sides. Let h be the gap between the plates after they have been separated. We choose $h \ll b$, so that the field experienced by the negative plate due to the positive plate is $E_+ = 2\pi k\sigma$. The charge of the negative plate is $-\sigma b^2$, so the magnitude of the force attracting it back toward the positive plate is (force) = (charge)(field) = $2\pi k\sigma^2 b^2$. The amount of work done in separating the plates is (work) = (force)(distance) = $2\pi k\sigma^2 b^2 h$. This is the amount of energy that has been stored in the field between the two plates, $U_e = 2\pi k\sigma^2 b^2 h = 2\pi k\sigma^2 v$, where v is the volume of the region between the plates.

We want to equate this to $U_e = aE^2 v$. (We can write U_e and v rather than $\mathrm{d}U_e$ and $\mathrm{d}v$, since the field is constant in the region between the plates.) The field between the plates has contributions from both plates, $E = E_+ + E_- = 4\pi k\sigma$. (We only used half this value in the computation of the work done on the moving plate, since the moving plate can't make a force on itself. Mathematically, each plate is in a region where its own field is reversing directions, so we can think of its own contribution to the field as being zero within itself.) We then have $aE^2 v = a \cdot 16\pi^2 k^2\sigma^2 \cdot v$, and setting this equal to $U_e = 2\pi k\sigma^2 v$ from the result of the work computation, we find $a = 1/8\pi k$. Our final result is as follows:

The electric energy possessed by an electric field $\mathbf{E}$ occupying an infinitesimal volume of space $\mathrm{d}v$ is given by

$$\mathrm{d}U_e = \frac{1}{8\pi k} E^2 \mathrm{d}v \qquad ,$$

where $E^2 = \mathbf{E}\cdot\mathbf{E}$ is the square of the magnitude of the electric field.

This is reminiscent of how waves behave: the energy content of a wave is typically proportional to the square of its amplitude.

Self-Check

We can think of the quantity $\mathrm{d}U_e/\mathrm{d}v$ as the *energy density* due to the electric field, i.e. the number of joules per cubic meter needed in order to create that field. (a) How does this quantity depend on the components of the field vector, E_x, E_y, and E_z? (b) Suppose we have a field with $E_x \neq 0$, E_y=0, and E_z=0. What would happen to the energy density if we reversed the sign of E_x? ▷ Answer, p. 709

A numerical example *example 14*

▷ A capacitor has plates whose areas are 10^{-4} m^2, separated by a gap of 10^{-5} m. A 1.5-volt battery is connected across it. How much energy is sucked out of the battery and stored in the electric field between the plates? (A real capacitor typically has an insulating material between the plates whose molecules interact electrically with the charge in the plates. For this example, we'll assume that there is just a vacuum in between the plates. The plates are also typically rolled up rather than flat.)

▷ To connect this with our previous calculations, we need to find the charge density on the plates in terms of the voltage we were given. Our previous examples were based on the assumption that the gap between the plates was small compared to the size of the plates. Is this valid here? Well, if the plates were square, then the area of 10^{-4} m^2 would imply that their sides were 10^{-2} m in length. This is indeed very large compared to the gap of 10^{-5} m, so this assumption appears to be valid (unless, perhaps, the plates have some very strange, long and skinny shape).

Based on this assumption, the field is relatively uniform in the whole volume between the plates, so we can use a single symbol, E, to represent its magnitude, and the relation $E = \mathrm{d}V/\mathrm{d}x$ is equivalent to $E = \Delta V/\Delta x = (1.5\ \mathrm{V})/(\mathrm{gap}) = 1.5 \times 10^5$ V/m.

Since the field is uniform, we can dispense with the calculus, and replace $\mathrm{d}U_e = (1/8\pi k)E^2\mathrm{d}v$ with $U_e = (1/8\pi k)E^2 v$. The volume equals the area multiplied by the gap, so we have

$$\begin{aligned} U_e &= (1/8\pi k)E^2(\mathrm{area})(\mathrm{gap}) \\ &= \frac{1}{8\pi \times 9\ \times 10^9\ \mathrm{N{\cdot}m^2/C^2}}(1.5 \times 10^5\ \mathrm{V/m})^2(10^{-4}\ \mathrm{m}^2)(10^{-5}\ \mathrm{m}) \\ &= 1 \times 10^{-10}\ \mathrm{J} \end{aligned}$$

Self-Check

Show that the units in the preceding example really do work out to be joules. ▷ Answer, p. 709

Why k is on the bottom *example 15*

It may also seem strange that the constant k is in the denominator of the equation $\mathrm{d}U_e = (1/8\pi k)E^2\mathrm{d}v$. The Coulomb constant k tells us how strong electric forces are, so shouldn't it be on top? No. Consider, for instance, an alternative universe in which electric forces are twice as strong as in ours. The numerical value of k is doubled. Because k is doubled, all the electric field strengths are doubled as well, which quadruples the quantity E^2. In the expression $E^2/8\pi k$, we've quadrupled something on top and doubled something on the bottom, which makes the energy twice as big. That makes perfect sense.

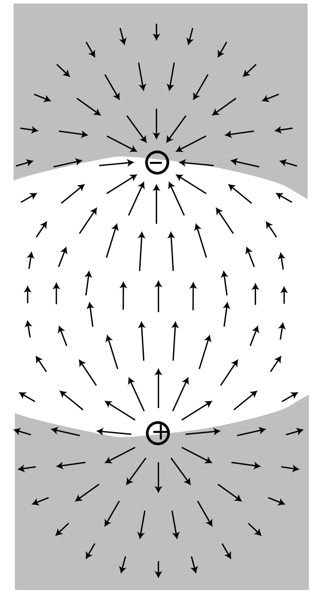
Potential energy of a pair of opposite charges — *example 16*

Imagine taking two opposite charges, b, that were initially far apart and allowing them to come together under the influence of their electrical attraction.

According to our old approach, electrical energy is lost because the electric force did positive work as it brought the charges together. (This makes sense because as they come together and accelerate it is their electrical energy that is being lost and converted to kinetic energy.)

By the new method, we must ask how the energy stored in the electric field has changed. In the region indicated approximately by the shading in the figure, the superposing fields of the two charges undergo partial cancellation because they are in opposing directions. The energy in the shaded region is reduced by this effect. In the unshaded region, the fields reinforce, and the energy is increased.

It would be quite a project to do an actual numerical calculation of the energy gained and lost in the two regions (this is a case where the old method of finding energy gives greater ease of computation), but it is fairly easy to convince oneself that the energy is less when the charges are closer. This is because bringing the charges together shrinks the high-energy unshaded region and enlarges the low-energy shaded region.



b / Example 16.

A spherical capacitor — *example 17*

▷ A spherical capacitor, c, consists of two concentric spheres of radii a and b. Find the energy required to charge up the capacitor so that the plates hold charges $+q$ and $-q$.

▷ On page 65, I proved that for *gravitational* forces, the interaction of a spherical shell of mass with other masses outside it is the same as if the shell's mass was concentrated at its center. On the interior of such a shell, the forces cancel out exactly. Since gravity and the electric force both vary as $1/r^2$, the same proof carries over immediately to electrical forces. The magnitude of the outward electric field contributed by the charge $+q$ of the central sphere is therefore

$$|\mathbf{E}_+| = \begin{cases} 0, & r < a \\ kq/r^2, & r > a \end{cases} \quad ,$$

where r is the distance from the center. Similarly, the magnitude of the *inward* field contributed by the outside sphere is

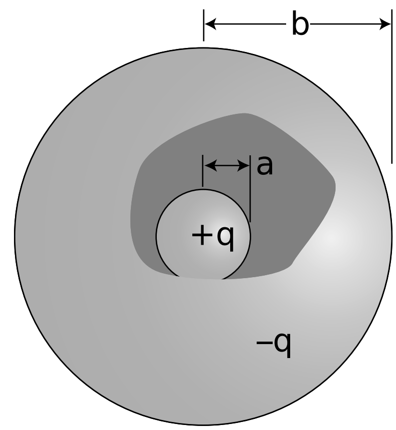
$$|\mathbf{E}_-| = \begin{cases} 0, & r < b \\ kq/r^2, & r > b \end{cases} \quad .$$

In the region outside the whole capacitor, the two fields are equal in magnitude, but opposite in direction, so they cancel. We then have for the total field

$$|\mathbf{E}| = \begin{cases} 0, & r < a \\ kq/r^2, & a < r < b \\ 0, & r > b \end{cases} \quad ,$$

so to calculate the energy, we only need to worry about the region $a < r < b$. The energy density in this region is

$$\begin{aligned} \frac{\mathrm{d}U_e}{\mathrm{d}v} &= \frac{1}{8\pi k}E^2 \\ &= \frac{kq^2}{8\pi}r^{-4} \quad . \end{aligned}$$



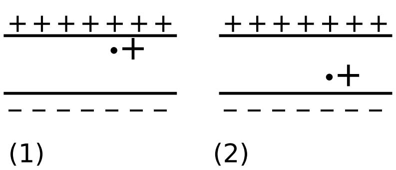
c / Example B. Part of the outside sphere has been drawn as if it is transparent, in order to show the inside sphere.

This expression only depends on r, so the energy density is constant across any sphere of radius r. We can slice the region $a < r < b$ into concentric spherical layers, like an onion, and the energy within one such layer, extending from r to $r + \mathrm{d}r$ is

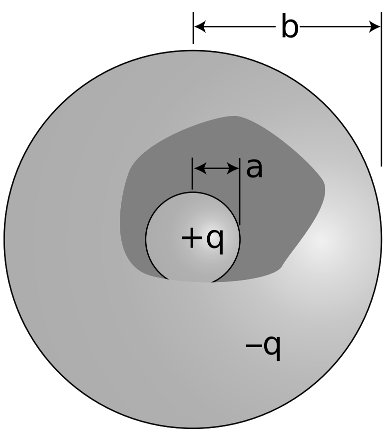
$$\begin{aligned} \mathrm{d}U_e &= \frac{\mathrm{d}U_e}{\mathrm{d}v}\mathrm{d}v \\ &= \frac{\mathrm{d}U_e}{\mathrm{d}v}(\text{area of shell})(\text{thickness of shell}) \\ &= (\frac{kq^2}{8\pi}r^{-4})(4\pi r^2)(\mathrm{d}r) \\ &= \frac{kq^2}{2}r^{-2}\mathrm{d}r \qquad . \end{aligned}$$

Integrating over all the layers to find the total energy, we have

$$\begin{aligned} U_e &= \int \mathrm{d}U_e \\ &= \int_a^b \frac{kq^2}{2}r^{-2}\mathrm{d}r \\ &= -\frac{kq^2}{2}r^{-1}\Big|_a^b \\ &= \frac{kq^2}{2}\left(\frac{1}{a}-\frac{1}{b}\right) \end{aligned}$$



d / Discussion question A.



e / Discussion question B.

Discussion Questions

A The figure shows a positive charge in the gap between two capacitor plates. Compare the energy of the electric fields in the two cases. Does this agree with what you would have expected based on your knowledge of electrical forces?

B The figure shows a spherical capacitor. In the text, the energy stored in its electric field is shown to be

$$U_e = \frac{kq^2}{2}\left(\frac{1}{a}-\frac{1}{b}\right) \qquad .$$

What happens if the difference between b and a is very small? Does this make sense in terms of the mechanical work needed in order to separate the charges? Does it make sense in terms of the energy stored in the electric field? Should these two energies be added together?

Similarly, discuss the cases of $b \to \infty$ and $a \to 0$.

C Criticize the following statement: “A solenoid makes a charge in the space surrounding it, which dissipates when you release the energy.”

D In example 16 on page 477, I argued that for the charges shown in the figure, the fields contain less energy when the charges are closer together, because the region of cancellation expanded, while the region of reinforcing fields shrank. Perhaps a simpler approach is to consider the two extreme possibilities: the case where the charges are infinitely far apart, and the one in which they are at zero distance from each other, i.e. right on top of each other. Carry out this reasoning for the case of (1) a positive charge and a negative charge of equal magnitude, (2) two positive charges of equal magnitude, (3) the gravitational energy of two equal masses.

10.4.2 Gravitational field energy

Example B depended on the close analogy between electric and gravitational forces. In fact, every argument, proof, and example discussed so far in this section is equally valid as a gravitational example, provided we take into account one fact: only positive mass exists, and the gravitational force between two masses is attractive. This is the opposite of what happens with electrical forces, which are repulsive in the case of two positive charges. As a consequence of this, we need to assign a *negative* energy density to the gravitational field! For a gravitational field, we have

$$\mathrm{d}U_g = -\frac{1}{8\pi G} g^2 \mathrm{d}v \qquad ,$$

where $g^2 = \mathbf{g} \cdot \mathbf{g}$ is the square of the magnitude of the gravitational field.

10.4.3 Magnetic field energy

So far we've only touched in passing on the topic of magnetic fields, which will deal with in detail in chapter 11. Magnetism is an interaction between moving charge and moving charge, i.e. between currents and currents. Since a current has a direction in space,[3] while charge doesn't, we can anticipate that the mathematical rule connecting a magnetic field to its source-currents will have to be completely different from the one relating the electric field to its source-charges. However, if you look carefully at the argument leading to the relation $\mathrm{d}U_e/\mathrm{d}v = E^2/8\pi k$, you'll see that these mathematical details were only necessary to the part of the argument in which we fixed the constant of proportionality. To establish $\mathrm{d}U_e/\mathrm{d}v \propto E^2$, we only had to use three simple facts:

- The field is proportional to the source.
- Forces are proportional to fields.
- Field contributed by multiple sources add like vectors.

All three of these statements are true for the magnetic field as well, so without knowing anything more specific about magnetic fields — not even what units are used to measure them! — we can state with certainty that the energy density in the magnetic field is proportional to the square of the magnitude of the magnetic field.

[3] Current is a scalar, since the definition $I = \mathrm{d}q/\mathrm{d}t$ is the derivative of a scalar. However, there is a closely related quantity called the current *density*, $\mathbf{J}$, which is a vector, and $\mathbf{J}$ is in fact the more fundamentally important quantity.

10.5 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 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.

10.5.1 Capacitance and inductance

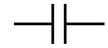
In a mechanical oscillation, energy is exchanged repetitively between potential and kinetic forms, and may also be siphoned off in the form of heat dissipated by friction. In an electrical circuit, resistors are the circuit elements that dissipate heat. What are the electrical analogs of storing and 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

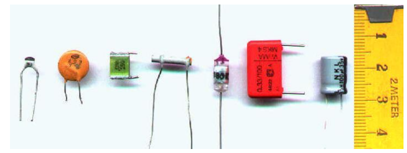
A capacitor's energy exists in its surrounding electric fields. It is proportional to the square of the field strength, which is proportional to the charges on the plates. If we assume the plates carry charges that are the same in magnitude, $+q$ and $-q$, then the energy stored in the capacitor must be proportional to q^2. For historical reasons, we write the constant of proportionality as $1/2C$,

$$U_C = \frac{1}{2C} q^2 \qquad .$$

The constant C is a geometrical property of the capacitor, called its capacitance.



a / The symbol for a capacitor.



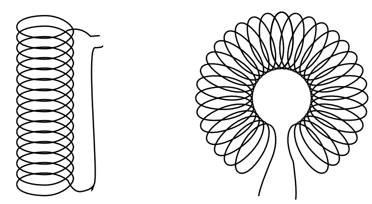
b / Some capacitors.

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 show the symbol used in schematics to represent a capacitor.

A parallel-plate capacitor *example 18*

▷ 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?

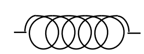
▷ Since the plates are metal, the charges on each plate are free to move, and will tend to cluster themselves more densely near the edges due to the mutual repulsion of the other charges in the same plate. However, it turns out that if the gap is small, this is a small effect, so we can get away with assuming uniform charge density on each plate. The result of example 14 then applies, and for the region between the plates, we have $E = 4\pi k\sigma = 4\pi kq/A$ and $U_e = (1/8\pi k)E^2Ah$. 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 $C = A/4\pi kh$.



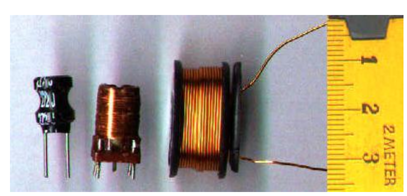
c / Two common geometries for inductors. The cylindrical shape on the left is called a solenoid.

Inductors

Any current will create a magnetic field, so in fact every current-carrying wire in a circuit acts as an inductor! However, this type of "stray" inductance is typically negligible, just as we can usually ignore the stray resistance of our wires and only take into account the actual resistors. To store any appreciable amount of magnetic energy, one usually uses a coil of wire designed specifically to be an inductor. All the loops' contribution to the magnetic field add together to make a stronger field. Unlike capacitors and resistors, practical inductors are easy to make by hand. One can for instance spool some wire around a short wooden dowel. An inductor like this, in the form cylindrical coil of wire, is called a solenoid, 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.



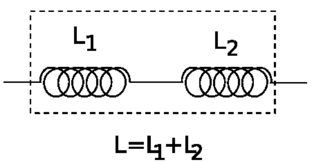
e / Some inductors.

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

$$U_L = \frac{L}{2} I^2 \qquad .$$

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 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!"

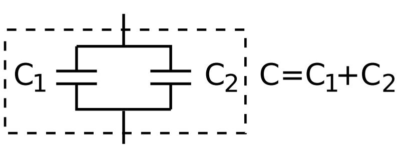
There is a lumped circuit approximation for inductors, just like the one for capacitors (p. 471). For a capacitor, this means assuming that the electric fields is completely internal, so that components only interact via currents that flow through wires, not due to the physical overlapping of their fields in space. Similarly for an inductor, the lumped circuit approximation is the assumption that the magnetic fields are completely internal.



f / Inductances in series add.

Identical inductances in series — *example 19*

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.)



g / Capacitances in parallel add.

Identical capacitances in parallel — *example 20*

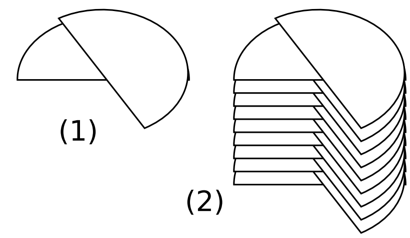
When two identical capacitances are placed in parallel, any charge deposited at the terminals of the combined double capacitor will divide itself evenly between the two parts. The electric fields surrounding each capacitor will be half the intensity, and therefore store one quarter the energy. Two capacitors, each storing one quarter the energy, give half the total energy storage. Since capacitance is inversely related to energy storage, this implies that identical capacitances in parallel give double the capacitance. In general, capacitances in parallel add. This is unlike the behavior of inductors and resistors, for which series configurations give addition.

This is consistent with the result of example 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 30 and 32.

A variable capacitor *example 21*

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.

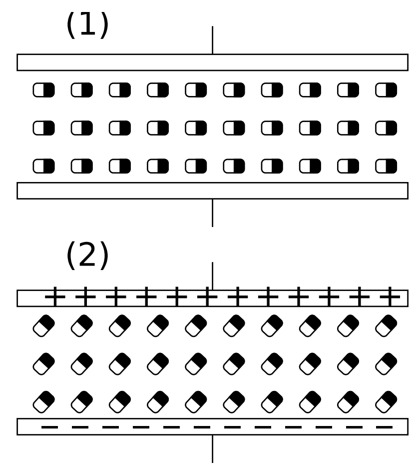
Discussion Questions

A Suppose that two parallel-plate capacitors are wired in parallel, and are placed very close together, side by side, so that the lumped circuit approximation is not very accurate. Will the resulting capacitance be too small, or too big? Could you twist the circuit into a different shape and make the effect be the other way around, or make the effect vanish? How about the case of two inductors in series?

B Most practical capacitors do not have an air gap or vacuum gap between the plates; instead, they have an insulating substance called a dielectric. We can think of the molecules in this substance as dipoles that are free to rotate (at least a little), but that are not free to move around, since it is a solid. The figure shows a highly stylized and unrealistic way of visualizing this. We imagine that all the dipoles are intially 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?

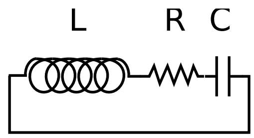
Now redo the analysis in terms of the mechanical work needed in order to charge up the plates.



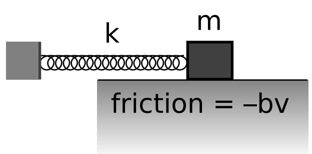
i / Discussion question B.

10.5.2 Oscillations

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.



j / A series LRC circuit.



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,

$$U_{spring} = \frac{1}{2}kx^2 \qquad (1)$$

$$K = \frac{1}{2}mv^2 \qquad . \qquad (2)$$

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 $\mathrm{d}t$, divided by $\mathrm{d}t$, $P = W/\mathrm{d}t = F\mathrm{d}x/\mathrm{d}t = Fv = -bv^2$, so

$$\text{rate of heat dissipation} = -bv^2 \qquad . \qquad (3)$$

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, p. 709

In the circuit, the stored forms of energy are

$$U_C = \frac{1}{2C}q^2 \qquad (1')$$

$$U_L = \frac{1}{2}LI^2 \qquad , \qquad (2')$$

and the rate of heat dissipation in the resistor is

$$\text{rate of heat dissipation} = -RI^2 \qquad . \qquad (3')$$

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, p. 709

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[4] $\omega \approx \sqrt{k/m}$, we can immediately solve the electrical version by analogy, giving

$$\omega \approx \frac{1}{\sqrt{LC}} \qquad .$$

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.

Tuning a radio receiver — *example 22*
A radio receiver uses this kind of circuit to pick out the desired station. Since the receiver resonates at a particular frequency, stations whose frequencies are far off will not excite any response in the circuit. The value of R has to be small enough so that only one station at a time is picked up, but big enough so that the tuner isn't too touchy. The resonant frequency can be tuned by adjusting either L or C, but variable capacitors are easier to build than variable inductors.

A numerical calculation — *example 23*
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?

[4] 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.

▷ Solving for L, we have

$$\begin{aligned}L &= \frac{1}{C\omega^2}\\ &= \frac{1}{(10^{-6}\ \mathrm{F})(2\pi\times 10^3\ \mathrm{s}^{-1})^2}\\ &= 2.5\times 10^{-3}\ \mathrm{F}^{-1}\mathrm{s}^2\end{aligned}$$

Checking that these really are the same units as henries is a little tedious, but it builds character:

$$\begin{aligned}\mathrm{F}^{-1}\mathrm{s}^2 &= (\mathrm{C}^2/\mathrm{J})^{-1}\mathrm{s}^2\\ &= \mathrm{J}\cdot\mathrm{C}^{-2}\mathrm{s}^2\\ &= \mathrm{J}/\mathrm{A}^2\\ &= \mathrm{H}\end{aligned}$$

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.

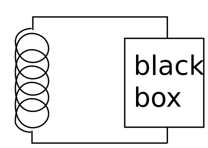
The only thing that might seem spooky here is that we used to speak as if the current in the inductor caused the magnetic field, but now it sounds as if the field causes the current. Actually this is symptomatic of the elusive nature of cause and effect in physics. It's equally valid to think of the 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 $q + \mathrm{d}q$:

$$\begin{aligned}
V_C &= (U_{q+\mathrm{d}q} - U_q)/\mathrm{d}q \\
&= \frac{\mathrm{d}U_C}{\mathrm{d}q} \\
&= \frac{\mathrm{d}}{\mathrm{d}q}\left(\frac{1}{2C}q^2\right) \\
&= \frac{q}{C}
\end{aligned}$$

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

$$\begin{aligned} |IV| &= \left|\frac{\mathrm{d}U_L}{\mathrm{d}t}\right| \\ &= \left|\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}LI^2\right)\right| \\ &= \left|LI\frac{\mathrm{d}I}{\mathrm{d}t}\right| \qquad , \end{aligned}$$

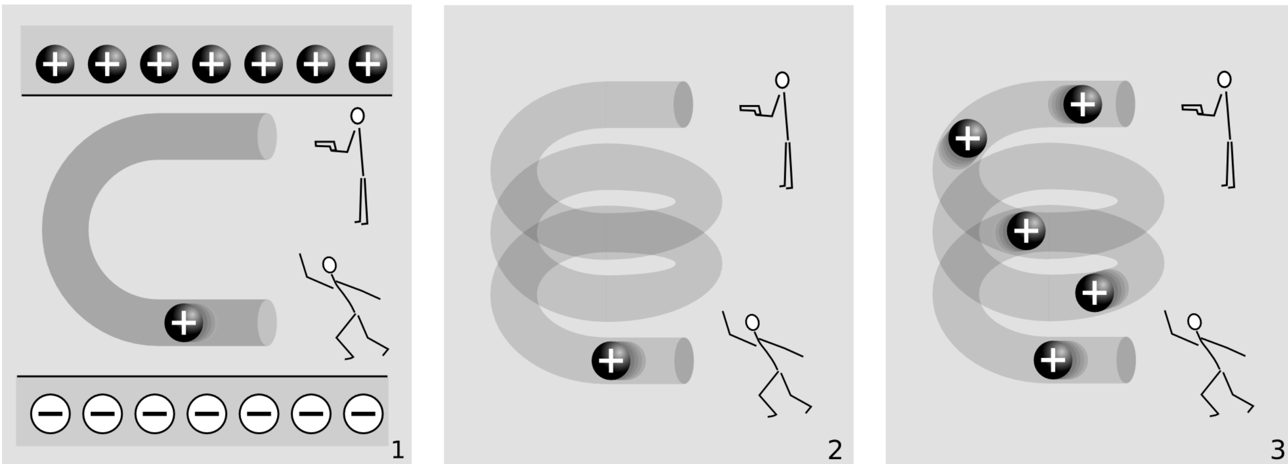
or

$$|V| = \left|L\frac{\mathrm{d}I}{\mathrm{d}t}\right| \qquad ,$$

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, $\mathrm{d}I/\mathrm{d}t$ is zero.

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.

Ballasts *example 24*

In a gas discharge tube, such as a neon sign, enough voltage is applied to a tube full of gas to ionize some of the atoms in the gas. Once ions have been created, the voltage accelerates them, and they strike other atoms, ionizing them as well and resulting in a chain reaction. This is a spark, like a bolt of lightning. But once the spark starts up, the device begins to act as though it has no resistance: more and more current flows, without the need to apply any more voltage. The power, $P = IV$, would grow without limit, and the tube would burn itself out.

The simplest solution is to connect an inductor, known as the "ballast," in series with the tube, and run the whole thing on an AC voltage. During each cycle, as the voltage reaches the point where the chain reaction begins, there is a surge of current, but the inductor resists such a sudden change of current, and the energy that would otherwise have burned out the bulb is instead channeled into building a magnetic field.

A common household fluorescent lightbulb consists of a gas discharge tube in which the glass is coated with a fluorescent material. The gas in the tube emits ultraviolet light, which is absorbed by the coating, and the coating then glows in the visible spectrum.

Until recently, it was common for a 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

A What happens when the physicist at the bottom in figure m/3 starts getting tired, and decreases the current?

10.5.4 Decay

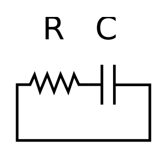
Up until now I've soft-pedaled the fact that by changing the characteristics of an oscillator, it is possible to produce non-oscillatory behavior. For example, imagine taking the mass-on-a-spring system and making the spring weaker and weaker. In the limit of small k, it's as though there was no spring 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 134.

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

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:



o / An RC circuit.

$$V_R = IR$$
$$V_C = q/C$$

The loop rule tells us

$$V_R + V_C = 0 \qquad ,$$

and combining the three equations results in a relationship between q and I:

$$I = -\frac{1}{RC}q$$

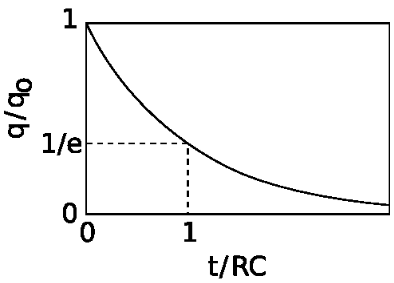
The negative sign tells us that the current tends to reduce the charge on the capacitor, i.e. to discharge it. It makes sense that the current is proportional to q: if q is large, then the attractive forces between the $+q$ and $-q$ charges on the plates of the capacitor are large, and charges will flow 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 $\mathrm{d}q/\mathrm{d}t$, we have

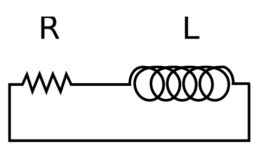
$$\frac{\mathrm{d}q}{\mathrm{d}t} = -\frac{1}{RC}q \qquad .$$

We need a function $q(t)$ whose derivative equals itself, but multiplied by a negative constant. A function of the form ae^t, where $e = 2.718...$ 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

$$q = q_\mathrm{o} \exp\left(-\frac{t}{RC}\right) \qquad .$$



p / Over a time interval RC, the charge on the capacitor is reduced by a factor of e.



q / An RL circuit.

The RL circuit

The RL circuit, q, can be attacked by similar methods, and it can easily be shown that it gives

$$I = I_\mathrm{o} \exp\left(-\frac{R}{L}t\right) \qquad .$$

The RL time constant equals L/R.

Death by solenoid; spark plugs *example 25*

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!

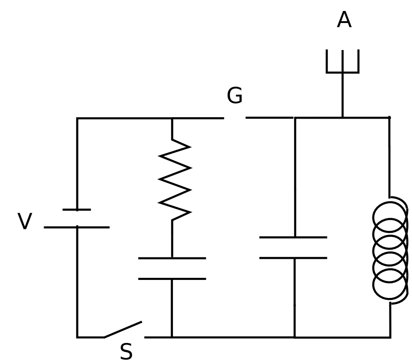
The way out of this conundrum is to recognize that the open gap in the circuit has a resistance which is large, but not infinite. This large resistance causes the RL time constant L/R to be very small. The current thus continues to flow for a very brief time, and flows straight across the air gap where 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 $V_L = \mathrm{d}I/\mathrm{d}t$: since the time constant is short, the time derivative $\mathrm{d}I/\mathrm{d}t$ is large.

This is exactly how a car's spark plugs work. Another application is to electrical safety: it can be dangerous to break an inductive circuit suddenly, because so much energy is released in a short time. There is also no guarantee that the spark will discharge across the air gap; it might go through your body instead, since your body might have a lower resistance.

A spark-gap radio transmitter *example 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).



r / Example 26.

Discussion Questions

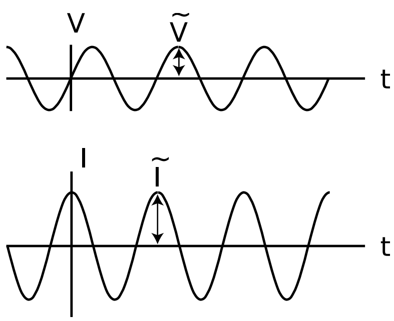
A A gopher gnaws through one of the wires in the DC lighting system in your front yard, and the lights turn off. At the instant when the circuit becomes open, we can consider the bare ends of the wire to be like the plates of a capacitor, with an air gap (or gopher gap) between them. What kind of capacitance value are we talking about here? What would this tell you about the *RC* time constant?

10.5.5 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

$$I = \frac{V}{R} \qquad .$$



s / In a capacitor, the current is 90° ahead of the voltage in phase.

In the purely capacitive case, the relation $V = q/C$ lets us calculate

$$\begin{aligned} I &= \frac{\mathrm{d}q}{\mathrm{d}t} \\ &= C\frac{\mathrm{d}V}{\mathrm{d}t} \qquad . \end{aligned}$$

If the voltage varies as, for example, $V(t) = \tilde{V}\sin(\omega t)$, then the current will be $I(t) = \omega C\tilde{V}\cos(\omega t)$, so the maximum current is $\tilde{I} = \omega C\tilde{V}$. By analogy with Ohm's law, we can then write

$$\tilde{I} = \frac{\tilde{V}}{Z_C} \qquad ,$$

where the quantity

$$Z_C = \frac{1}{\omega C} \qquad , \qquad \text{[impedance of a capacitor]}$$

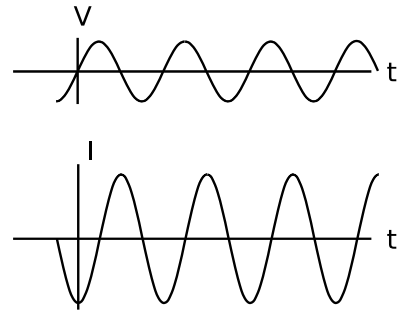
having units of ohms, is called the *impedance* of the capacitor at this frequency. Note that it is only the *maximum* current, $\tilde{I}$, that is proportional to the *maximum* voltage, $\tilde{V}$, so the capacitor is not behaving like a resistor. The maxima of V and I occur at different times, as shown in figure s. It makes sense that the impedance becomes infinite at zero frequency. Zero frequency means that it would take an infinite time before the voltage would change by any amount. In other words, this is like a situation where the capacitor has been connected across the terminals of a battery and been allowed to settle down to a state where there is constant charge on both terminals. Since the electric fields between the plates are constant, there is no energy being added to or taken out of the field. A capacitor that can't exchange energy with any other circuit component is nothing more than a broken (open) circuit.

Self-Check
Why can't a capacitor have its impedance printed on it along with its capacitance? ▷ Answer, p. 709

Similar math (but this time with an integral instead of a derivative) gives

$$Z_L = \omega L \qquad \text{[impedance of an inductor]}$$

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.



t / The current through an inductor lags behind the voltage by a phase angle of 90°.

The phase relationships shown in figures s and t 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. 577).

Low-pass and high-pass filters *example 27*
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, and (3) impedances in parallel and series don't combine according to the same rules as resistances. It is possible, however, to get get around the third limitation, as discussed in subsection .

Discussion Question

A Figure s on page 494 shows the voltage and current for a capacitor. Sketch the *q*-*t* graph, and use it to give a physical explanation of the phase relationship between the voltage and current. For example, why is the current zero when the voltage is at a maximum or minimum?

B Figure t on page 495 shows the voltage and current for an inductor. The power is considered to be positive when energy is being put into the inductor's magnetic field. Sketch the graph of the power, and then the graph of *U*, the energy stored in the magnetic field, and use it to give a physical explanation of the *P*-*t* graph. In particular, discuss why the frequency is doubled on the *P*-*t* graph.

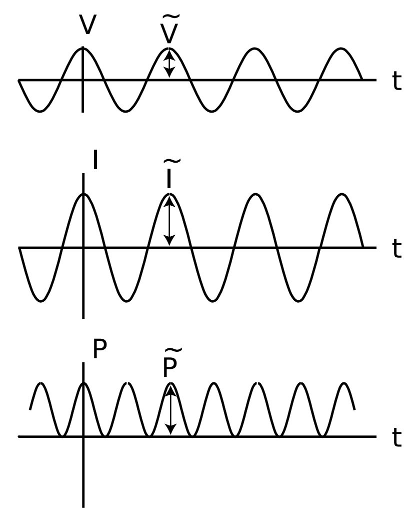
C Relate the features of the graph in figure t on page 495 to the story told in cartoons in figure m/2-3 on page 489.

10.5.6 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 $\tilde{V}$. In the special case of a resistor, we know that I and V are in phase. For example, if V varies as $\tilde{V}\cos\omega t$, then I will be a cosine as well, $\tilde{I}\cos\omega t$. The power is then $\tilde{I}\tilde{V}\cos^2\omega t$, which is always positive,[5] and varies between 0 and $\tilde{I}\tilde{V}$. Even if the time variation was $\cos\omega t$ or $\sin(\omega t+\pi/4)$, we would still have a maximum power of $\tilde{I}\tilde{V}$, 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.[6] Since the circuit oscillates at a frequency[7] 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.



u / 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

$$\begin{aligned} E &= \int \mathrm{d}E \\ &= \int_0^T \frac{\mathrm{d}E}{\mathrm{d}t}\mathrm{d}t \qquad , \end{aligned}$$

[5] A resistor always turns electrical energy into heat. It never turns heat into electrical energy!

[6] 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.

[7] Note that this time we "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.

where $T = 2\pi/\omega$ is the period.

$$\begin{aligned}
E &= \int_0^T P\mathrm{d}t \\
&= \int_0^T P\mathrm{d}t \\
&= \int_0^T \tilde{I}\tilde{V}\cos^2\omega t\mathrm{d}t \\
&= \tilde{I}\tilde{V}\int_0^T \cos^2\omega t\mathrm{d}t \\
&= \tilde{I}\tilde{V}\int_0^T \frac{1}{2}\left(1+\cos 2\omega t\right)\mathrm{d}t
\end{aligned}$$

The reason for using the trig identity $\cos^2 x = (1+\cos 2x)/2$ in the last step is that it lets us get the answer without doing a hard integral. Over the course of one full cycle, the quantity $\cos 2\omega t$ goes positive, negative, positive, and negative again, so the integral of it is zero. We then have

$$\begin{aligned}
E &= \tilde{I}\tilde{V}\int_0^T \frac{1}{2}\mathrm{d}t \\
&= \frac{\tilde{I}\tilde{V}T}{2}
\end{aligned}$$

The average power is

$$\begin{aligned}
P_{av} &= \frac{\text{energy transferred in one full cycle}}{\text{time for one full cycle}} \\
&= \frac{\tilde{I}\tilde{V}T/2}{T} \\
&= \frac{\tilde{I}\tilde{V}}{2} \qquad ,
\end{aligned}$$

i.e. the average is half the maximum. The power varies from 0 to $\tilde{I}\tilde{V}$, 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

$$P_{av} = \frac{\tilde{I}\tilde{V}}{2} \qquad \text{[average power in a resistor]}$$

for a resistor.

RMS quantities

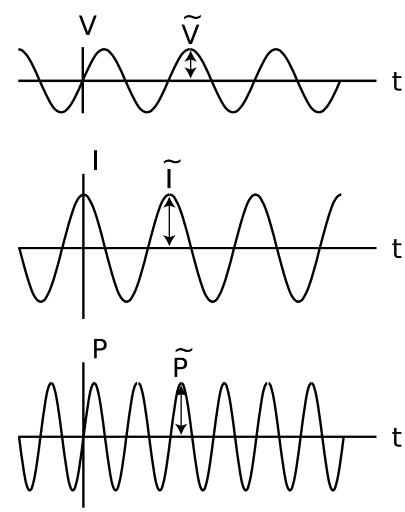
Suppose one day the electric company decided to start supplying your electricity as DC rather than AC. How would the DC voltage have to be related to the amplitude $\tilde{V}$ of the AC voltage previously used if they wanted your lightbulbs to have the same brightness as before? The resistance of the 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, $P_{av} = \tilde{I}\tilde{V}/2 = \tilde{V}^2/2R$. 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 V) $\times \sqrt{2} \approx 160$ V. The reason for referring to $\tilde{V}/\sqrt{2}$ as "the" voltage is that people who are naive about AC circuits can plug $\tilde{V}/\sqrt{2}$ into a familiar DC equation like $P = V^2/R$ and get the right *average* answer. The quantity $\tilde{V}/\sqrt{2}$ 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 $\tilde{V}/\sqrt{2}$. 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, $\tilde{V}\cos\omega t$, then the relation $I = C\mathrm{d}V/\mathrm{d}t$ tells us that the current will be some constant multiplied by minus the sine, $-\tilde{V}\sin\omega t$. The integral we did in the case of a resistor now becomes

$$E = \int_0^T -\tilde{I}\tilde{V}\sin\omega t\cos\omega t\mathrm{d}t \qquad ,$$

and based on figure v, 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!



v / Power in a capacitor: the rate at which energy is being stored in (+) or removed from (-) the electric field.

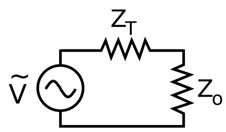
Why is this? It makes sense if you think in terms of energy. A resistor converts electrical energy to heat, never the other way around. A capacitor, however, merely stores electrical energy in an electric field and then gives it back. For a capacitor,

$$P_{av} = 0 \qquad \text{[average power in a capacitor]}$$

Notice that although the average power is zero, the power at any given instant is *not* typically zero, as shown in figure v. 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.



w / We wish to maximize the power delivered to the load, Z_o, by adjusting its impedance.

10.5.7 Impedance Matching

Figure w shows a commonly encountered situation: we wish to maximize the average power, P_{av}, delivered to the load for a fixed value of $\tilde{V}$, 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. Using the methods of subsection 10.5.9, however, it turns out that we can easily generalize our result to deal with any kind of impedance 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

$$\tilde{I} = \frac{\tilde{V}}{Z_\text{o}+Z_T} \qquad ,$$

so

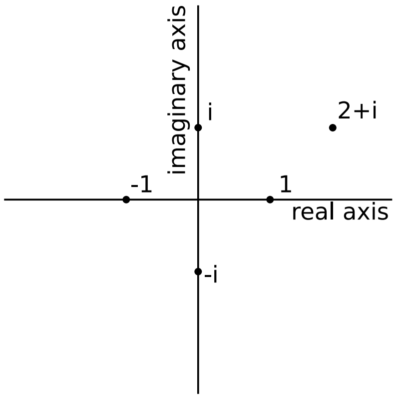
$$\begin{aligned}
P_{av} &= \tilde{I}\tilde{V}/2 \\
&= \tilde{I}^2 Z_\text{o}/2 \\
&= \frac{\tilde{V}^2 Z_\text{o}}{(Z_\text{o}+Z_T)^2}/2 \qquad .
\end{aligned}$$

The maximum of this expression occurs where the derivative is zero,

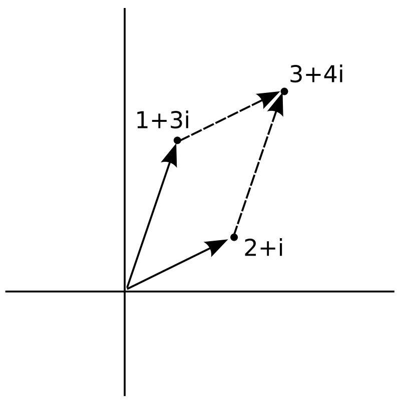
$$\begin{aligned}
0 &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}Z_\text{o}}\left[\frac{\tilde{V}^2 Z_\text{o}}{(Z_\text{o}+Z_T)^2}\right] \\
0 &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}Z_\text{o}}\left[\frac{Z_\text{o}}{(Z_\text{o}+Z_T)^2}\right] \\
0 &= (Z_\text{o}+Z_T)^{-2} - 2Z_\text{o}(Z_\text{o}+Z_T)^{-3} \\
0 &= (Z_\text{o}+Z_T) - 2Z_\text{o} \\
Z_\text{o} &= Z_T
\end{aligned}$$

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, $\tilde{I}^2 Z_o$, decreases in proportion to Z_o.

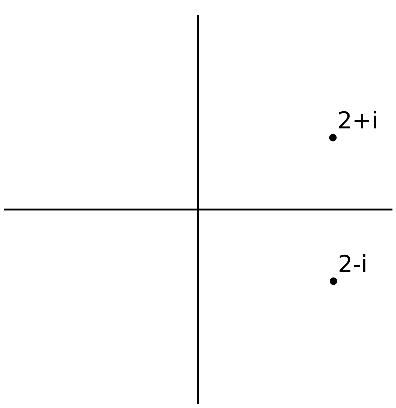
Maximizing the power transmission by matching Z_T to Z_o is called *impedance matching*. For example, an 8-ohm home stereo speaker will be correctly matched to a home stereo amplifier with an internal impedance of 8 ohms, and 4-ohm car speakers will be correctly matched to a car stereo with a 4-ohm internal 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.



x / Visualizing complex numbers as points in a plane.



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



z / A complex number and its conjugate.

10.5.8 Review of Complex Numbers

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.

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^*.

The complex numbers obey all the same rules of arithmetic as the reals, except that they can't be ordered along a single line. That is, it's not possible to say whether one complex number is greater than another. We can compare them in terms of their magnitudes (their distances from the origin), but two distinct complex numbers may have the same magnitude, so, for example, we can't say whether 1 is greater than i or i is greater than 1.

A square root of i *example 28*

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

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

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

Example 28 showed one method of multiplying complex numbers. However, there is another nice interpretation of complex multiplication. We define the argument of a complex number as its angle in the complex plane, measured counterclockwise from the positive real axis. Multiplying two complex numbers then corresponds to multiplying their magnitudes, and adding their arguments.

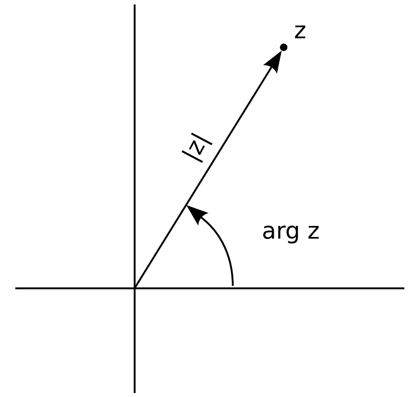
Self-Check

Using this interpretation of multiplication, how could you find the square roots of a complex number? ▷ Answer, p. 709

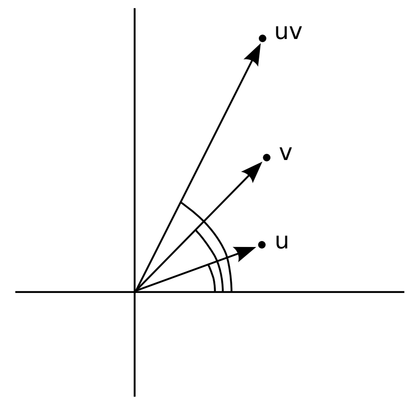
An identity — *example 29*

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. [8]

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.



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



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

[8] 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.

Discussion Questions

A Find $\arg i$, $\arg(-i)$, and $\arg 37$, where $\arg z$ denotes the argument of the complex number z.

B Visualize the following multiplications in the complex plane using the interpretation of multiplication in terms of multiplying magnitudes and adding arguments: $(i)(i) = -1$, $(i)(-i) = 1$, $(-i)(-i) = -1$.

C If we visualize z as a point in the complex plane, how should we visualize $-z$?

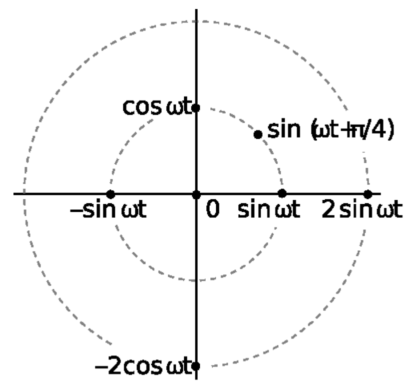
D Find four different complex numbers z such that $z^4 = 1$.

E Compute the following:

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

10.5.9 Complex Impedance

How do impedances combine in series and parallel? As I've warned you already, you can't just treat them as resistors for these purposes. Figure ac shows a useful way to visualize what's going on. When a circuit is being driven at a frequency ω, we use points in the plane to represent sinusoidal functions with various phases and amplitudes.



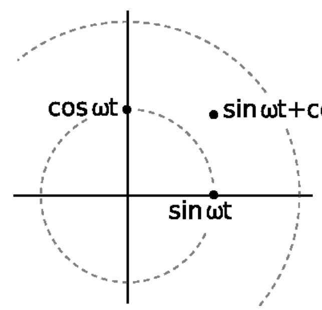
ac / Representing functions with points in polar coordinates.

A single capacitor *example 30*

▷ If the current through a capacitor is $I(t) = \tilde{I}\cos\omega t$, then the voltage is $V(t) = (\tilde{I}/\omega C)\sin\omega t$. More generally, the voltage is behind of the current by 90 degrees ($\pi/2$) in phase. How can we visualize this rule using a polar coordinate plot?

▷ The voltage is represented by a point that is 90 degrees clockwise from the one representing the current, and to get the radial coordinate of the point representing the voltage, we multiply the current's radial coordinate by the impedance $Z = 1/\omega C$.

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. 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. A less trivial example is $\cos\omega t + \sin\omega t = \sqrt{2}\sin(\omega t + \pi/4)$, which is suggestive when we visualize it in figure ad.



ad / Adding two sinusoidal functions.

Self-Check

Which of the following functions can be represented in this way? $\cos(6t-4)$, $\cos^2 t$, $\tan t$ ▷ Answer, p. 710

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 $\cos\omega t$. Our examples of adding functions have all been consistent with the normal method of adding complex numbers, and it can be proved that this method of addition works in general. Furthermore, the rule for finding the voltage across a capacitor corresponds to *multiplication* by the complex number $-i/\omega C$. We thus have two types of complex numbers: those that represent sinusoidal functions of time, and those that represent impedances. Rather than inventing a new set of notation for these quantities, we simply redefine symbols like $\tilde{V}$, $\tilde{I}$, and Z as complex numbers.

With this convention, the impedances of resistors, capacitors, and inductors are

$$
\begin{aligned}
Z_R &= R \\
Z_C &= -\frac{i}{\omega C} \\
Z_L &= i\omega L \qquad ,
\end{aligned}
$$

and these complex impedances, unlike our old real versions, combine in parallel and in series just like resistances. To prove this, we would only need to recycle our old proofs, but with the understanding that the numbers are all complex now. Those proofs were based on the loop and junction rules, which are still true when we represent voltages and currents as complex numbers.

Series impedance *example 31*

▷ 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

$$
\begin{aligned}
Z &= Z_L + Z_C \\
&= i\omega L - \frac{i}{\omega C} \qquad .
\end{aligned}
$$

The current is then

$$
\tilde{I} = \frac{\tilde{V}}{i\omega L - i/\omega C} \qquad .
$$

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 g on page 129).

Resonance with damping — *example 32*

▷ What is the amplitude of the current in a series LRC circuit?

▷ Generalizing from example 31, we add a third, real impedance:

$$\begin{aligned}|\tilde{I}| &= \frac{|\tilde{V}|}{|Z|}\\ &= \frac{|\tilde{V}|}{|R+i\omega L-i/\omega C|}\\ &= \frac{|\tilde{V}|}{\sqrt{R^2+(\omega L-1/\omega C)^2}}\end{aligned}$$

This result would have taken pages of algebra without the complex number technique!

A second-order stereo crossover filter — *example 33*

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, ae/1, gives a total impedance that at high frequencies is dominated by the inductor's, so the current is proportional to ω^{-1}, and the power drawn by the woofer is proportional to ω^{-2}.

A second-order filter, like ae/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

$$\tilde{V}_d = \tilde{I}_L Z \qquad ,$$

and we also have

$$\tilde{V}_L = \tilde{I}_L Z_L \qquad .$$

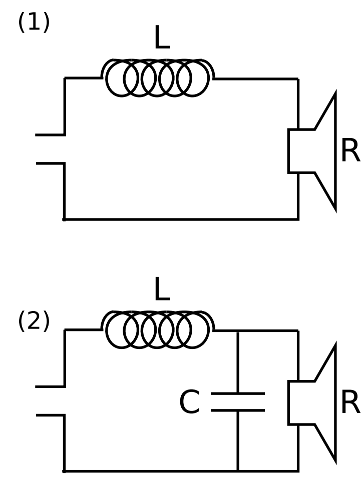
The loop rule, applied to the outer perimeter of the circuit, gives

$$\tilde{V}_d = \tilde{V}_L + \tilde{V}_R \qquad .$$

Straightforward algebra now results in

$$\tilde{V}_R = \frac{\tilde{V}_d}{1+Z_L/Z_C+Z_L/Z_R} \qquad .$$

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}.



ae / Example 33.

10.6 Fields by Gauss' Law

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:

$$E_{outward,\ on\ side\ 1} + E_{outward,\ on\ side\ 2} = 4\pi k\sigma \qquad .$$

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:

$$\left(E_{outward,\ on\ side\ 1} + E_{outward,\ on\ side\ 2}\right) A = 4\pi k\sigma A$$

[9] no relation to the human mathematician of the same name

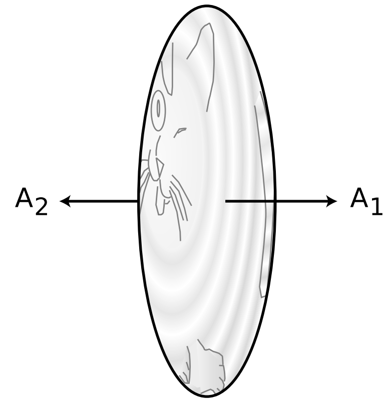
or

$$E_{outward,\ on\ side\ 1}A + E_{outward,\ on\ side\ 2}A = 4\pi kq \qquad ,$$

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 $\mathbf{A}$. 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 $u_{parallel\ to\ v}|\mathbf{v}|$, and she can therefore rewrite her equation as

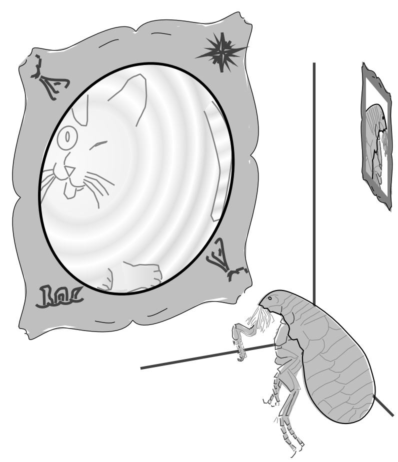
$$\mathbf{E}_1 \cdot \mathbf{A}_1 + \mathbf{E}_2 \cdot \mathbf{A}_2 = 4\pi kq \qquad .$$

The quantity on the left side of this equation is called the *flux* through the surface, written Φ.



a / The area vector is defined to be perpendicular to the surface, in the outward direction. Its magnitude tells how much the area is.

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.

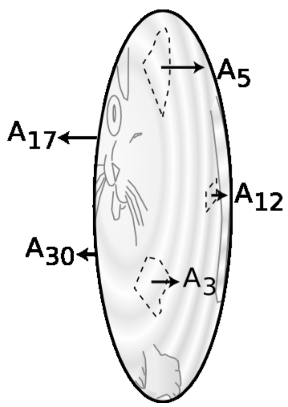


b / Gauss contemplates a map of the known world.

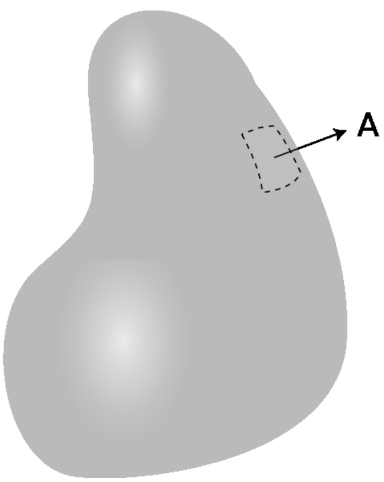
Of course it is possible that the electric charge of planet Flatcat is not perfectly uniform, but that isn't a problem. As discussed in subsection 10.3.2, as long as one is very close to the surface, the field only depends on the *local* charge density. In fact, a side-benefit of Gauss's program of exploration 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:

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

Gauss herself leads one of the expeditions, which heads due east, toward the distant Tail Kingdom, known only from fables and the



c / Each part of the surface has its own area vector. Note the differences in lengths of the vectors, corresponding to the unequal areas.



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

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.

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.

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

$$|\mathbf{E}| = \frac{kq_{total}}{r^2} \qquad ,$$

where r is the radius of Flatcat. The flux is then

$$\Phi = \sum \mathbf{E}_j \cdot \mathbf{A}_j \qquad ,$$

and since the $\mathbf{E}_j$ and $\mathbf{A}_j$ vectors are parallel, the dot product equals $|\mathbf{E}_j||\mathbf{A}_j|$, so

$$\Phi = \sum \frac{kq_{total}}{r^2}|\mathbf{A}_j| \qquad .$$

But the field strength is always the same, so we can take it outside the sum, giving

$$\begin{aligned}\Phi &= \frac{kq_{total}}{r^2}\sum|\mathbf{A}_j| \\ &= \frac{kq_{total}}{r^2}A_{total} \\ &= \frac{kq_{total}}{r^2}4\pi r^2 \\ &= 4\pi kq_{total} \qquad .\end{aligned}$$

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 65), 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, $\Phi = 4\pi kq_{total}$. 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 $4\pi kq_{in}$, 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.

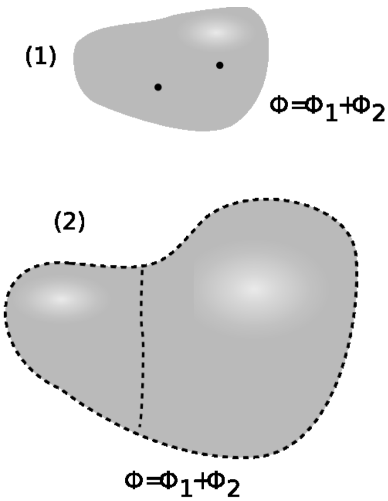
[10] Newton's human namesake actually proved this for gravity, not electricity, but they're both $1/r^2$ forces, so the proof works equally well in both cases.

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, p. 710

10.6.2 Additivity of flux



e / 1. The flux due to two charges equals the sum of the fluxes from each one. 2. When two regions are joined together, the flux through the new region equals the sum of the fluxes through the two parts.

Figure e shows two two different ways in which flux is additive. Figure e/1, additivity by charge, shows that we can break down a charge distribution into two or more parts, and the flux equals the sum of the fluxes due to the individual charges. This follows directly from the fact that the flux is defined in terms of a dot product, $\mathbf{E}\cdot\mathbf{A}$, and the dot product has the additive property $(\mathbf{a}+\mathbf{b})\cdot\mathbf{c} = \mathbf{a}\cdot\mathbf{c}+\mathbf{b}\cdot\mathbf{c}$.

To understand additivity of flux by region, e/2, we have to consider the parts of the two surfaces that were eliminated when they were joined together, like knocking out a wall to make two small apartments into one big one. Although the two regions shared this wall before it was removed, the area vectors were opposite: the direction that is outward from one region is inward with respect to the other. Thus if the field on the wall contributes positive flux to one region, it contributes an equal amount of negative flux to the other region, and we can therefore eliminate the wall to join the two regions, without changing the total flux.

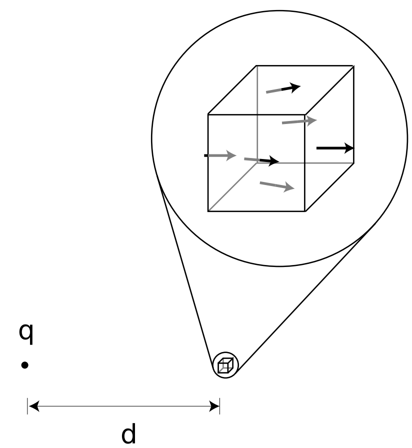
10.6.3 Zero flux from outside charges

A third important property of Gauss' theorem is that it only refers to the charge *inside* the region we choose to discuss. In other words, it asserts that any charge outside the region contributes zero to the flux. This makes at least some sense, because a charge outside the region will have field vectors pointing into the surface on one side, and out of the surface on the other. Certainly there should be at least partial cancellation between the negative (inward) flux on one side and the positive (outward) flux on the other. But why should this cancellation be exact?

To see the reason for this perfect cancellation, we can imagine space as being built out of tiny cubes, and we can think of any charge distribution as being composed of point charges. The additivity-by-charge property tells us that any charge distribution can be handled by considering its point charges individually, and the additivity-by-region property tells us that if we have a single point charge outside a big region, we can break the region down into tiny cubes. If we can prove that the flux through such a tiny cube really does cancel exactly, then the same must be true for any region, which we could build out of such cubes, and any charge distribution, which we can build out of point charges.

For simplicity, we will carry out this calculation only in the special case shown in figure f, where the charge lies along one axis of the cube. Let the sides of the cube have length $2b$, so that the area of each side is $(2b)^2 = 4b^2$. The cube extends a distance b above, below, in front of, and behind the horizontal x axis. There is a distance $d-b$ from the charge to the left side, and $d+b$ to the right side.

There will be one negative flux, through the left side, and five positive ones. Of these positive ones, the one through the right side is very nearly the same in magnitude as the negative flux through the left side, but just a little less because the field is weaker on the right, due to the greater distance from the charge. The fluxes through the other four sides are very small, since the field is nearly perpendicular to their area vectors, and the dot product $\mathbf{E}_j \cdot \mathbf{A}_j$ is zero if the two vectors are perpendicular. In the limit where b is very small, we can approximate the flux by evaluating the field at the center of each of the cube's six sides, giving



f / The flux through a tiny cube due to a point charge.

$$\begin{aligned}\Phi &= \Phi_{left} + 4\Phi_{side} + \Phi_{right} \\ &= |\mathbf{E}_{left}||\mathbf{A}_{left}|\cos 180\,° + 4|\mathbf{E}_{side}||\mathbf{A}_{side}|\cos\theta_{side} \\ &\quad + |\mathbf{E}_{right}||\mathbf{A}_{right}|\cos 0\,° \qquad ,\end{aligned}$$

and a little trig gives $\cos\theta_{side} \approx b/d$, so

$$\begin{aligned}\Phi &= -|\mathbf{E}_{left}||\mathbf{A}_{left}| + 4|\mathbf{E}_{side}||\mathbf{A}_{side}|\frac{b}{d} + |\mathbf{E}_{right}||\mathbf{A}_{right}| \\ &= \left(4b^2\right)\left(-|\mathbf{E}_{left}| + 4|\mathbf{E}_{side}|\frac{b}{d} + |\mathbf{E}_{right}|\right) \\ &= \left(4b^2\right)\left(-\frac{kq}{(d-b)^2} + 4\frac{kq}{d^2}\frac{b}{d} + \frac{kq}{(d+b)^2}\right) \\ &= \left(\frac{4kqb^2}{d^2}\right)\left(-\frac{1}{(1-b/d)^2} + \frac{4b}{d} + \frac{1}{(1+b/d)^2}\right) \qquad .\end{aligned}$$

Using the approximation $(1+\epsilon)^{-2} \approx 1-2\epsilon$ for small ϵ, this becomes

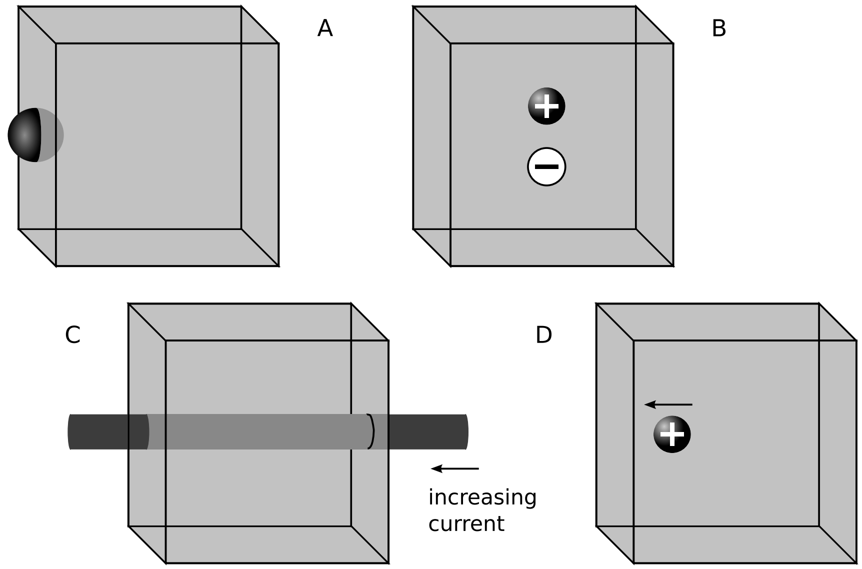
$$\begin{aligned}\Phi &= \left(\frac{4kqb^2}{d^2}\right)\left(-1 - \frac{2b}{d} + \frac{4b}{d} + 1 - \frac{2b}{d}\right) \\ &= 0 \qquad .\end{aligned}$$

Thus in the limit of a very small cube, $b \ll d$, we have proved that the flux due to this exterior charge is zero. The proof can be extended to the case where the charge is not along any axis of the cube,[11] and based on additivity we then have a proof that the flux due to an outside charge is always zero.

[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 35 on page 522.

Discussion Questions

g / Discussion question A-D.



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

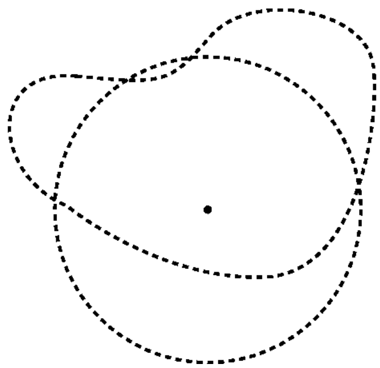
B The dipole is completely enclosed in the cube. What does Gauss's law say about the flux through the cube? If you imagine the dipole's field pattern, can you verify that this makes sense?

C The wire passes in through one side of the cube and out through the other. If the current through the wire is increasing, then the wire will act like an inductor, and there will be a voltage difference between its ends. (The inductance will be relatively small, since the wire isn't coiled up, and the ΔV will therefore also be fairly small, but still not zero.) The ΔV implies the existence of electric fields, and yet Gauss's law says the flux must be zero, since there is no charge inside the cube. Why isn't Gauss's law violated?

D The charge has been loitering near the edge of the cube, but is then suddenly hit with a mallet, causing it to fly off toward the left side of the cube. We haven't yet discussed in detail how disturbances in the electric and magnetic fields ripple outward through space, but it turns out that they do so at the speed of light. (In fact, that's what light is: ripples in the electric and magnetic fields.) Because the charge is closer to the left side of the cube, the change in the electric field occurs there before the information reaches the right side. This would seem certain to lead to a violation of Gauss's law. How can the ideas explored in discussion question C show the resolution to this paradox?

10.6.4 Proof of Gauss' theorem

With the computational machinery we've developed, it is now simple to prove Gauss' theorem. Based on additivity by charge, it suffices to prove the law for a point charge. We have already proved Gauss' law for a point charge in the case where the point charge is outside the region. If we can prove it for the inside case, then we're all done.



h / Completing the proof of Gauss' theorem.

If the charge is inside, we reason as follows. First, we forget about the actual Gaussian surface of interest, and instead construct a spherical one, centered on the charge. For the case of a sphere, we've already seen the proof written on a napkin by the flea named Newton (page 510). Now wherever the actual surface sticks out beyond the sphere, we glue appropriately shaped pieces onto the sphere. In the example shown in figure h, we have to add two Mickey Mouse ears. Since these added pieces do not contain the point charge, the flux through them is zero, and additivity of flux by region therefore tells us that the total flux is not changed when we make this alteration. Likewise, we need to chisel out any regions where the sphere sticks out beyond the actual surface. Again, there is no change in flux, since the region being altered doesn't contain the point charge. This proves that the flux through the Gaussian surface of interest is the same as the flux through the sphere, and since we've already proved that that flux equals $4\pi kq_{in}$, our proof of Gauss' theorem is complete.

Discussion Questions

A A critical part of the proof of Gauss' theorem was the proof that a tiny cube has zero flux through it due to an external charge. Discuss qualitatively why this proof would fail if Coulomb's law was a $1/r$ or $1/r^3$ law.

10.6.5 Gauss' law as a fundamental law of physics

Note that the proof of Gauss' theorem depended on the computation on the napkin discussed on page 10.6.1. The crucial point in this computation was that the electric field of a point charge falls off like $1/r^2$, and since the area of a sphere is proportional to r^2, the result is independent of r. The $1/r^2$ variation of the field also came into play on page 513 in the proof that the flux due to an outside charge is zero. In other words, if we discover some other force of nature which is proportional to $1/r^3$ or r, then Gauss' theorem will not apply to that force. Gauss' theorem is not true for nuclear forces, which fall off exponentially with distance. However, this is the *only* assumption we had to make about the nature of the field. Since gravity, for instance, also has fields that fall off as $1/r^2$, Gauss' theorem is equally valid for gravity — we just have to replace mass with charge, change the Coulomb constant k to the gravitational constant G, and insert a minus sign because the gravitational fields around a (positive) mass point inward.

Gauss' theorem can only be proved if we assume a $1/r^2$ field, and the converse is also true: any field that satisfies Gauss' theorem must be a $1/r^2$ field. Thus although we previously thought of Coulomb's law as the fundamental law of nature describing electric forces, it is equally valid to think of Gauss' theorem as the basic law of nature for electricity. From this point of view, Gauss' theorem is not a mathematical fact but an experimentally testable statement about nature, so we'll refer to it as Gauss' *law*, just as we speak of Coulomb's *law* or Newton's *law* of gravity.

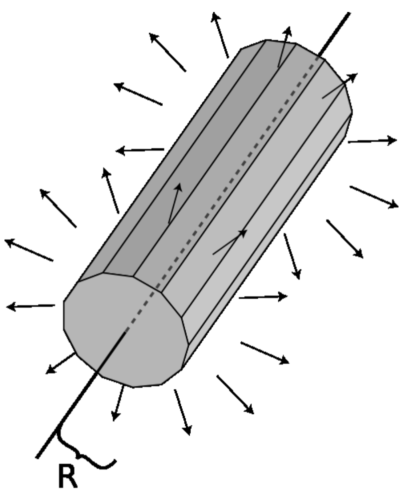
If Gauss' law is equivalent to Coulomb's law, why not just use Coulomb's law? First, there are some cases where calculating a field is easy with Gauss' law, and hard with Coulomb's law. More importantly, Gauss' law and Coulomb's law are only mathematically equivalent under the assumption that all our charges are standing still, and all our fields are constant over time, i.e. in the study of electrostatics, as opposed to electrodynamics. As we broaden our scope to study generators, inductors, transformers, and radio antennas, we will encounter cases where Gauss' law is valid, but Coulomb's law is not.

10.6.6 Applications

Often we encounter situations where we have a static charge distribution, and we wish to determine the field. Although superposition is a generic strategy for solving this type of problem, if the charge distribution is symmetric in some way, then Gauss' law is often a far easier way to carry out the computation.

Field of a long line of charge

Consider the field of an infinitely long line of charge, holding a uniform charge per unit length λ. Computing this field by brute-force superposition was fairly laborious (examples 10 on page 467 and 13 on page 473). With Gauss' law it becomes a very simple calculation.



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

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

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

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

$$\begin{aligned} 4\pi k q_{in} &= \sum \mathbf{E}_j \cdot \mathbf{A}_j \\ 4\pi k\lambda L &= \sum |\mathbf{E}_j||\mathbf{A}_j| \qquad . \end{aligned}$$

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

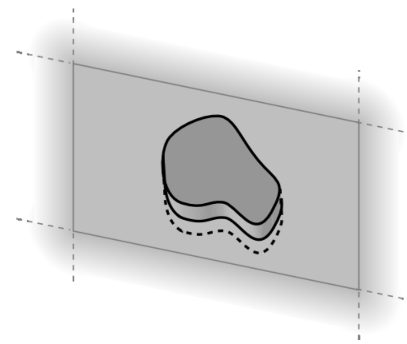
$$4\pi k\lambda L = |\mathbf{E}| \sum |\mathbf{A}_j|$$

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

$$\begin{aligned} 4\pi k\lambda L &= |\mathbf{E}| \times 2\pi RL \\ |\mathbf{E}| &= \frac{2k\lambda}{R} \end{aligned}$$

Field near a surface charge

As claimed earlier, the result $E = 2\pi k\sigma$ for the field near a charged surface is a special case of Gauss' law. We choose a Gaussian surface of the shape shown in figure j, known as a Gaussian pillbox. The exact shape of the flat end caps is unimportant.



j / Applying Gauss' law to an infinite charged surface.

The symmetry of the charge distribution tells us that the field points directly away from the surface, and is equally strong on both sides of the surface. This means that the end caps contribute equally to the flux, and the curved sides have zero flux through them. If the area of each end cap is A, then

$$4\pi k q_{in} = \mathbf{E}_1 \cdot \mathbf{A}_1 + \mathbf{E}_2 \cdot \mathbf{A}_2 \qquad ,$$

where the subscripts 1 and 2 refer to the two end caps. We have $\mathbf{A}_2 = -\mathbf{A}_1$, so

$$\begin{aligned} 4\pi k q_{in} &= \mathbf{E}_1 \cdot \mathbf{A}_1 - \mathbf{E}_2 \cdot \mathbf{A}_1 \\ 4\pi k q_{in} &= (\mathbf{E}_1 - \mathbf{E}_2) \cdot \mathbf{A}_1 \qquad , \end{aligned}$$

and by symmetry the magnitudes of the two fields are equal, so

$$\begin{aligned} 2|\mathbf{E}|A &= 4\pi k\sigma A \\ |\mathbf{E}| &= 2\pi k\sigma \end{aligned}$$

The symmetry between the two sides could be broken by the existence of other charges nearby, whose fields would add onto the field of the surface itself. Even then, Gauss's law still guarantees

$$4\pi k q_{in} = (\mathbf{E}_1 - \mathbf{E}_2) \cdot \mathbf{A}_1 \qquad ,$$

or

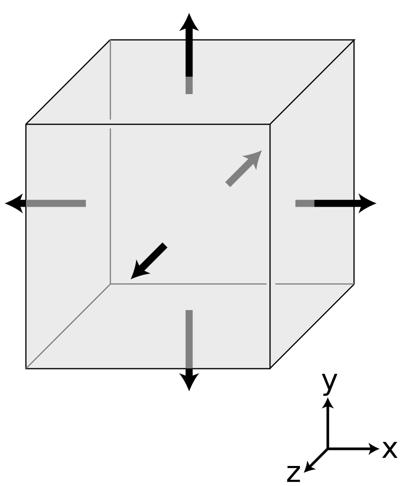
$$|\mathbf{E}_{\perp,1} - \mathbf{E}_{\perp,2}| = 4\pi k\sigma \qquad ,$$

where the subscript $\perp$ indicates the component of the field parallel to the surface (i.e., parallel to the area vectors). In other words, the electric field changes discontinuously when we pass through a charged surface; the discontinuity occurs in the component of the field perpendicular to the surface, and the amount of discontinuous change is $4\pi k\sigma$. This is a completely general statement that is true near any charged surface, regardless of the existence of other charges nearby.

10.7 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.



a / A tiny cubical Gaussian surface.

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 512. 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 $x+\mathrm{d}x$, from y to $y+\mathrm{d}y$, and from y to $y+\mathrm{d}y$.

The sides at x and $x+\mathrm{d}x$ have area vectors $-\mathrm{d}y\mathrm{d}z\hat{\mathbf{x}}$ and $\mathrm{d}y\mathrm{d}z\hat{\mathbf{x}}$, respectively. The flux through the side at x is $-E_x(x)\mathrm{d}y\mathrm{d}z$, and the flux through the opposite side, at $x+\mathrm{d}x$ is $E_x(x+\mathrm{d}x)\mathrm{d}y\mathrm{d}z$. The sum of these is $(E_x(x+\mathrm{d}x)-E_x(x))\mathrm{d}y\mathrm{d}z$, 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 $E_x(x+\mathrm{d}x)-E_x(x)$ can be rewritten as $\mathrm{d}E_x = (\mathrm{d}E_x)/(\mathrm{d}x)\mathrm{d}x$, so the contribution to the flux from these two sides of the cube ends up being

$$\frac{\mathrm{d}E_x}{\mathrm{d}x}\mathrm{d}x\mathrm{d}y\mathrm{d}z \qquad .$$

Doing the same for the other sides, we end up with a total flux

$$\begin{aligned}\mathrm{d}\Phi &= \left(\frac{\mathrm{d}E_x}{\mathrm{d}x}+\frac{\mathrm{d}E_y}{\mathrm{d}y}+\frac{\mathrm{d}E_z}{\mathrm{d}z}\right)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\ &= \left(\frac{\mathrm{d}E_x}{\mathrm{d}x}+\frac{\mathrm{d}E_y}{\mathrm{d}y}+\frac{\mathrm{d}E_z}{\mathrm{d}z}\right)\mathrm{d}v \qquad ,\end{aligned}$$

where $\mathrm{d}v$ 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,

$$\mathrm{d}\Phi = \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}\right)\mathrm{d}v \qquad .$$

Using Gauss' law,

$$4\pi k q_{in} = \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}\right)\mathrm{d}v \qquad ,$$

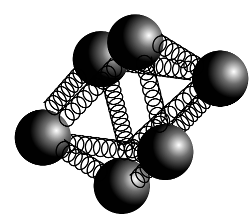
and we introduce the notation ρ (Greek letter rho) for the charge per unit volume, giving

$$4\pi k\rho = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \qquad .$$

The quantity on the right is called the *divergence* of the electric field, written $\operatorname{div}\mathbf{E}$. Using this notation, we have

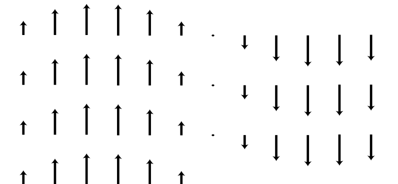
$$\operatorname{div}\mathbf{E} = 4\pi k\rho \qquad .$$

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 div **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 $\operatorname{div}\mathbf{E} < 0$. What if the field is constant? We know based on the definition of the divergence that we should have $\operatorname{div}\mathbf{E} = 0$ 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.



c / Example 34.

Divergence of a sine wave *example 34*

▷ 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

$$\mathbf{E} = (\sin Kx)\,\hat{\mathbf{y}} \qquad ,$$

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

$$\mathbf{E} = \frac{\partial E_y}{\partial y} \qquad ,$$

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!

Electric field of a point charge — *example 35*

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

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

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

$$\begin{aligned}\mathbf{E} &= \frac{kq}{r^3}\hat{\mathbf{r}} \\ &= \frac{kq(x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}})}{\left(x^2+y^2+z^2\right)^{3/2}} \qquad .\end{aligned}$$

The three terms in the divergence are all similar, e.g.

$$\begin{aligned}\frac{\partial E_x}{\partial x} &= kq\frac{\partial}{\partial x}\left[\frac{x}{\left(x^2+y^2+z^2\right)^{3/2}}\right] \\ &= kq\left[\frac{1}{\left(x^2+y^2+z^2\right)^{3/2}} - \frac{3}{2}\,\frac{2x^2}{\left(x^2+y^2+z^2\right)^{5/2}}\right] \\ &= kq\left(r^{-3} - 3x^2r^{-5}\right) \qquad .\end{aligned}$$

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.

The field inside a uniform sphere of charge *example 36*

▷ 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

$$\mathbf{E} = br^p\hat{\mathbf{r}} \qquad ,$$

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 35, we rewrite $\hat{\mathbf{r}}$ as $\mathbf{r}/r$, and to simplify the writing we define $n = p - 1$, so

$$\mathbf{E} = br^n\mathbf{r} \qquad .$$

Gauss' law in differential form is

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

so we want a field whose divergence is constant. For a field of the form we guessed, the divergence has terms in it like

$$\begin{aligned} \frac{\partial E_x}{\partial x} &= \frac{\partial}{\partial x}\left(br^n x\right) \\ &= b\left(nr^{n-1}\frac{\partial r}{\partial x}x + r^n\right) \end{aligned}$$

The partial derivative $\partial r/\partial x$ is easily calculated to be x/r, so

$$\frac{\partial E_x}{\partial x} = b\left(nr^{n-2}x^2 + r^n\right)$$

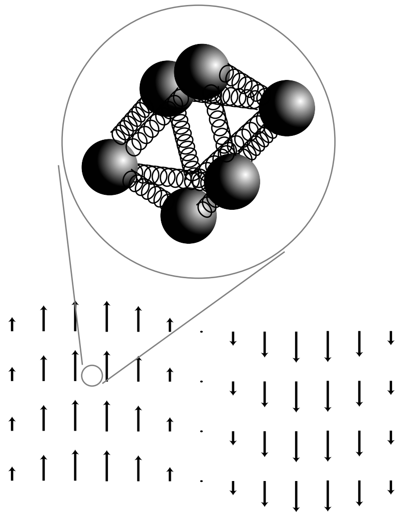
Adding in similar expressions for the other two terms in the divergence, and making use of $x^2 + y^2 + z^2 = r^2$, we have

$$\text{div}\,\mathbf{E} = b(n+3)r^n \qquad .$$

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

$$\mathbf{E} = \frac{4\pi k\rho}{3}r\,\hat{\mathbf{r}} \qquad .$$

The field is zero at the center, and gets stronger and stronger as we approach the surface.



d / Discussion question A.

Discussion Questions

A As suggested by the figure, discuss the results you would get by inserting the div-meter at various locations in the sine-wave field.

This chapter is summarized on page 734. Notation and terminology are tabulated on pages 718-719.

Problems

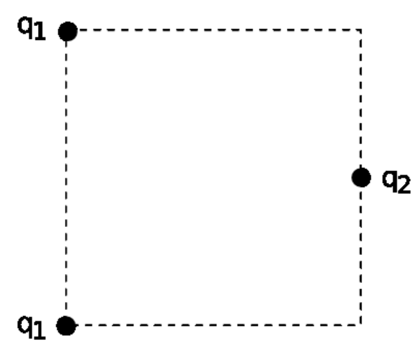
The symbols √, , etc. are explained on page 534.

1 (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. √
(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.)

2 Show that the alternative definition of the magnitude of the electric field, $|\mathbf{E}| = \tau / D_t \sin\theta$, has units that make sense.

3 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.

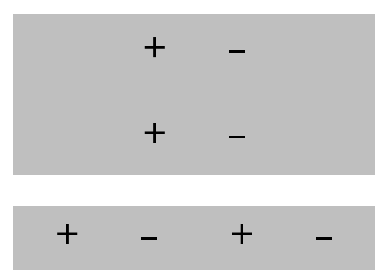
4 Three charges are arranged on a square as shown. All three charges are positive. What value of q_2/q_1 will produce zero electric field at the center of the square? √



Problem 4.

5 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. √

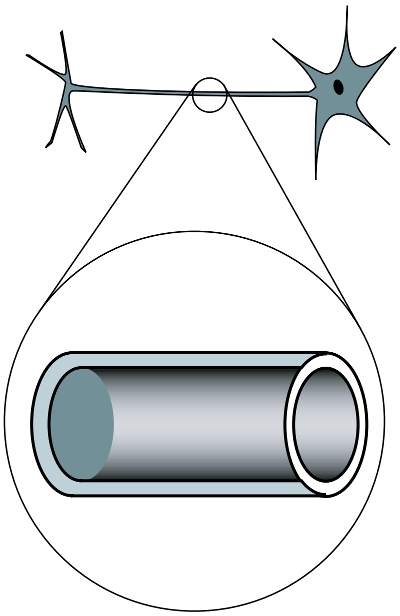
6 Compare the two dipole moments.



Problem 6.

7 Find an arrangement of charges that has zero total charge and zero dipole moment, but that will make nonvanishing electric fields.

8 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? ✓



Problem 8.

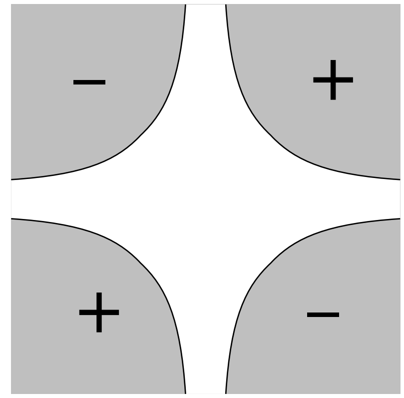
9 The gap between the electrodes in an automobile engine's spark plug is 0.060 cm. To produce an electric spark in a gasoline-air mixture, an electric field of 3.0×10^6 V/m must be achieved. On starting a car, what minimum voltage must be supplied by the ignition circuit? Assume the field is uniform. ✓
(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?

10 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. ✓

11 (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. ▷ Hint, p. 705

12 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, p. 716

13 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 $+V_o$, and the other two at $-V_o$, 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 $V = V_o b^{-2} xy$ 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." √

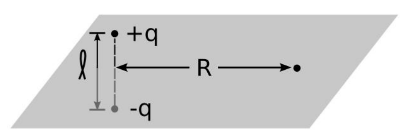


Problem 13.

14 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 11), 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 7). 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? √

15 (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. √
(b) Write down another voltage that gives exactly the same field.

16 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).



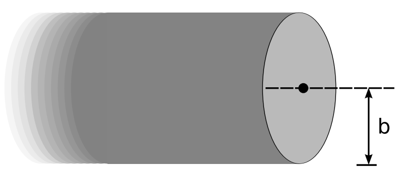
Problem 16.

17 (a) As suggested in example 9 on page 466, 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 469.

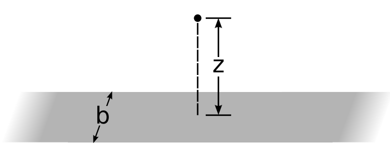
18 As suggested in example 11 on page 468, show that you can get the same result for the on-axis field by differentiating the voltage

19 (a) Example 10 on page 467 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 Q is not changing, so as L gets bigger, the charge is getting spread more thinly. ▷ Answer, p. 712
(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 $\mathrm{d}x$, and extends from x to $x + \mathrm{d}x$. 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.



Problem 20.

20 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?



Problem 21.

21 (a) Show that the field found in example 10 on page 467 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. √
(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 19a, which is given on page 712.

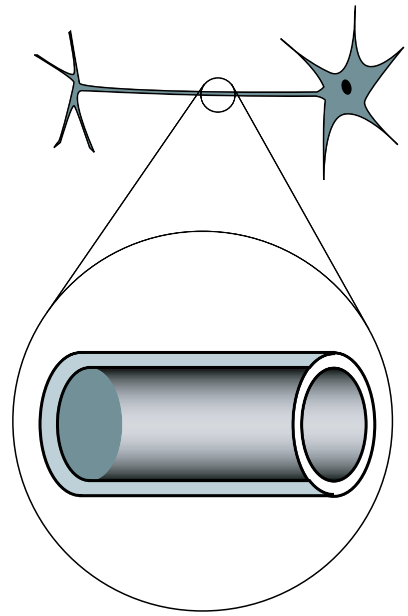
22 Find the voltage at the edge of a uniformly charged disk. (Define $V = 0$ to be infinitely far from the disk.)
√ ▷ Hint, p. 705

23 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. √

(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?

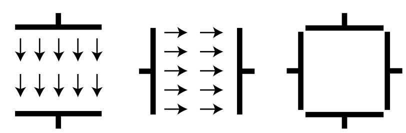
(c) Make an order-of-magnitude estimate of the energy released by a typical lightning bolt, assuming reasonable values for its size and altitude. E_c is about 10^6 V/m.



Problem 24.

24 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.) ▷ Hint, p. 705 √

(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?



Problem 25.

25 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. √

(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. √

(c) Calculate the energy of the put-together configuration. Does assembling them like this release energy, consume energy, or neither? √

26 (a) Show that the energy in the electric field of a point charge is infinite! Does the integral diverge at small distances, at large distances, or both? ▷ Hint, p. 705

(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 36 shows this using a different technique.)

Remark: 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. ✓

27 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? ✓

28 (a) Show that the equation $V_L = L\, \mathrm{d}I/\mathrm{d}t$ has the right units.
(b) Verify that RC has units of time.
(c) Verify that L/R has units of time.

29 Find the energy stored in a capacitor in terms of its capacitance and the voltage difference across it. ✓

30 Find the inductance of two identical inductors in parallel.

31 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.

32 (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?

33 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.) ✓

34 Starting from the relation $V = L\mathrm{d}I/\mathrm{d}t$ for the voltage difference across an inductor, show that an inductor has an impedance equal to $L\omega$.

35 (a) Use complex number techniques to rewrite the function $f(t) = 4\sin\omega t + 3\cos\omega t$ in the form $A\sin(\omega t + \delta)$. ✓
(b) Verify the result using the trigonometric identity $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \sin\beta\cos\alpha$.

36 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? √
(c) What is the resonant frequency f_{res} corresponding to your answer in part b? √

37 (a) Find the parallel impedance of a 37 kΩ resistor and a 1.0 nF capacitor at $f = 1.0 \times 10^4$ Hz. √
(b) A voltage with an amplitude of 1.0 mV drives this impedance at this frequency. What is the amplitude of the current drawn from the voltage source, what is the current's phase angle with respect to the voltage, and does it lead the voltage, or lag behind it? √

38 (a) Use Gauss' law to find the fields inside and outside an infinite cylindrical surface with radius b and uniform surface charge density σ. √
(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 519).
(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. √

39 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 38.) √

40 This is an alternative approach to problem 39, 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. √
(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. ✓

41 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

$$\begin{aligned} x &= u + p \\ y &= v + q \\ z &= w + r \qquad , \end{aligned}$$

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.

42 Using a techniques similar to that of problem 41, 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.

43 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?

44 (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 $\mathbf{E} = bxz\hat{\mathbf{z}}$.

45 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.

46 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.

47 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.

48 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.

49 (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. $\checkmark$

50 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.

51 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.

52 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.

53 Redo the calculation of example 5 using a different origin for the coordinate system, and show that you get the same result.

54 Calculate the quantity i^i.

Key to symbols:
easy typical challenging difficult very difficult
$\surd$ An answer check is available at www.lightandmatter.com.

Exercises

Exercise 10A: Field Vectors

Apparatus:

3 solenoids

DC power supply

compass

ruler

cut-off plastic cup

At this point you've studied the gravitational field, **g**, and the electric field, **E**, but not the magnetic field, **B**. However, they all have some of the same mathematical behavior: they act like vectors. Furthermore, magnetic fields are the easiest to manipulate in the lab. Manipulating gravitational fields directly would require futuristic technology capable of moving planet-sized masses around! Playing with electric fields is not as ridiculously difficult, but static electric charges tend to leak off through your body to ground, and static electricity effects are hard to measure numerically. Magnetic fields, on the other hand, are easy to make and control. Any moving charge, i.e. any current, makes a magnetic field.

A practical device for making a strong magnetic field is simply a coil of wire, formally known as a solenoid. The field pattern surrounding the solenoid gets stronger or weaker in proportion to the amount of current passing through the wire.

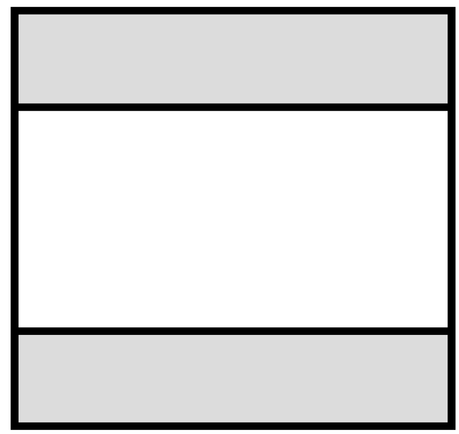
1. With a single solenoid connected to the power supply and laid with its axis horizontal, use a magnetic compass to explore the field pattern inside and outside it. The compass shows you the field vector's direction, but not its magnitude, at any point you choose. Note that the field the compass experiences is a combination (vector sum) of the solenoid's field and the earth's field.

2. What happens when you bring the compass extremely far away from the solenoid?

What does this tell you about the way the solenoid's field varies with distance?

Thus although the compass doesn't tell you the field vector's magnitude numerically, you can get at least some general feel for how it depends on distance.

3. Make a sea-of-arrows sketch of the magnetic field in the horizontal plane containing the solenoid's axis. The length of each arrow should at least approximately reflect the strength of the magnetic field at that point.



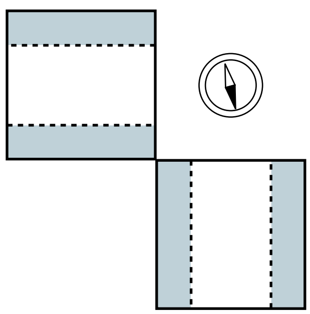
Does the field seem to have sources or sinks?

4. What do you think would happen to your sketch if you reversed the wires?

Try it.

5. Now hook up the two solenoids in parallel. You are going to measure what happens when their two fields combine in the at a certain point in space. As you've seen already, the solenoids' nearby fields are much stronger than the earth's field; so although we now theoretically have three fields involved (the earth's plus the two solenoids'), it will be safe to ignore the earth's field. The basic idea here is to place the solenoids with their axes at some angle to each other, and put the compass at the intersection of their axes, so that it is the same distance from each solenoid. Since the geometry doesn't favor either solenoid, the only factor that would make one solenoid influence the compass more than the other is current. You can use the cut-off plastic cup as a little platform to bring the compass up to the same level as the solenoids' axes.

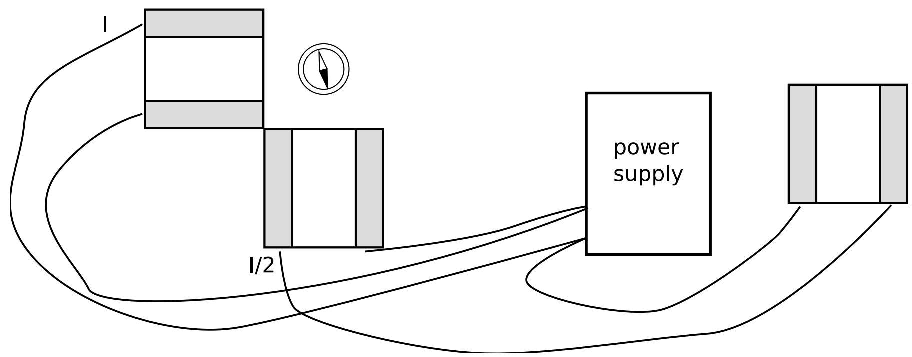
a)What do you think will happen with the solenoids' axes at 90 degrees to each other, and equal currents? Try it. Now represent the vector addition of the two magnetic fields with a diagram. Check your diagram with your instructor to make sure you're on the right track.



b) Now try to make a similar diagram of what would happen if you switched the wires on one of the solenoids.

After predicting what the compass will do, try it and see if you were right.

c)Now suppose you were to go back to the arrangement you had in part a, but you changed one of the currents to half its former value. Make a vector addition diagram, and use trig to predict the angle.



Try it. To cut the current to one of the solenoids in half, an easy and accurate method is simply to put the third solenoid in series with it, and put that third solenoid so far away that its magnetic field doesn't have any significant effect on the compass.

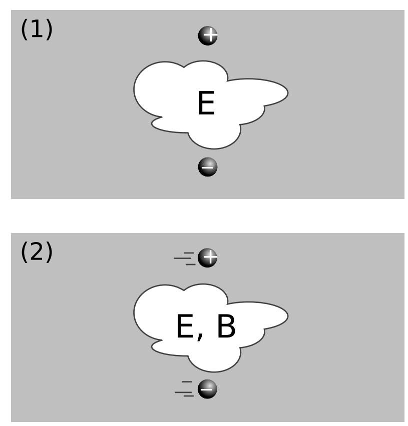
Chapter 11
Electromagnetism

Think not that I am come to destroy the law, or the prophets: I am not come to destroy, but to fulfill. *Matthew 5:17*

11.1 More About the Magnetic Field

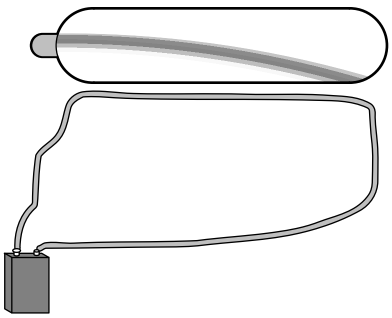
11.1.1 Magnetic forces

In this chapter, I assume you know a few basic ideas about Einstein's theory of relativity, as described in section 7.1. 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.



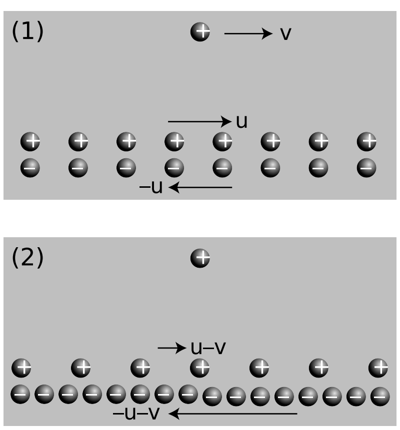
a / The pair of charged particles, as seen in two different frames of reference.

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. 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, **E**, and a magnetic field, **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.



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 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.



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.)

To see why this must occur based on relativity, we make the mathematically idealized model shown in figure c. 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

[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.

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,

[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 chapter 5 of E.M. Purcell, *Electricity and Magnetism*, McGraw Hill, 1963.

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

$$\begin{aligned}\lambda_{total,\ 2} &\approx \lambda\left[1+\frac{(u-v)^2}{2c^2}\right] - \lambda\left[1+\frac{(-u-v)^2}{2c^2}\right]\\ &= \frac{-2\lambda uv}{c^2} \qquad .\end{aligned}$$

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

$$E = \frac{4k\lambda uv}{c^2 R} \qquad ,$$

and the force acting on the lone charge q is

$$F = \frac{4k\lambda quv}{c^2 R} \qquad .$$

In frame 1, the current is $I = 2\lambda_1 u$ (see homework problem 9), 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

$$F = \frac{2kIqv}{c^2 R} \qquad .$$

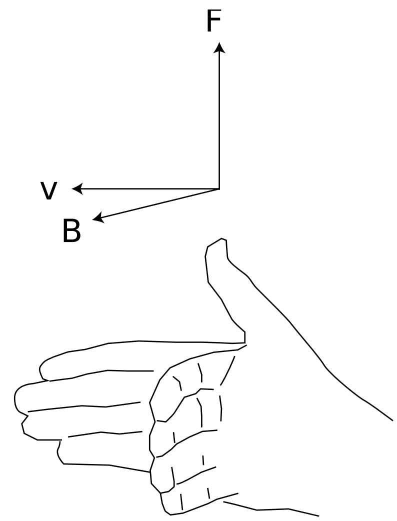
11.1.2 The magnetic field

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 $k/c^2 = 10^{-7}\ \text{N}\cdot\text{s}^2/\text{C}^2$ exactly, as a matter of definition.



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.

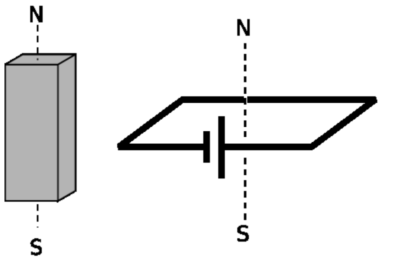
Naively, we could try to work by analogy with the electric field, and define the magnetic field as the magnetic force per unit charge. However, if we think of the lone charge in our example as the test charge, we'll find that this approach fails, because the force depends not just on the test particle's 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, $\mathbf{v}$, 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 $\mathbf{B}$, and the force on a test particle moving through a magnetic field is proportional both to the $\mathbf{B}$ vector and to the particle's own $\mathbf{v}$ vector. In other words, the magnetic force vector $\mathbf{F}$ is found by some sort of vector multiplication of the vectors $\mathbf{v}$ and $\mathbf{B}$. As proved on page 694, 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, $\mathbf{B}$, as the vector that determines the force on a charged particle according to the following rule:

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B} \qquad \text{[definition of the magnetic field]}$$

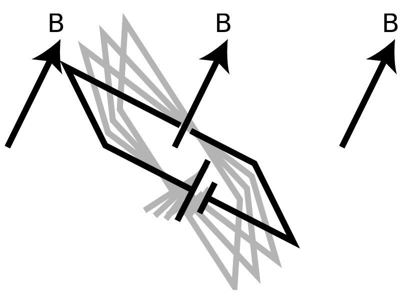
From this definition, we see that the magnetic field's units are $\text{N}\cdot\text{s}/\text{C}\cdot\text{m}$, which are usually abbreviated as teslas, $1\ \text{T} = 1\ \text{N}\cdot$

s/C·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 / 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.

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.

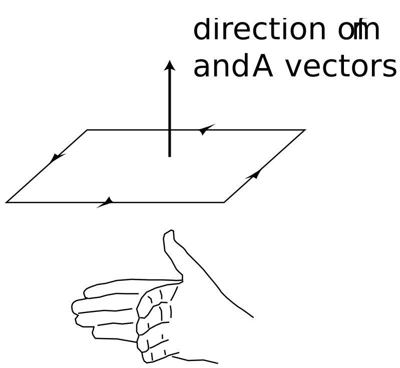


f / A dipole tends to align itself to the surrounding magnetic field.

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 460, we gave an alternative definition of the electric field in terms of the torque on an electric dipole.



g / The **m** and **A** vectors.

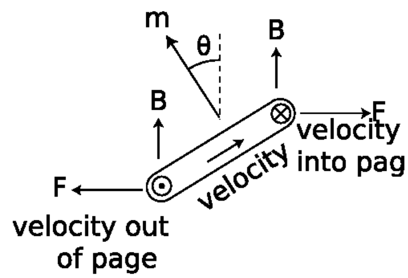
To define the strength of a magnetic field, however, we need some way of defining the strength of a test dipole, i.e. we need a definition of the magnetic dipole moment. We could use an iron permanent magnet constructed according to certain specifications, but such an object is really an extremely complex 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, e.

Applying $\mathbf{F} = \mathbf{v} \times \mathbf{B}$, we find that such a loop, when placed in a magnetic field, f, experiences a torque that tends to align plane so that its interior "face" points in a certain direction. 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

$$\mathbf{m} = I\mathbf{A} \qquad ,$$

where the direction of the vectors is defined as shown in figure g.



h / 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.

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

$$|B| = \frac{\tau}{|\mathbf{m}_t|\sin\theta} \qquad ,$$

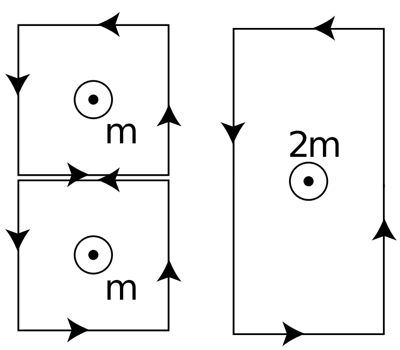
where θ is the angle between the dipole vector and the field.

$\odot$ out of the page

$\otimes$ into the page

i / 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's show that this is consistent with the previous definition, using the geometry shown in figure h. The velocity vector that point in and out of the page are shown using the convention defined in figure i. 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

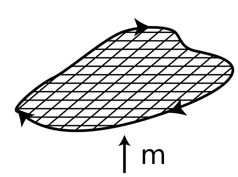


j / Dipole vectors can be added.

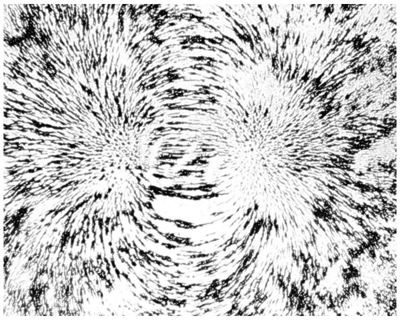
$$\begin{aligned} |\boldsymbol{\tau}| &= 2\frac{h}{2}|\mathbf{F}|\sin\theta \\ &= h\,qvB\,\sin\theta \qquad , \end{aligned}$$

and substituting $q = \lambda h$ and $v = m/h^2\lambda$, we have

$$\begin{aligned} |\boldsymbol{\tau}| &= h\,\lambda h\,\frac{m}{h^2\lambda}B\sin\theta \\ &= mB\sin\theta \qquad , \end{aligned}$$



k / An irregular loop can be broken up into little squares.



l / 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.

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 j. 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 k, so the dipole moment of any planar current loop can be calculated based on its area, regardless of its shape.

The magnetic dipole moment of an atom — *example 1*

Let's make an order-of-magnitude estimate of the magnetic dipole moment of an atom. A hydrogen atom is about 10^{-10} m in diameter, and the electron moves at speeds of about $10^{-2}c$. We don't know the shape of the orbit, and indeed it turns out that according to the principles of quantum mechanics, the electron doesn't even have a well-defined orbit, but if we're brave, we can still estimate the dipole moment using the cross-sectional area of the atom, which will be on the order of $(10^{-10}\ \text{m})^2 = 10^{-20}\ \text{m}^2$. The electron is a single particle, not a steady current, but again we throw caution to the winds, and estimate the current it creates as $e/\Delta t$, where Δt, the time for one orbit, can be estimated by dividing the size of the atom by the electron's velocity. (This is only a rough estimate, and we don't know the shape of the orbit, so it would be silly, for instance, to bother with multiplying the diameter by π based on our intuitive visualization of the electron as moving around the circumference of a circle.) The result for the dipole moment is $m \sim 10^{-23}\ \text{A}\cdot\text{m}^2$.

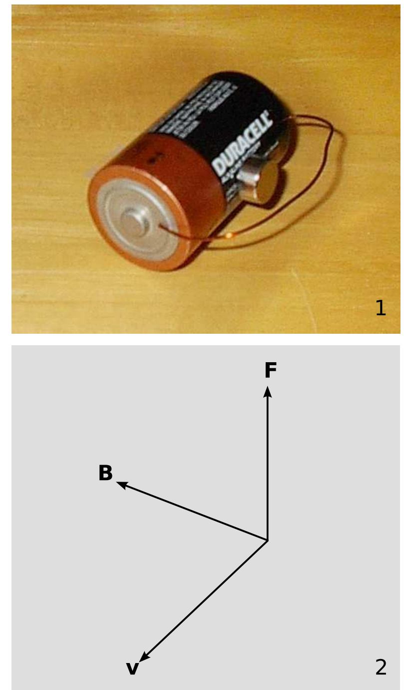
Should we be impressed with how small this dipole moment is, or with how big it is, considering that it's being made by a single atom? Very large or very small numbers are never very interesting by themselves. To get a feeling for what they mean, we need to compare them to something else. An interesting comparison here is to think in terms of the total number of atoms in a typical object, which might be on the order of 10^{26} (Avogadro's number). Suppose we had this many atoms, with their moments all aligned. The total dipole moment would be on the order of $10^3\ \text{A}\cdot\text{m}^2$, which is a pretty big number. To get a dipole moment this strong using human-scale devices, we'd have to send a thousand amps of current through a one-square meter loop of wire! The insight to be gained here is that, even in a permanent magnet, we must not have all the atoms perfectly aligned, because that would cause more spectacular magnetic effects than we really observe. Apparently, nearly all the atoms in such a magnet are oriented randomly, and do not contribute to the magnet's dipole moment.

Discussion Questions

A The physical situation shown in figure c on page 540 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 **E** and **B**? How does this compare with the forces?

B 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 **E** and **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



m / Example 2.

Magnetic levitation *example 2*

In figure m, 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.

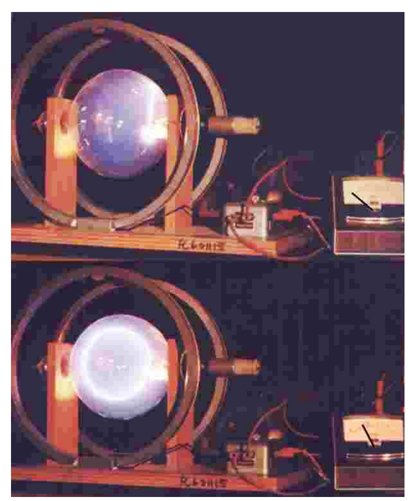
From the photo, it's possible to find the direction of the magnetic field made by the permanent magnet. The electrons in the copper wire are negatively charged, so they flow from the negative (flat) terminal of the battery to the positive terminal (the one with the bump, in front). As the electrons pass 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 543), the field must be toward the battery.

A circular orbit *example 3*

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.

Hallucinations during an MRI scan *example 4*

During an MRI scan of the head, the patient's nervous system is exposed to intense magnetic fields. The average velocities of the charge-carrying ions in the nerve cells is fairly low, but if the patient moves her head suddenly, the velocity can be high enough that the magnetic field makes significant forces on the ions. This can result in visual and auditory hallucinations, e.g., frying bacon sounds.



n / Magnetic forces cause a beam of electrons to move in a circle.

Self-Check

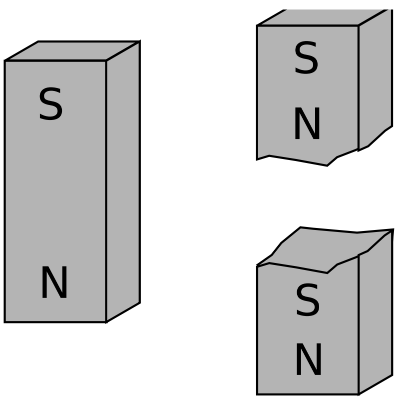
In what direction is the magnetic field in figure n? Don't forget that the beam is made of electrons, which are negatively charged! ▷ Answer, p. 710

Homework problem 13 is a quantitative analysis of circular orbits.

A velocity filter *example 5*

Suppose you see the electron beam in figure n, 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 14 is an analysis of this application.

11.1.4 No magnetic monopoles

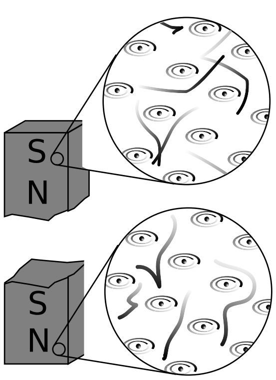


o / You can't isolate the poles of a magnet by breaking it in half.

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, o, you will find that you have simply made two smaller two-ended objects.

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, p.



p / 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.

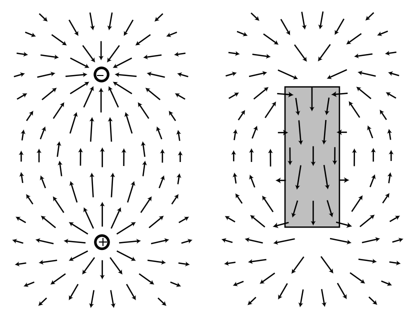
Electric and magnetic forces seem similar in many ways. Both act at a distance, both can be either attractive or repulsive, and both are intimately related to the property of matter called charge.

(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 1960s 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.

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 q. 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.



q / Magnetic fields have no sources or sinks.

Self-Check

Draw a Gaussian surface on the electric dipole field of figure q that has nonzero electric flux through it, and then draw a similar surface on the magnetic field pattern. What happens? ▷ Answer, p. 710

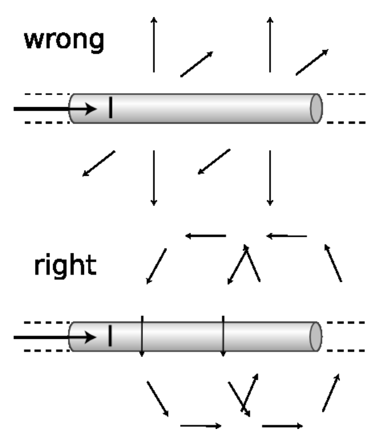
The field of a wire *example 6*

▷ On page 542, we showed that a long, straight wire carrying current I exerts a magnetic force

$$F = \frac{2kIqv}{c^2 R}$$

on a particle with charge q moving parallel to the wire with velocity v. What, then, is the magnetic field of the wire?

▷ 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.



r / Example 6.

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 r. 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

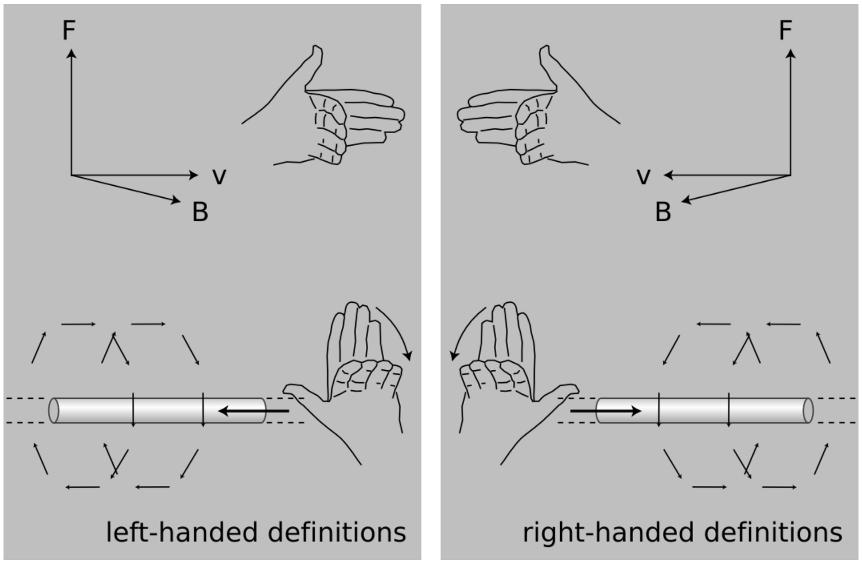
$$\begin{aligned}|B| &= \frac{|F|}{|q||v|}\\ &= \frac{2kI}{c^2R}\end{aligned}$$

11.1.5 Symmetry and handedness

The physicist Richard Feynman helped to get me hooked on physics with an educational film containing the following puzzle. 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?

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 r 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

[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.



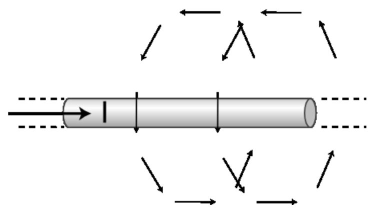
s / Left-handed and right-handed definitions.

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 s.

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. An asymmetry would be grotesque. Whatever their aesthetic feelings, they had to change their opinions about reality when experiments showed that the weak nuclear force violates right-left symmetry! 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.

11.2 Magnetic Fields by Superposition

11.2.1 Superposition of straight wires

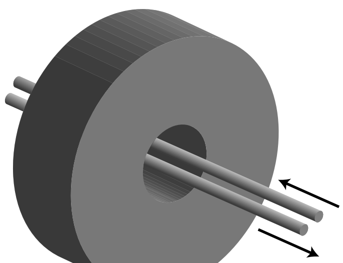


a / The magnetic field of a long, straight wire.

In chapter 10, one of the most important goals was to learn how to calculate the electric field for a given charge distribution. The corresponding problem for magnetism would be to calculate the magnetic field arising from a given set of currents. So far, however, we only know how to calculate the magnetic field of a long, straight wire,

$$B = \frac{2kI}{c^2 R} \qquad ,$$

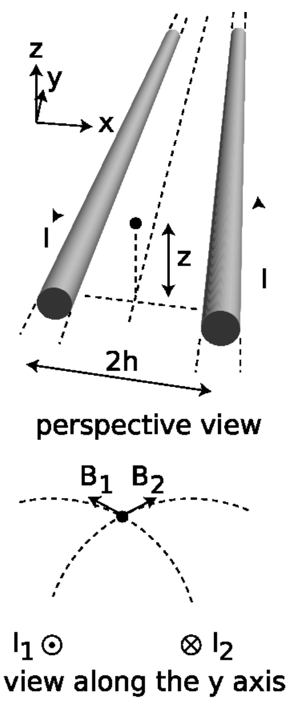
with the geometry shown in figure a. Whereas a charge distribution can be broken down into individual point charges, most currents cannot be broken down into a set of straight-line currents. Nevertheless, let's see what we can do with the tools that we have.



b / A ground fault interrupter.

A ground fault interrupter *example 7*

Electric current in your home is supposed to flow out of one side of the outlet, through an appliance, and back into the wall through the other side of the outlet. If that's not what happens, then we have a problem — the current must be finding its way to ground through some other path, perhaps through someone's body. If you have outlets in your home that have "test" and "reset" buttons on them, they have a safety device built into them that is meant to protect you in this situation. The ground fault interrupter (GFI) shown in figure b, routes the outgoing and returning currents through two wires that lie very close together. The clockwise and counterclockwise fields created by the two wires combine by vector addition, and normally cancel out almost exactly. However, if current is not coming back through the circuit, a magnetic field is produced. The doughnut-shaped collar detects this field (using an effect called induction, to be discussed in section 11.5), and sends a signal to a logic chip, which breaks the circuit within about 25 milliseconds.



c / Example 8.

An example with vector addition *example 8*

▷ Two long, straight wires each carry current I parallel to the y axis, but in opposite directions. They are separated by a gap $2h$ in the x direction. Find the magnitude and direction of the magnetic field at a point located at a height z above the plane of the wires, directly above the center line.

▷ The magnetic fields contributed by the two wires add like vectors, which means we can add their x and z components. The x components cancel by symmetry. The magnitudes of the individual fields are equal,

$$B_1 = B_2 = \frac{2kI}{c^2 R} \qquad ,$$

so the total field in the z direction is

$$B_z = 2\frac{2kI}{c^2 R}\sin\theta \qquad ,$$

where θ is the angle the field vectors make above the x axis. The sine of this angle equals h/R, so

$$B_z = \frac{4kIh}{c^2 R^2}.$$

(Putting this explicitly in terms of z gives the less attractive form $B_z = 4kIh/c^2(h^2+z^2)$.)

At large distances from the wires, the individual fields are mostly in the $\pm x$ direction, so most of their strength cancels out. It's not surprising that the fields tend to cancel, since the currents are in opposite directions. What's more interesting is that not only is the field weaker than the field of one wire, it also falls off as R^{-2} rather than R^{-1}. If the wires were right on top of each other, their currents would cancel each other out, and the field would be zero. From far away, the wires appear to be almost on top of each other, which is what leads to the more drastic R^{-2} dependence on distance.

Self-Check

In example 8, what is the field right between the wires, at $z = 0$, and how does this simpler result follow from vector addition? ▷ Answer, p. 710

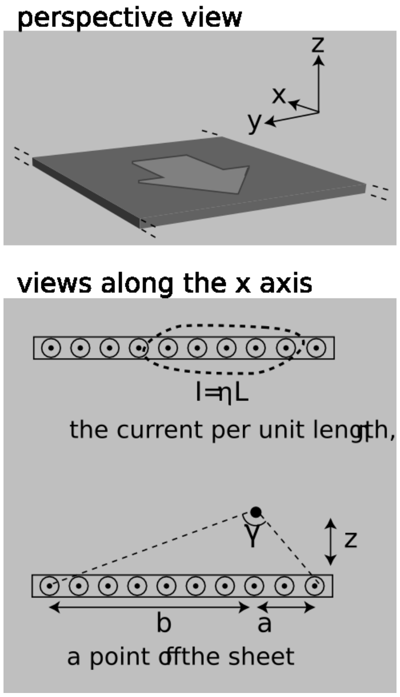
An alarming infinity

An interesting aspect of the R^{-2} dependence of the field in example 8 is the energy of the field. We've already established in the preceding chapter that the energy density of the magnetic field must be proportional to the square of the field strength, B^2, the same as for the gravitational and electric fields. Suppose we try to calculate the energy per unit length stored in the field of a *single* wire. We haven't yet found the proportionality factor that goes in front of the B^2, but that doesn't matter, because the energy per unit length turns out to be infinite! To see this, we can construct concentric cylindrical shells of length L, with each shell extending from R to $R+\mathrm{d}R$. The volume of the shell equals its circumference times its thickness times its length, $\mathrm{d}v = (2\pi R)(\mathrm{d}R)(L) = 2\pi L\mathrm{d}R$. For a single wire, we have $B \sim R^{-1}$, so the energy density is proportional to R^{-2}, and the energy contained in each shell varies as $R^{-2}\mathrm{d}v \sim R^{-1}\mathrm{d}r$. Integrating this gives a logarithm, and as we let R approach infinity, we get the logarithm of infinity, which is infinite.

Taken at face value, this result would imply that electrical currents could never exist, since establishing one would require an infinite amount of energy per unit length! In reality, however, we would be dealing with an electric *circuit*, which would be more like the two wires of example 8: current goes out one wire, but comes back through the other. Since the field really falls off as R^{-2}, we have an energy density that varies as R^{-4}, which does *not* give infinity when integrated out to infinity. (There is still an infinity at $R = 0$, but this doesn't occur for a real wire, which has a finite diameter.)

Still, one might worry about the physical implications of the single-wire result. For instance, suppose we turn on an electron gun, like the one in a TV tube. It takes perhaps a microsecond for the beam to progress across the tube. After it hits the other side of the tube, a return current is established, but at least for the first microsecond, we have only a single current, not two. Do we have infinite energy in the resulting magnetic field? No. It takes time for electric and magnetic field disturbances to travel outward through space, so during that microsecond, the field spreads only to some finite value of R, not $R = \infty$.

This reminds us of an important fact about our study of magnetism so far: we have only been considering situations where the currents and magnetic fields are constant over time. The equation $B = 2kI/c^2R$ was derived under this assumption. This equation is only valid if we assume the current has been established and flowing steadily for a long time, and if we are talking about the field at a point in space at which the field has been established for a long time. The generalization to time-varying fields is nontrivial, and qualitatively new effects will crop up. We have already seen one example of this on page 489, where we inferred that an inductor's time-varying magnetic field creates an electric field — an electric field which is not created by any charges anywhere. Effects like these will be discussed in section 11.5



d / A sheet of charge.

A sheet of current

There is a saying that in computer science, there are only three nice numbers: zero, one, and however many you please. In other words, computer software shouldn't have arbitrary limitations like a maximum of 16 open files, or 256 e-mail messages per mailbox. When superposing the fields of long, straight wires, the really interesting cases are one wire, two wires, and infinitely many wires. With an infinite number of wires, each carrying an infinitesimal current, we can create sheets of current, as in figure d. Such a sheet has a certain amount of current per unit length, which we notate η (Greek letter eta). The setup is similar to example 8, except that all the currents are in the same direction, and instead of adding up two fields, we add up an infinite number of them by doing an integral. For the y component, we have

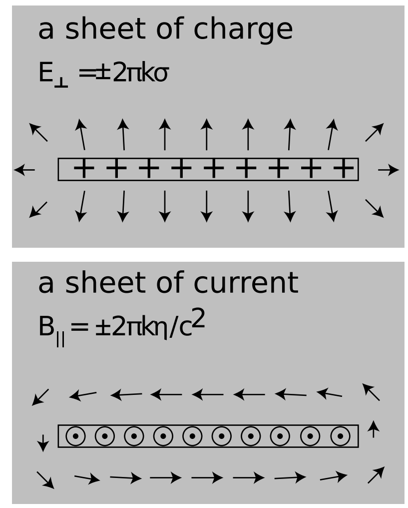
$$\begin{aligned}
B_y &= \int \frac{2k\,\mathrm{d}I}{c^2R}\cos\theta \\
&= \int_{-a}^{b} \frac{2k\eta\,\mathrm{d}y}{c^2R}\cos\theta \\
&= \frac{2k\eta}{c^2}\int_{-a}^{b}\frac{\cos\theta}{R}\,\mathrm{d}y \\
&= \frac{2k\eta}{c^2}\int_{-a}^{b}\frac{z\,\mathrm{d}y}{y^2+z^2} \\
&= \frac{2k\eta}{c^2}\left(\tan^{-1}\frac{b}{z} - \tan^{-1}\frac{-a}{z}\right) \\
&= \frac{2k\eta\gamma}{c^2} \qquad ,
\end{aligned}$$

where in the last step we have used the identity $\tan^{-1}(-x) = -\tan^{-1}x$, combined with the relation $\tan^{-1}b/z + \tan^{-1}a/z = \gamma$, which can be verified with a little geometry and trigonometry. The calculation of B_z is left as an exercise (problem 17). More interesting is what happens underneath the sheet: by the right-hand rule, all the currents make rightward contributions to the field there, so B_y abruptly reverses itself as we pass through the sheet.

Close to the sheet, the angle γ approaches π, so we have

$$B_y = \frac{2\pi k\eta}{c^2} \qquad .$$

Figure e shows the similarity between this result and the result for a sheet of charge. In one case the sources are charges and the field is electric; in the other case we have currents and magnetic fields. In both cases we find that the field changes suddenly when we pass through a sheet of sources, and the amount of this change doesn't depend on the size of the sheet. It was this type of reasoning that eventually led us to Gauss' law in the case of electricity, and in section 11.3 we will see that a similar approach can be used with magnetism. The difference is that, whereas Gauss' law involves the flux, a measure of how much the field *spreads out*, the corresponding law for magnetism will measure how much the field *curls*.



e / A sheet of charge and a sheet of current.

Is it just dumb luck that the magnetic-field case came out so similar to the electric field case? Not at all. We've already seen that what one observer perceives as an electric field, another observer may perceive as a magnetic field. An observer flying along above a charged sheet will say that the charges are in motion, and will therefore say that it is both a sheet of current and a sheet of charge. Instead of a pure electric field, this observer will experience a combination of an electric field and a magnetic one. (We could also construct an example like the one in figure c on page 540, in which the field was purely magnetic.)

11.2.2 Energy in the magnetic field

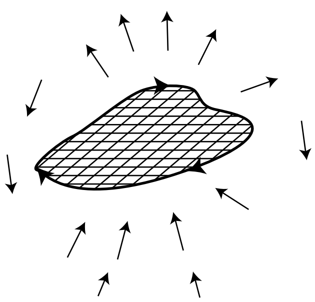
In section 10.4, I've already argued that the energy density of the magnetic field must be proportional to $|\mathbf{B}|^2$, which we can write as B^2 for convenience. To pin down the constant of proportionality, we now need to do something like the argument on page 10.4: find one example where we can calculate the mechanical work done by the magnetic field, and equate that to the amount of energy lost by the field itself. The easiest example is two parallel sheets of charge, with their currents in opposite directions. Homework problem 42 is such a calculation, which gives the result

$$\mathrm{d}U_m = \frac{c^2}{8\pi k}B^2\mathrm{d}v \qquad .$$

11.2.3 Superposition of dipoles

To understand this subsection, you'll have to have studied section 4.2.4, on iterated integrals.

The distant field of a dipole, in its midplane



f / The field of any planar current loop can be found by breaking it down into square dipoles.

Most current distributions cannot be broken down into long, straight wires, and subsection 11.2.1 has exhausted most of the interesting cases we can handle in this way. A much more useful building block is a square current loop. We have already seen how the dipole moment of an irregular current loop can be found by breaking the loop down into square dipoles (figure k on page 545), because the currents in adjoining squares cancel out on their shared edges. Likewise, as shown in figure f, if we could find the magnetic field of a square dipole, then we could find the field of any planar loop of current by adding the contributions to the field from all the squares.

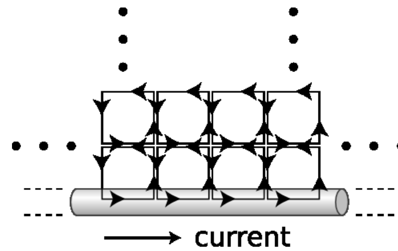
The field of a square-loop dipole is very complicated close up, but luckily for us, we only need to know the current at distances that are large compared to the size of the loop, because we're free to make the squares on our grid as small as we like. The *distant* field of a square dipole turns out to be simple, and is no different from the distant field of any other dipole with the same dipole moment. We can also save ourselves some work if we only worry about finding the field of the dipole in its own plane, i.e. the plane perpendicular to its dipole moment. By symmetry, the field in this plane cannot have any component in the radial direction (inward toward the dipole, or outward away from it); it is perpendicular to the plane, and in the opposite direction compared to the dipole vector. (The field *inside* the loop is in the same direction as the dipole vector, but we're interested in the distant field.) Letting the dipole vector be along the z axis, we find that the field in the $x-y$ plane is of the form $B_z = f(r)$, where $f(r)$ is some function that depends only on r, the distance from the dipole.

We can pin down the result even more without any math. We know that the magnetic field made by a current always contains a factor of k/c^2, which is the coupling constant for magnetism. We also know that the field must be proportional to the dipole moment, $m = IA$. Fields are always directly proportional to currents, and the proportionality to area follows because dipoles add according to their area. For instance, a square dipole that is 2 micrometers by 2 micrometers in size can be cut up into four dipoles that are 1 micrometer on a side. This tells us that our result must be of the form $B_z = (k/c^2)(IA)g(r)$. Now if we multiply the quantity $(k/c^2)(IA)$ by the function $g(r)$, we have to get units of teslas, and this only works out if $g(r)$ has units of m^{-3} (homework problem 22), so our result must be of the form

$$B_z = \frac{\beta k I A}{c^2 r^3} \qquad ,$$

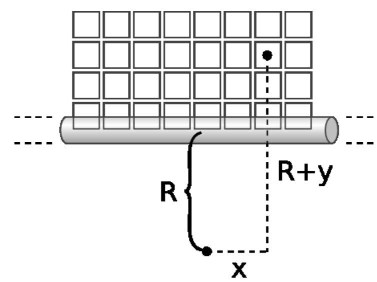
where β is a unitless constant. Thus our only task is to determine β, and we will have determined the field of the dipole.

If we wanted to, we could simply build a dipole, measure its field, and determine β empirically. Better yet, we can get an exact result if we take a current loop whose field we know exactly, break it down into infinitesimally small squares, integrate to find the total field, set this result equal to the known expression for the field of the loop, and solve for β. There's just one problem here. We don't yet know an expression for the field of *any* current loop of *any* shape — all we know is the field of a long, straight wire. Are we out of luck? No, because, as shown in figure g, we can make a long, straight wire by putting together square dipoles! Any square dipole away from the edge has all four of its currents canceled by its neighbors. The only currents that don't cancel are the ones on the edge, so by superimposing all the square dipoles, we get a straight-line current.



g / A long, straight current-carrying wire can be constructed by filling half of a plane with square dipoles.

This might seem strange. If the squares on the interior have all their currents canceled out by their neighbors, why do we even need them? Well, we need the squares on the edge in order to make the straight-line current. We need the second row of squares to cancel out the currents at the top of the first row of squares, and so on.



h / Setting up the integral.

Integrating as shown in figure h, we have

$$B_z = \int_{y=0}^{\infty}\int_{x=-\infty}^{\infty} \mathrm{d}B_z \qquad ,$$

where $\mathrm{d}B_z$ is the contribution to the total magnetic field at our point of interest, which lies a distance R from the wire.

$$\begin{aligned}
B_z &= \int_{y=0}^{\infty}\int_{x=-\infty}^{\infty} \frac{\beta k I\,\mathrm{d}A}{c^2 r^3} \\
&= \frac{\beta k I}{c^2}\int_{y=0}^{\infty}\int_{x=-\infty}^{\infty} \frac{1}{[x^2+(R+y)^2]^{3/2}}\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{\beta k I}{c^2 R^3}\int_{y=0}^{\infty}\int_{x=-\infty}^{\infty} \left[\left(\frac{x}{R}\right)^2+\left(1+\frac{y}{R}\right)^2\right]^{-3/2}\,\mathrm{d}x\,\mathrm{d}y
\end{aligned}$$

This can be simplified with the substitutions $x = Ru$, $y = Rv$, and $\mathrm{d}x\,\mathrm{d}y = R^2\,\mathrm{d}u\,\mathrm{d}v$:

$$B_z = \frac{\beta k I}{c^2 R}\int_{v=0}^{\infty}\int_{u=-\infty}^{\infty} \frac{1}{[u^2+(1+v)^2]^{3/2}}\,\mathrm{d}u\,\mathrm{d}v$$

The u integral is of the form $\int_{-\infty}^{\infty}(u^2+b)^{-3/2}\,\mathrm{d}u = 2/b^2$, so

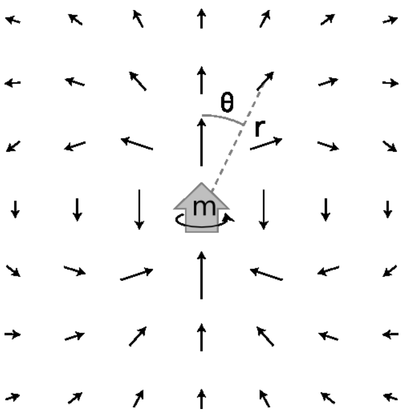
$$B_z = \frac{\beta k I}{c^2 R}\int_{v=0}^{\infty}\frac{1}{(1+v)^2}\,\mathrm{d}v \qquad ,$$

and the remaining v integral is equals 2, so

$$B_z = \frac{2\beta kI}{c^2R} \qquad .$$

This is the field of a wire, which we already know equals $2kI/c^2R$, so we have β=1. Remember, the point of this whole calculation was not to find the field of a *wire*, which we already knew, but to find the unitless constant β in the expression for the field of a *dipole*. The distant field of a dipole, in its midplane, is therefore $B_z = \beta kIA/c^2r^3 = kIA/c^2r^3$, or, in terms of the dipole moment,

$$B_z = \frac{km}{c^2r^3} \qquad .$$



i / The field of a dipole.

The distant field of a dipole, out of its midplane

What about the field of a magnetic dipole outside of the dipole's midplane? Let's compare with an electric dipole. An electric dipole, unlike a magnetic one, can be built out of two opposite monopoles, i.e. charges, separated by a certain distance, and it is then straightforward to show by vector addition that the field of an electric dipole is

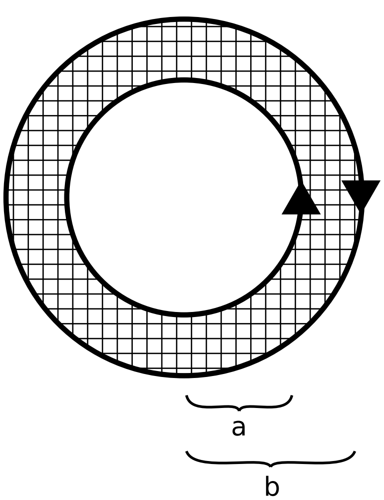
$$\begin{aligned} E_z &= kD\left(3\cos^2\theta - 1\right)r^{-3} \\ E_R &= kD\left(3\sin\theta\ \cos\theta\right)r^{-3} \qquad , \end{aligned}$$

where r is the distance from the dipole to the point of interest, θ is the angle between the dipole vector and the line connecting the dipole to this point, and E_z and E_R are, respectively, the components of the field parallel to and perpendicular to the dipole vector.

In the midplane, θ equals $\pi/2$, which produces $E_z = -kDr^{-3}$ and $E_R = 0$. This is the same as the field of a *magnetic* dipole in its midplane, except that the electric coupling constant k replaces the magnetic version k/c^2, and the electric dipole moment D is substituted for the magnetic dipole moment m. It is therefore reasonable to conjecture that by using the same presto-change-o recipe we can find the field of a magnetic dipole outside its midplane:

$$\begin{aligned} B_z &= \frac{km}{c^2}\left(3\cos^2\theta - 1\right)r^{-3} \\ B_R &= \frac{km}{c^2}\left(3\sin\theta\ \cos\theta\right)r^{-3} \qquad . \end{aligned}$$

This turns out to be correct. [5]



j / Example 9.

[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.

Concentric, counterrotating currents *example 9*

▷ Two concentric circular current loops, with radii a and b, carry the same amount of current I, but in opposite directions. What is the field at the center?

▷ We can produce these currents by tiling the region between the circles with square current loops, whose currents all cancel each other except at the inner and outer edges. The flavor of the calculation is the same as the one in which we made a line of current by filling a half-plane with square loops. The main difference is that this geometry has a different symmetry, so it will make more sense to use polar coordinates instead of x and y. The field at the center is

$$\begin{aligned} B_z &= \int \frac{kI}{c^2 r^3}\,\mathrm{d}A \\ &= \int_{r=a}^{b} \frac{kI}{c^2 r^3} \cdot 2\pi r \mathrm{d}r \\ &= \frac{2\pi k I}{c^2}\left(\frac{1}{a}-\frac{1}{b}\right) \quad . \end{aligned}$$

The positive sign indicates that the field is out of the page.

Field at the center of a circular loop *example 10*

▷ What is the magnetic field at the center of a circular current loop of radius a, which carries a current I?

▷ This is like example 9, but with the outer loop being very large, and therefore too distant to make a significant field at the center. Taking the limit of that result as b approaches infinity, we have

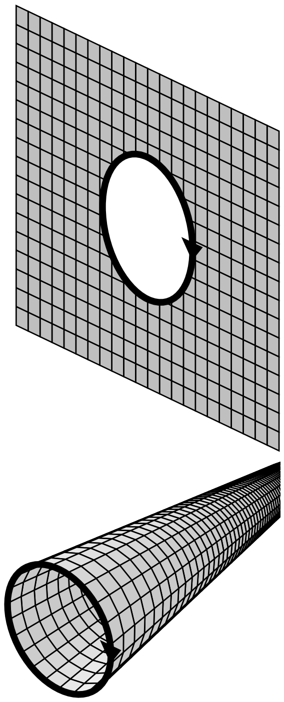
$$B_z = \frac{2\pi k I}{c^2 a}$$

Comparing the results of examples 9 and 10, we see that the directions of the fields are both out of the page. In example 9, the outer loop has a current in the opposite direction, so it contributes a field that is into the page. This, however, is weaker than the field due to the inner loop, which dominates because it is less distant.

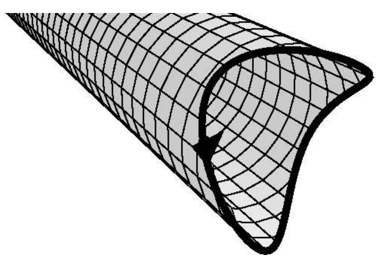
11.2.4 The Biot-Savart law (optional)

In section 11.2.3 we developed a method for finding the field due to a given current distribution by tiling a plane with square dipoles. This method has several disadvantages:

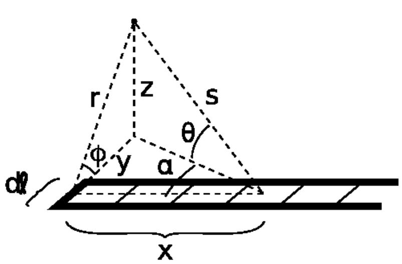
- The currents all have to lie in a single plane, and the point at which we're computing the field must be in that plane as well.
- We need to do integral over an area, which means one integral inside another, e.g. $\int\int \ldots \mathrm{d}x\mathrm{d}y$. That can get messy.
- It's physically bizarre to have to construct square dipoles in places where there really aren't any currents.



k / Two ways of making a current loop out of square dipoles.



l / The new method can handle non-planar currents.



m / The field of an infinite U.

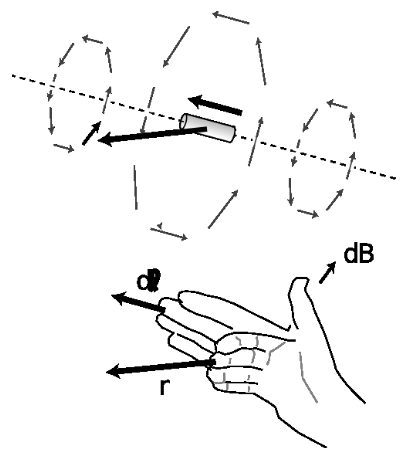
Figure k shows the first step in eliminating these defects: instead of spreading our dipoles out in a plane, we bring them out along an axis. As shown in figure l, this eliminates the restriction to currents that lie in a plane. Now we have to use the general equations for a dipole field from page 558, rather than the simpler expression for the field in the midplane of a dipole. This increase in complication is more than compensated for by a fortunate feature of the new geometry, which is that the infinite tube can be broken down into strips, and we can find the field of such a strip for once and for all. This means that we no longer have to do one integral inside another. The derivation of the most general case is a little messy, so I'll just present the case shown in figure m, where the point of interest is assumed to lie in the $y-z$ plane. Intuitively, what we're really finding is the field of the short piece of length $\mathrm{d}\ell$ on the end of the U; the two long parallel segments are going to be canceled out by their neighbors when we assemble all the strips to make the tube. We expect that the field of this end-piece will form a field pattern that circulates around the y axis, so at the point of interest, it's really the x component of the field that we want to compute:

$$\begin{aligned}
\mathrm{d}B_x &= \int \mathrm{d}B_R \cos\alpha \\
&= \int \frac{kI\,\mathrm{d}\ell\,\mathrm{d}x}{c^2 s^3}(3\sin\theta\cos\theta\cos\alpha) \\
&= \frac{3kI\,\mathrm{d}\ell}{c^2}\int_0^\infty \frac{1}{s^3}\left(\frac{xz}{s^2}\right)\,\mathrm{d}x \\
&= \frac{3kIz\,\mathrm{d}\ell}{c^2}\int_0^\infty \frac{x}{(x^2+r^2)^{5/2}}\,\mathrm{d}x \\
&= \frac{kI\,\mathrm{d}\ell\,z}{c^2r^3} \\
&= \frac{kI\,\mathrm{d}\ell\,\sin\phi}{c^2r^2}
\end{aligned}$$

In the more general case, l,the current loop is not planar, the point of interest is not in the end-planes of the U's, and the U shapes have their ends staggered, so the end-piece $d\ell$ is not the only part of each U whose current is not canceled. Without going into the gory details, the correct general result is as follows:

$$dB = \frac{kI\, d\boldsymbol{\ell} \times \mathbf{r}}{c^2 r^3} \qquad ,$$

which is known as the Biot-Savart law. (It rhymes with “leo bazaar.” Both t’s are silent.) The distances $d\ell$ and r are now defined as vectors, $d\boldsymbol{\ell}$ and $\mathbf{r}$, which point, respectively, in the direction of the current in the end-piece and the direction from the end-piece to the point of interest. The new equation looks different, but it is consistent with the old one. The vector cross product $d\boldsymbol{\ell} \times \mathbf{r}$ has a magnitude $r\, d\ell\, \sin\phi$, which cancels one of r’s in the denominator and makes the $d\boldsymbol{\ell} \times \mathbf{r}/r^3$ into a vector with magnitude $d\ell\, \sin\phi/r^2$.

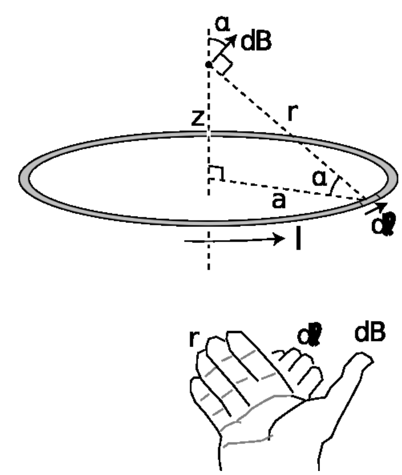


n / The geometry of the Biot-Savart law. The small arrows show the result of the Biot-Savart law at various positions relative to the current segment $d\ell$. The Biot-Savart law involves a cross product, and the right-hand rule for this cross product is demonstrated for one case.

The field at the center of a circular loop — *example 11*

Previously we had to do quite a bit of work (examples 9 and 10), to calculate the field at the center of a circular loop of current of radius a. It’s much easier now. Dividing the loop into many short segments, each $d\ell$ is perpendicular to the $\mathbf{r}$ vector that goes from it to the center of the circle, and every $\mathbf{r}$ vector has magnitude a. Therefore every cross product $d\boldsymbol{\ell} \times \mathbf{r}$ has the same magnitude, $a\, d\ell$, as well as the same direction along the axis perpendicular to the loop. The field is

$$\begin{aligned} B &= \int \frac{kIa\, d\ell}{c^2 a^3} \\ &= \frac{kI}{c^2 a^2} \int d\ell \\ &= \frac{kI}{c^2 a^2}(2\pi a) \\ &= \frac{2\pi kI}{c^2 a} \end{aligned}$$



o / Example 12.

Out-of-the-plane field of a circular loop — *example 12*

▷ What is the magnetic field of a circular loop of current at a point on the axis perpendicular to the loop, lying a distance z from the loop’s center?

▷ Again, let’s write a for the loop’s radius. The $\mathbf{r}$ vector now has magnitude $\sqrt{a^2+z^2}$, but it is still perpendicular to the $d\ell$ vector. By symmetry, the only nonvanishing component of the field is along the z axis,

$$\begin{aligned} B_z &= \int |d\mathbf{B}| \cos\alpha \\ &= \int \frac{kIr\, d\ell}{c^2 r^3} \frac{a}{r} \\ &= \frac{kIa}{c^2 r^3} \int d\ell \\ &= \frac{2\pi kIa^2}{c^2 (a^2+z^2)^{3/2}} \qquad . \end{aligned}$$

Is it the field of a particle?

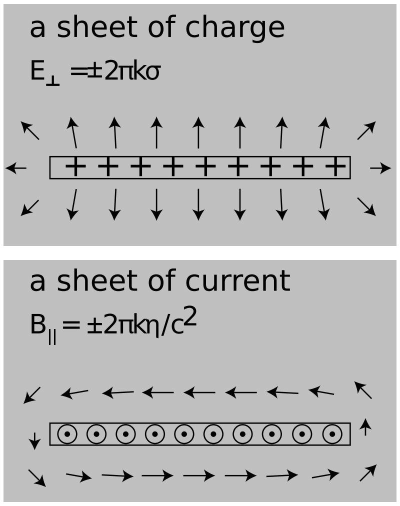
We have a simple equation, based on Coulomb's law, for the electric field surrounding a charged particle. Looking at figure n, we can imagine that if the current segment $d\ell$ was very short, then it might only contain one electron. It's tempting, then, to interpret the Biot-Savart law as a similar equation for the magnetic field surrounding a moving charged particle. Tempting but wrong! Suppose you stand at a certain point in space and watch a charged particle move by. It has an electric field, and since it's moving, you will also detect a magnetic field on top of that. Both of these fields change over time, however. Not only do they change their magnitudes and directions due to your changing geometric relationship to the particle, but they are also time-delayed, because disturbances in the electromagnetic field travel at the speed of light, which is finite. The fields you detect are the ones corresponding to where the particle used to be, not where it is now. Coulomb's law and the Biot-Savart law are both false in this situation, since neither equation includes time as a variable. It's valid to think of Coulomb's law as the equation for the field of a stationary charged particle, but not a moving one. The Biot-Savart law fails completely as a description of the field of a charged particle, since stationary particles don't make magnetic fields, and the Biot-Savart law fails in the case where the particle is moving.

If you look back at the long chain of reasoning that led to the Biot-Savart law, it all started from the relativistic arguments at the beginning of this chapter, where we assumed a steady current in an infinitely long wire. Everything that came later was built on this foundation, so all our reasoning depends on the assumption that the currents are steady. In a steady current, any charge that moves away from a certain spot is replaced by more charge coming up behind it, so even though the charges are all moving, the electric and magnetic fields they produce are constant. Problems of this type are called electrostatics and magnetostatics problems, and it is only for these problems that Coulomb's law and the Biot-Savart law are valid.

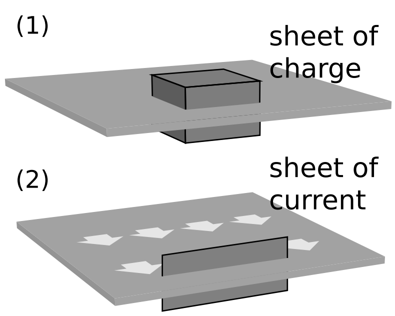
You might think that we could patch up Coulomb's law and the Biot-Savart law by inserting the appropriate time delays. However, we've already seen a clear example of a phenomenon that wouldn't be fixed by this patch: on page 489, we found that a changing magnetic field creates an electric field. Induction effects like these also lead to the existence of light, which is a wave disturbance in the electric and magnetic fields. We could try to apply another band-aid fix to Coulomb's law and the Biot-Savart law to make them deal with induction, but it won't work.

So what *are* the fundamental equations that describe how sources give rise to electromagnetic fields? We've already encountered two of them: Gauss' law for electricity and Gauss' law for magnetism.

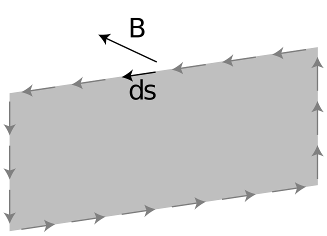
Experiments show that these are valid in all situations, not just static ones. But Gauss' law for magnetism merely says that the magnetic flux through a closed surface is zero. It doesn't tell us how to make magnetic fields using currents. It only tells us that we *can't* make them using magnetic monopoles. The following section develops a new equation, called Ampère's law, which is equivalent to the Biot-Savart law for magnetostatics, but which, unlike the Biot-Savart law, can easily be extended to nonstatic situations.



a / The electric field of a sheet of charge, and the magnetic field of a sheet of current.



b / A Gaussian surface and an Ampèrian surface.



c / The definition of the circulation, Γ.

11.3 Magnetic Fields by Ampère's Law

11.3.1 Ampère's law

As discussed at the end of subsection 11.2.4, our goal now is to find an equation for magnetism that, unlike the Biot-Savart law, will not end up being a dead end when we try to extend it to nonstatic situations.[6] Experiments show that Gauss' law is valid in both static and nonstatic situations, so it would be reasonable to look for an approach to magnetism that is similar to the way Gauss' law deals with electricity.

How can we do this? Figure a, reproduced from page e, is our roadmap. Electric fields spread out from charges. Magnetic fields curl around currents. In figure b/1, we define a Gaussian surface, and we define the flux in terms of the electric field pointing out through this surface. In the magnetic case, b/2, we define a surface, called an Ampèrian surface, and we define a quantity called the circulation, Γ (uppercase Greek gamma), in terms of the magnetic field that points along the edge of the Ampèrian surface, c. We break the edge into tiny parts $\mathbf{s}_j$, and for each of these parts, we define a contribution to the circulation using the dot product of d$\mathbf{s}$ with the magnetic field:

$$\Gamma = \sum \mathbf{s}_j \cdot \mathbf{B}_j$$

The circulation is a measure of how curly the field is. Like a Gaussian surface, an Ampèrian surface is purely a mathematical construction. It is not a physical object.

In figure b/2, the field is perpendicular to the edges on the ends, but parallel to the top and bottom edges. A dot product is zero when the vectors are perpendicular, so only the top and bottom edges contribute to Γ. Let these edges have length s. Since the field is constant along both of these edges, we don't actually have to break them into tiny parts; we can just have $\mathbf{s}_1$ on the top edge, pointing to the left, and $\mathbf{s}_2$ on the bottom edge, pointing to the right. The vector $\mathbf{s}_1$ is in the same direction as the field $\mathbf{B}_1$, and $\mathbf{s}_2$ is in the same direction as $\mathbf{B}_2$, so the dot products are simply equal to the products of the vectors' magnitudes. The resulting circulation is

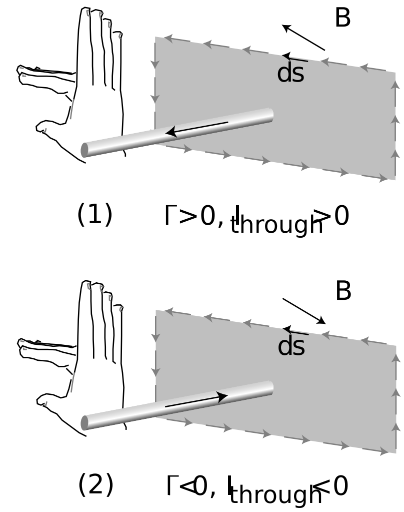
$$\begin{aligned}\Gamma &= |\mathbf{s}_1||\mathbf{B}_1| + |\mathbf{s}_2||\mathbf{B}_2| \\ &= \frac{2\pi k\eta s}{c^2} + \frac{2\pi k\eta s}{c^2} \\ &= \frac{4\pi k\eta s}{c^2} \qquad .\end{aligned}$$

But ηs is (current/length)(length), i.e. it is the amount of current that pierces the Ampèrian surface. We'll call this current $I_{through}$.

[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.

We have found one specific example of the general law of nature known as Ampère's law:

$$\Gamma = \frac{4\pi k}{c^2} I_{through}$$



d / Positive and negative signs in Ampère's law.

Positive and negative signs

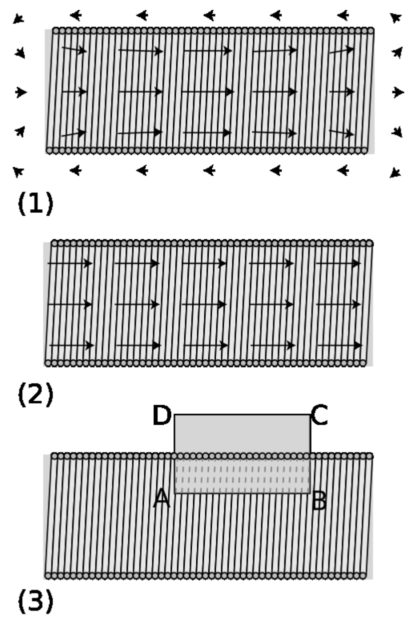
Figures d/1 and d/2 show what happens to the circulation when we reverse the direction of the current $I_{through}$. Reversing the current causes the magnetic field to reverse itself as well. The dot products occurring in the circulation are all negative in d/2, so the total circulation is now negative. To preserve Ampère's law, we need to define the current in d/2 as a negative number. In general, determine these plus and minus signs using the right-hand rule shown in the figure. As the fingers of your hand sweep around in the direction of the **s** vectors, your thumb defines the direction of current which is positive. Choosing the direction of the thumb is like choosing which way to insert an ammeter in a circuit: on a digital meter, reversing the connections gives readings which are opposite in sign.

A solenoid — *example 13*

▷ What is the field inside a long, straight solenoid of length ℓ and radius a, and having N loops of wire evenly wound along it, which carry a current I?

▷ This is an interesting example, because it allows us to get a very good approximation to the field, but without some experimental input it wouldn't be obvious what approximation to use. Figure e/1 shows what we'd observe by measuring the field of a real solenoid. The field is nearly constant inside the tube, as long as we stay far away from the mouths. The field outside is much weaker. For the sake of an approximate calculation, we can idealize this field as shown in figure e/2. Of the edges of the Ampèrian surface shown in e/3, only AB contributes to the flux — there is zero field along CD, and the field is perpendicular to edges BC and DA. Ampère's law gives

$$\begin{aligned}\Gamma &= \frac{4\pi k}{c^2} I_{through}\\ (B)(\text{length of AB}) &= \frac{4\pi k}{c^2}(\eta)(\text{length of AB})\\ B &= \frac{4\pi k\eta}{c^2}\\ &= \frac{4\pi kNI}{c^2\ell}\end{aligned}$$



e / Example 13: a cutaway view of a solenoid.

Self-Check

What direction is the current in figure e? ▷ Answer, p. 710

Self-Check

Based on how ℓ entered into the derivation in example 13, how should it be interpreted? Is it the total length of the wire? ▷ Answer, p. 710

Self-Check

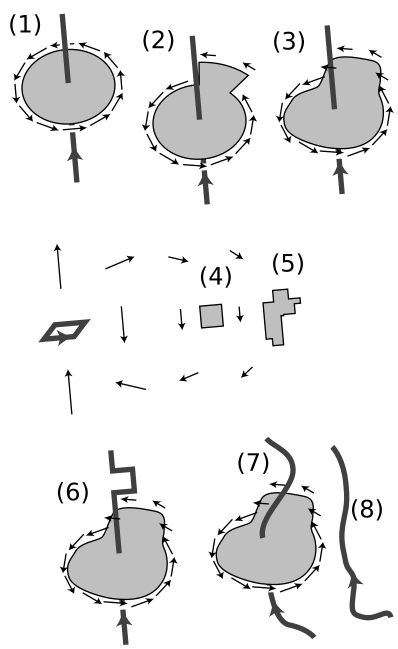
Surprisingly, we never needed to know the radius of the solenoid in example 13. Why is it physically plausible that the answer would be independent of the radius? ▷ Answer, p. 710

Example 13 shows how much easier it can sometimes be to calculate a field using Ampère's law rather than the approaches developed previously in this chapter. However, if we hadn't already known something about the field, we wouldn't have been able to get started. In situations that lack symmetry, Ampère's law may make things harder, not easier. Anyhow, we will have no choice in nonstatic cases, where Ampère's law is true, and static equations like the Biot-Savart law are false.

11.3.2 A quick and dirty proof

Here's an informal sketch for a proof of Ampère's law, with no pretensions to rigor. Even if you don't care much for proofs, it would be a good idea to read it, because it will help to build your ability to visualize how Ampère's law works.

First we establish by a direct computation (homework problem 31) that Ampère's law holds for the geometry shown in figure f/1, a circular Ampèrian surface with a wire passing perpendicularly through its center. If we then alter the surface as in figure f/2, Ampère's law still works, because the straight segments, being perpendicular to the field, don't contribute to the circulation, and the new arc makes the same contribution to the circulation as the old one it replaced, because the weaker field is compensated for by the greater length of the arc. It is clear that by a series of such modifications, we could mold the surface into any shape, f/3.



f / A proof of Ampère's law.

Next we prove Ampère's law in the case shown in figure f/4: a small, square Ampèrian surface subject to the field of a distant square dipole. This part of the proof can be most easily accomplished by the methods of section 11.4. It should, for example, be plausible in the case illustrated here. The field on the left edge is stronger than the field on the right, so the overall contribution of these two edges to the circulation is slightly counterclockwise. However, the field is not quite perpendicular to the top and bottoms edges, so they both make small clockwise contributions. The clockwise and counterclockwise parts of the circulation end up canceling each other out. Once Ampère's law is established for a square surface like f/4, it follows that it is true for an irregular surface like f/5, since we can build such a shape out of squares, and the circulations are additive when we paste the surfaces together this way.

By pasting a square dipole onto the wire, f/6, like a flag attached to a flagpole, we can cancel out a segment of the wire's current and create a detour. Ampère's law is still true because, as shown in the last step, the square dipole makes zero contribution to the circulation. We can make as many detours as we like in this manner, thereby morphing the wire into an arbitrary shape like f/7.

What about a wire like f/8? It doesn't pierce the Ampèrian surface, so it doesn't add anything to $I_{through}$, and we need to show that it likewise doesn't change the circulation. This wire, however, could be built by tiling the half-plane on its right with square dipoles, and we've already established that the field of a distant dipole doesn't contribute to the circulation. (Note that we couldn't have done this with a wire like f/7, because some of the dipoles would have been right on top of the Ampèrian surface.)

If Ampère's law holds for cases like f/7 and f/8, then it holds for any complex bundle of wires, including some that pass through the Ampèrian surface and some that don't. But we can build just about any static current distribution we like using such a bundle of wires, so it follows that Ampère's law is valid for any static current distribution.

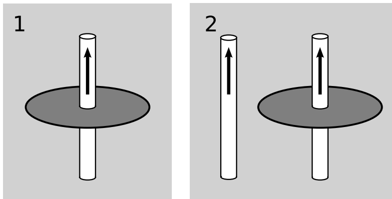
11.3.3 Maxwell's equations for static fields

Static electric fields *don't* curl the way magnetic fields do, so we can state a version of Ampère's law for electric fields, which is that the circulation of the electric field is zero. Summarizing what we know so far about static fields, we have

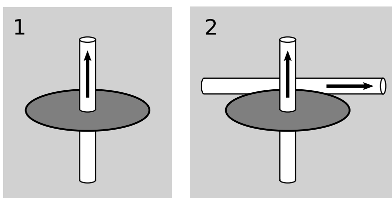
$$\begin{aligned}\Phi_E &= 4\pi k q_{in}\\ \Phi_B &= 0\\ \Gamma_E &= 0\\ \Gamma_B &= \frac{4\pi k}{c^2} I_{through} \qquad .\end{aligned}$$

This set of equations is the static case of the more general relations known as Maxwell's equations. On the left side of each equation, we have information about a field. On the right is information about the field's sources.

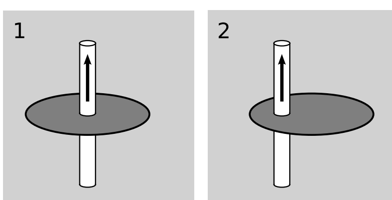
It is vitally important to realize that these equations are only true for statics. They are incorrect if the distribution of charges or currents is changing over time. For example, we saw on page 489 that the changing magnetic field in an inductor gives rise to an electric field. Such an effect is completely inconsistent with the static version of Maxwell's equations; the equations don't even refer to time, so if the magnetic field is changing over time, they will not do anything special. The extension of Maxwell's equations to nonstatic fields is discussed in section 11.6.



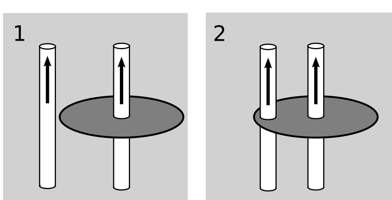
g / Discussion question A.



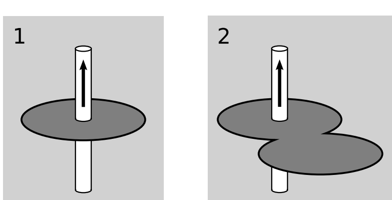
h / Discussion question B.



i / Discussion question C.



j / Discussion question D.



k / Discussion question E.

Discussion Questions

A Figure g/1 shows a wire with a circular Ampèrian surface drawn around its waist; in this situation, Ampère's law can be verified easily based on the equation for the field of a wire. In panel 2, a second wire has been added. Explain why it's plausible that Ampère's law still holds.

B Figure h is like figure g, but now the second wire is perpendicular to the first, and lies in the plane of, and outside of, the Ampèrian surface. Carry out a similar analysis.

C This discussion question is similar to questions A and B, but now the Ampèrian surface has been moved off center.

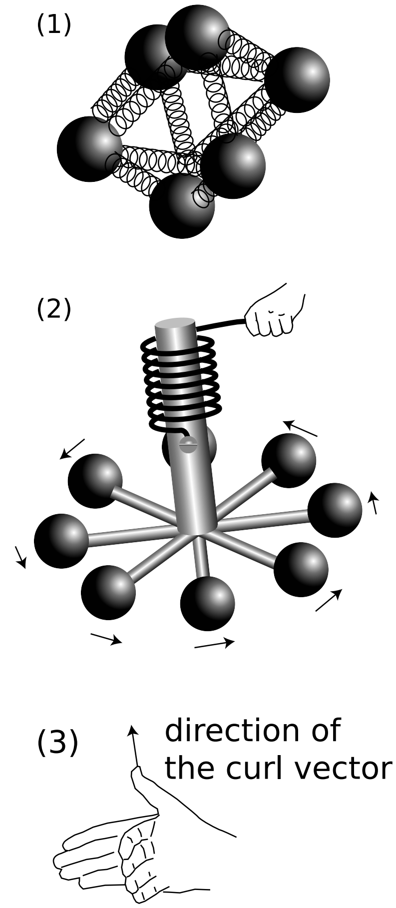
D The left-hand wire has been nudged over a little. Analyze as before.

E You know what to do.

11.4 Ampère's Law in Differential Form (optional)

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. 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.



a / The div-meter, 1, and the curl-meter, 2 and 3.

Gauss' law in differential form relates 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, $\mathrm{d}\Phi$ through a tiny cubical Gaussian surface containing a volume $\mathrm{d}v$. By analogy, we imagine a tiny square Ampèrian surface with infinitesimal area $\mathrm{d}\mathbf{A}$. We assume this surface has been oriented in order to get the maximum circulation. The area vector $\mathrm{d}\mathbf{A}$ will then be in the same direction as the one defined in figure a/3. Ampère's law is

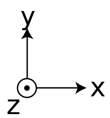
$$\mathrm{d}\Gamma = \frac{4\pi k}{c^2}\,\mathrm{d}I_{through} \qquad .$$

We define a current density per unit area, $\mathbf{j}$, which is a vector pointing in the direction of the current and having magnitude $\mathbf{j} = \mathrm{d}I/|\mathrm{d}\mathbf{A}|$. In terms of this quantity, we have

$$\begin{aligned}
\mathrm{d}\Gamma &= \frac{4\pi k}{c^2}\, j|\mathbf{j}|\,|\mathrm{d}\mathbf{A}| \\
\frac{\mathrm{d}\Gamma}{|\mathrm{d}\mathbf{A}|} &= \frac{4\pi k}{c^2}\,|\mathbf{j}|
\end{aligned}$$

With this motivation, we define the magnitude of the curl as

$$|\mathrm{curl}\,\mathbf{B}| = \frac{\mathrm{d}\Gamma}{|\mathrm{d}\mathbf{A}|} \qquad .$$

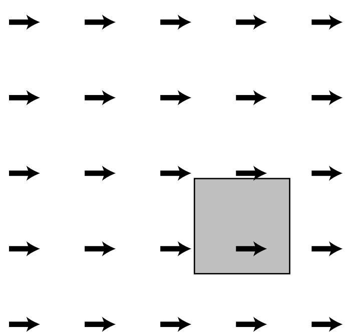


b / The coordinate system used in the following examples.

Note that the curl, just like a derivative, has a differential divided by another differential. In terms of this definition, we find Ampère's law in differential form:

$$\operatorname{curl} \mathbf{B} = \frac{4\pi k}{c^2}\mathbf{j}$$

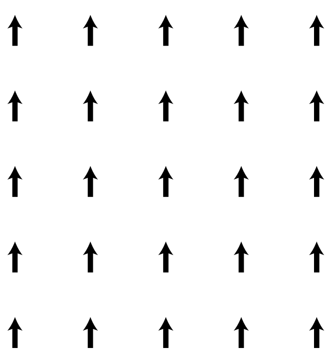
The complete set of Maxwell's equations in differential form is collected on page 699.


c / The field $\hat{\mathbf{x}}$.

11.4.2 Properties of the curl operator

The curl is a derivative.

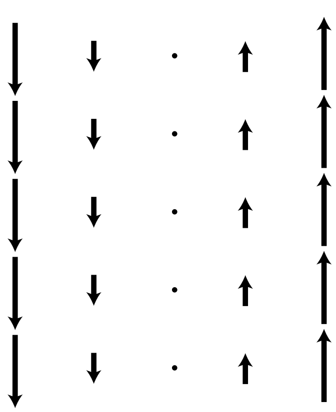
As an example, let's calculate the curl of the field $\hat{\mathbf{x}}$ shown in figure c. For our present purposes, it doesn't really matter whether this is an electric or a magnetic field; we're just getting out feet wet with the curl as a mathematical definition. Applying the definition of the curl directly, we construct an Ampèrian surface in the shape of an infinitesimally small square. Actually, since the field is uniform, it doesn't even matter very much whether we make the square finite or infinitesimal. The right and left edges don't contribute to the circulation, since the field is perpendicular to these edges. The top and bottom do contribute, but the top's contribution is clockwise, i.e. into the page according to the right-hand rule, while the bottom contributes an equal amount in the counterclockwise direction, which corresponds to an out-of-the-page contribution to the curl. They cancel, and the circulation is zero. We could also have determined this by imagining a curl-meter inserted in this field: the torques on it would have canceled out.


d / The field $\hat{\mathbf{y}}$.

It makes sense that the curl of a constant field is zero, because the curl is a kind of derivative. The derivative of a constant is zero.

The curl is rotationally invariant.

Figure c looks just like figure c, but rotated by 90 degrees. Physically, we could be viewing the same field from a point of view that was rotated. Since the laws of physics are the same regardless of rotation, the curl must be zero here as well. In other words, the curl is rotationally invariant. If a certain field has a certain curl vector, then viewed from some other angle, we simply see the same field and the same curl vector, viewed from a different angle. A zero vector viewed from a different angle is still a zero vector.

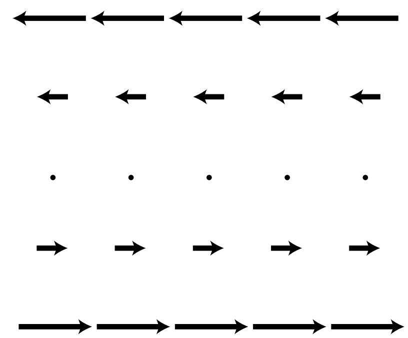

e / The field $x\hat{\mathbf{y}}$.

As a less trivial example, let's compute the curl of the field $\mathbf{F} = x\hat{\mathbf{y}}$ shown in figure e, at the point $(x = 0, y = 0)$. The circulation around a square of side s centered on the origin can be approximated by evaluating the field at the midpoints of its sides,

$$\begin{array}{llll} x=s/2 & y=0 & \mathbf{F}=(s/2)\hat{\mathbf{y}} & \mathbf{s}_1\cdot\mathbf{F}=s^2/2 \\ x=0 & y=s/2 & \mathbf{F}=0 & \mathbf{s}_2\cdot\mathbf{F}=0 \\ x=-s/2 & y=0 & \mathbf{F}=-(s/2)\hat{\mathbf{y}} & \mathbf{s}_3\cdot\mathbf{F}=s^2/2 \\ x=0 & y=-s/2 & \mathbf{F}=0 & \mathbf{s}_4\cdot\mathbf{F}=0 \end{array} \quad ,$$

which gives a circulation of s^2, and a curl with a magnitude of $s^2/\text{area}=s^2/s^2=1$. By the right-hand rule, the curl points out of the page, i.e. along the positive z axis, so we have

$$\operatorname{curl} x\hat{\mathbf{y}}=\hat{\mathbf{z}} \qquad .$$



f / The field $-y\hat{\mathbf{x}}$.

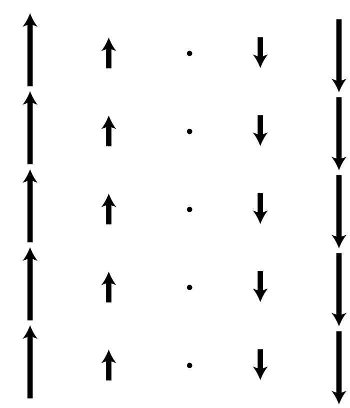
Now consider the field $-y\hat{\mathbf{x}}$, shown in figure f. This is the same as the previous field, but with your book rotated by 90 degrees about the z axis. Rotating the result of the first calculation, $\hat{\mathbf{z}}$, about the z axis doesn't change it, so the curl of this field is also $\hat{\mathbf{z}}$.

Scaling

When you're taking an ordinary derivative, you have the rule

$$\frac{\mathrm{d}}{\mathrm{d}x}[cf(x)]=c\frac{\mathrm{d}}{\mathrm{d}x}f(x) \qquad .$$

In other words, multiplying a function by a constant results in a derivative that is multiplied by that constant. The curl is a kind of derivative operator, and the same is true for a curl.



g / Example 14.

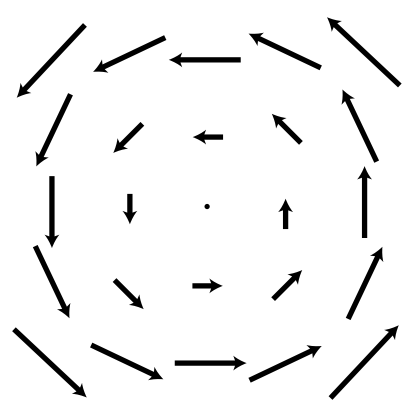
Multiplying the field by -1. *example 14*

▷ What is the curl of the field $-x\hat{\mathbf{y}}$ at the origin?

▷ Using the scaling property just discussed, we can make this into a curl that we've already calculated:

$$\begin{aligned}\operatorname{curl}(-x\hat{\mathbf{y}}) &= -\operatorname{curl}(x\hat{\mathbf{y}})\\ &= -\hat{\mathbf{z}}\end{aligned}$$

This is in agreement with the right-hand rule.



h / Example 15.

The curl is additive.

We have only calculated each field's curl at the origin, but each of these fields actually has the same curl everywhere. In example 14, for instance, it is obvious that the curl is constant along any vertical line. But even if we move along the x axis, there is still an imbalance between the torques on the left and right sides of the curl-meter. More formally, suppose we start from the origin and move to the left by one unit. We find ourselves in a region where the field is very much as it was before, except that all the field vectors have had one unit worth of $\hat{\mathbf{y}}$ added to them. But what do we get if we take the curl of $-x\hat{\mathbf{y}}+\hat{\mathbf{y}}$? The curl, like any god-fearing derivative operation, has the additive property

$$\operatorname{curl}(\mathbf{F}+\mathbf{G}) = \operatorname{curl}\mathbf{F} + \operatorname{curl}\mathbf{G} \qquad ,$$

so

$$\operatorname{curl}(-x\hat{\mathbf{y}}+\hat{\mathbf{y}}) = \operatorname{curl}(-x\hat{\mathbf{y}}) + \operatorname{curl}(\hat{\mathbf{y}}) \qquad .$$

But the second term is zero, so we get the same result as at the origin.

A field that goes in a circle — *example 15*

▷ What is the curl of the field $x\hat{\mathbf{y}} - y\hat{\mathbf{x}}$?

▷ Using the linearity of the curl, and recognizing each of the terms as one whose curl we have already computed, we find that this field's curl is a constant $2\hat{\mathbf{z}}$. This agrees with the right-hand rule.

The field inside a long, straight wire — *example 16*

▷ What is the magnetic field *inside* a long, straight wire in which the current density is j?

▷ Let the wire be along the z axis, so $\mathbf{j} = j\hat{\mathbf{z}}$. Ampère's law gives

$$\operatorname{curl}\mathbf{B} = \frac{4\pi k}{c^2} j\hat{\mathbf{z}} \qquad .$$

In other words, we need a magnetic field whose curl is a constant. We've encountered several fields with constant curls, but the only one that has the same symmetry as the cylindrical wire is $x\hat{\mathbf{y}} - y\hat{\mathbf{x}}$, so the answer must be this field or some constant multiplied by it,

$$\mathbf{B} = b\,(x\hat{\mathbf{y}} - y\hat{\mathbf{x}}) \qquad .$$

The curl of this field is $2b\hat{\mathbf{z}}$, so

$$2b = \frac{4\pi k}{c^2} j \qquad ,$$

and thus

$$\mathbf{B} = \frac{2\pi k}{c^2} j\,(x\hat{\mathbf{y}} - y\hat{\mathbf{x}}) \qquad .$$

The curl in component form

Now consider the field

$$F_x = ax + by + c$$
$$F_y = dx + ey + f \qquad ,$$

i.e.

$$\mathbf{F} = ax\hat{\mathbf{x}} + by\hat{\mathbf{x}} + c\hat{\mathbf{x}} + dx\hat{\mathbf{y}} + ey\hat{\mathbf{y}} + f\hat{\mathbf{y}} \qquad .$$

The only terms whose curls we haven't yet explicitly computed are the a, e, and f terms, and their curls turn out to be zero (homework problem 32). Only the b and d terms have nonvanishing curls. The curl of this field is

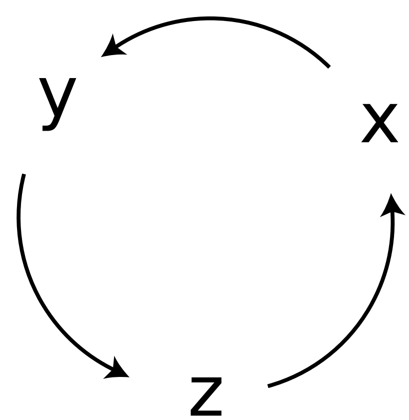
$$\begin{aligned}
\operatorname{curl}\mathbf{F} &= \operatorname{curl}(by\hat{\mathbf{x}}) + \operatorname{curl}(dx\hat{\mathbf{y}}) \\
&= b\operatorname{curl}(y\hat{\mathbf{x}}) + d\operatorname{curl}(x\hat{\mathbf{y}}) \qquad \text{[scaling]} \\
&= b(-\hat{\mathbf{z}}) + d(\hat{\mathbf{z}}) \qquad \text{[found previously]} \\
&= (d-b)\hat{\mathbf{z}} \qquad .
\end{aligned}$$

But *any* field in the $x-y$ plane can be approximated with this type of field, as long as we only need to get a good approximation within a small region. The infinitesimal Ampèrian surface occurring in the definition of the curl is tiny enough to fit in a pretty small region, so we can get away with this here. The d and b coefficients can then be associated with the partial derivatives $\partial F_y/\partial x$ and $\partial F_x/\partial y$. We therefore have

$$\operatorname{curl}\mathbf{F} = \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right)\hat{\mathbf{z}}$$

for any field in the $x-y$ plane. In three dimensions, we just need to generate two more equations like this by doing a cyclic permutation of the variables x, y, and z:

$$\begin{aligned}
(\operatorname{curl}\mathbf{F})_x &= \frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} \\
(\operatorname{curl}\mathbf{F})_y &= \frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x} \\
(\operatorname{curl}\mathbf{F})_z &= \frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}
\end{aligned}$$



i / A cyclic permutation of x, y, and z.

A sine wave — *example 17*

▷ Find the curl of the following electric field

$$\mathbf{E} = (\sin x)\hat{\mathbf{y}} \qquad ,$$

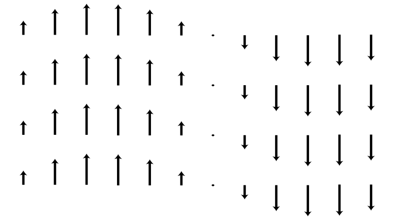
and interpret the result.

▷ The only nonvanishing partial derivative occurring in this curl is

$$\begin{aligned}
(\operatorname{curl}\mathbf{E})_z &= \frac{\partial E_y}{\partial x} \\
&= \cos x \qquad ,
\end{aligned}$$

so

$$\operatorname{curl}\mathbf{E} = \cos\hat{\mathbf{z}}$$

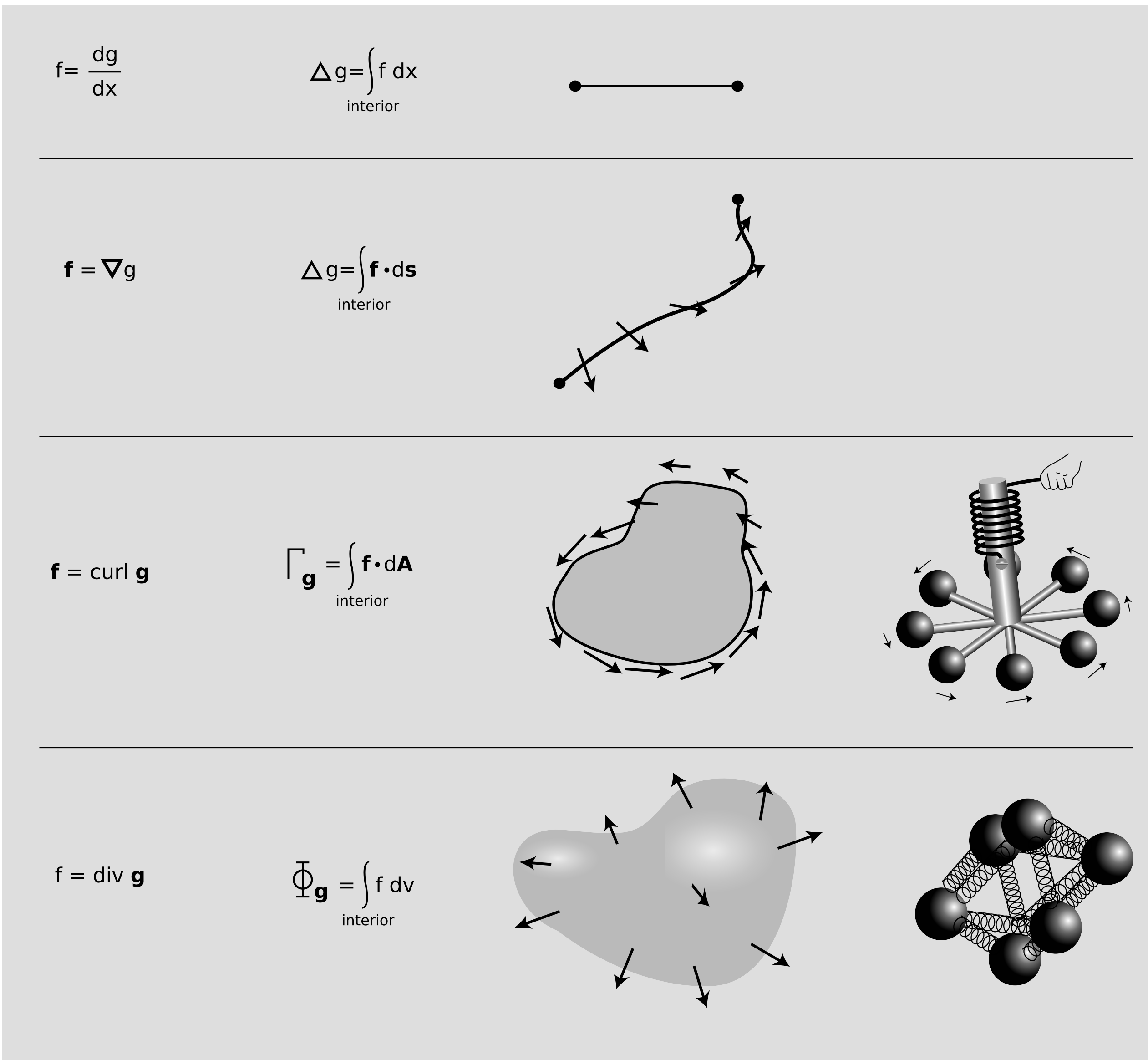


j / Example 17.

This is visually reasonable: the curl-meter would spin if we put its wheel in the plane of the page, with its axle poking out along the z axis. In some areas it would spin clockwise, in others counterclockwise, and this makes sense, because the cosine is positive in some placed and negative in others.

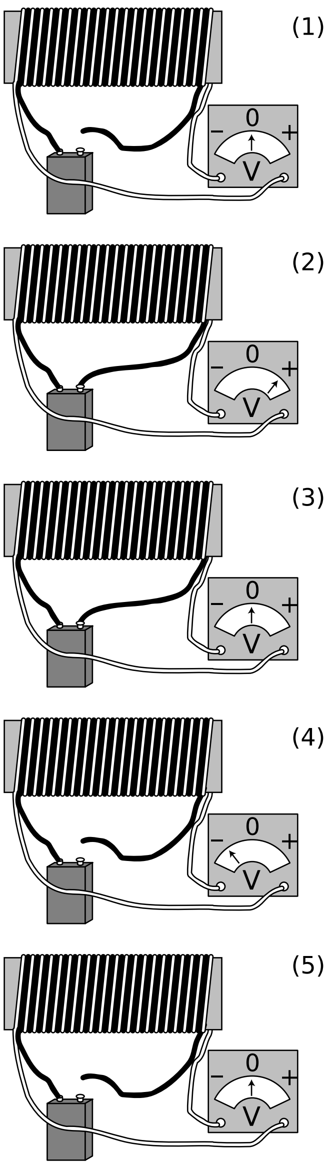
This is a perfectly reasonable field pattern: it the electric field pattern of a light wave! But Ampère's law for electric fields says the curl of **E** is supposed to be zero. What's going on? What's wrong is that we can't assume the *static* version of Ampère's law. All we've really proved is that this pattern is impossible as a static field: we can't have a light wave that stands still.

Figure k is a summary of the vector calculus presented in the optional sections of this book. The first column shows that one function is a related to another by a kind of differentiation. The second column states the fundamental theorem of calculus, which says that if you integrate the derivative over the interior of a region, you get some information about the original function at the boundary of that region.



k / A summary of the derivative, gradient, curl, and divergence.

a / Faraday on a British banknote.



b / Faraday's experiment, simplified and shown with modern equipment.

11.5 Induced Electric Fields

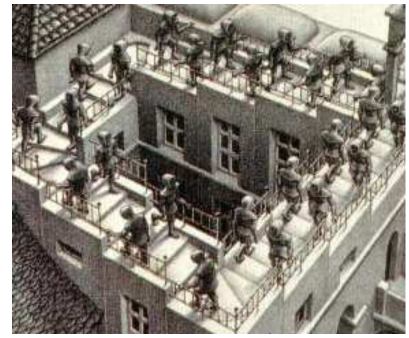
11.5.1 Faraday's experiment

Nature is simple, but the simplicity may not become evident until a hundred years after the discovery of some new piece of physics. We've already seen, on page 489, that the time-varying magnetic field in an inductor causes an electric field. This electric field is *not created by charges.* That argument, however, only seems clear with hindsight. The discovery of this phenomenon of induced electric fields — fields that are not due to charges — was a purely experimental accomplishment by Michael Faraday (1791-1867), the son of a blacksmith who had to struggle against the rigid class structure of 19th century England. Faraday, working in 1831, had only a vague and general idea that electricity and magnetism were related to each other, based on Oersted's demonstration, a decade before, that magnetic fields were caused by electric currents.

Figure b is a simplified drawing of the following experiment, as described in Faraday's original paper: "Two hundred and three feet of copper wire . . . were passed round a large block of wood; [another] two hundred and three feet of similar wire were interposed as a spiral between the turns of the first, and metallic contact everywhere prevented by twine [insulation]. One of these [coils] was connected with a galvanometer [voltmeter], and the other with a battery. . . When the contact was made, there was a sudden and very slight effect at the galvanometer, and there was also a similar slight effect when the contact with the battery was broken. But whilst the . . . current was continuing to pass through the one [coil], no . . . effect . . . upon the other [coil] could be perceived, although the active power of the battery was proved to be great, by its heating the whole of its own coil [through ordinary resistive heating] . . . "

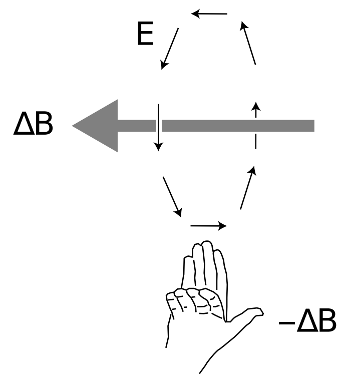
From Faraday's notes and publications, it appears that the situation in figure b/3 was a surprise to him, and he probably thought it would be a surprise to his readers, as well. That's why he offered evidence that the current was still flowing: to show that the battery hadn't just died. The induction effect occurred during the short time it took for the black coil's magnetic field to be established, b/2. Even more counterintuitively, we get an effect, equally strong but in the opposite direction, when the circuit is *broken*, b/4. The effect occurs only when the magnetic field is changing, and it appears to be proportional to the derivative $\partial\mathbf{B}/\partial\mathbf{t}$, which is in one direction when the field is being established, and in the opposite direction when it collapses.

The effect is proportional to $\partial\mathbf{B}/\partial\mathbf{t}$, but what *is* the effect? A voltmeter is nothing more than a resistor with an attachment for measuring the current through it. A current will not flow through a resistor unless there is some electric field pushing the electrons, so we conclude that the changing *magnetic field* has produced an *electric field* in the surrounding space. Since the white wire is not a perfect conductor, there must be electric fields in it as well. The remarkable thing about the circuit formed by the white wire is that as the electrons travel around and around, they are always being pushed forward by electric fields. This violates the loop rule, which says that when an electron makes a round trip, there is supposed to be just as much "uphill" (moving against the electric field) as "downhill" (moving with it). That's OK. The loop rule is only true for statics. Faraday's experiments show that an electron really can go around and around, and always be going "downhill," as in the famous drawing by M.C. Escher shown in figure c. That's just what happens when you have a curly field.



c / Detail from *Ascending and Descending*, M.C. Escher, 1960.

When a field is curly, we can measure its curliness using a circulation. Unlike the magnetic circulation Γ_B, the electric circulation Γ_E is something we can measure directly using ordinary tools. A circulation is defined by breaking up a loop into tiny segments, $d\mathbf{s}$, and adding up the dot products of these distance vectors with the field. But when we multiply electric field by distance, what we get is an indication of the amount of work per unit charge done on a test charge that has been moved through that distance. The work per unit charge has units of volts, and it can be measured using a voltmeter, as shown in figure e, where Γ_E equals the sum of the voltmeter readings. Since the electric circulation is directly measurable, most people who work with circuits are more familiar with it than they are with the magnetic circulation. They usually refer to Γ_E using the synonym "emf," which stands for "electromotive force," and notate it as $\mathcal{E}$. (This is an unfortunate piece of terminology, because its units are really volts, not newtons.) The term emf can also be used when the path is not a closed loop.



d / The relationship between the change in the magnetic field, and the electric field it produces.

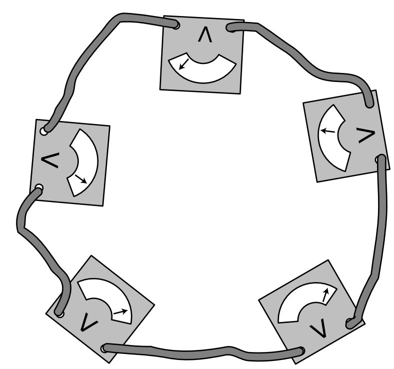
Faraday's experiment demonstrates a new relationship

$$\Gamma_E \propto -\frac{\partial B}{\partial t} \qquad ,$$

where the negative sign is a way of showing the observed left-handed relationship, d. This is similar to the structure of of Ampère's law:

$$\Gamma_B \propto I_{through} \qquad ,$$

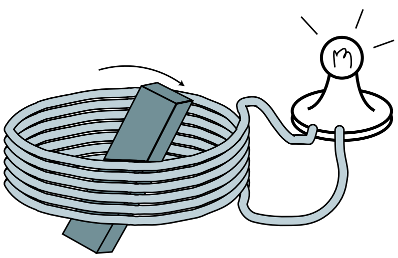
which also relates the curliness of a field to something that is going on nearby (a current, in this case).



e / The electric circulation is the sum of the voltmeter readings.

It's important to note that even though the emf, Γ_E, has units of volts, it isn't a voltage. A voltage is a measure of the electrical energy a charge has when it is at a certain point in space. The curly nature of nonstatic fields means that this whole concept becomes nonsense. In a curly field, suppose one electron stays at home while its friend goes for a drive around the block. When they are reunited, the one that went around the block has picked up some kinetic energy, while the one who stayed at home hasn't. We simply can't define an electrical energy $U_e = qV$ so that $U_e + K$ stays the same for each electron. No voltage pattern, V, can do this, because then it would predict the same kinetic energies for the two electrons, which is incorrect. When we're dealing with nonstatic fields, we need to think of the electrical energy in terms of the energy density of the fields themselves.

It might sound as though an electron could get a free lunch by circling around and around in a curly electric field, resulting in a violation of conservation of energy. The following examples, in addition to their practical interest, both show that energy is in fact conserved.



f / A generator.

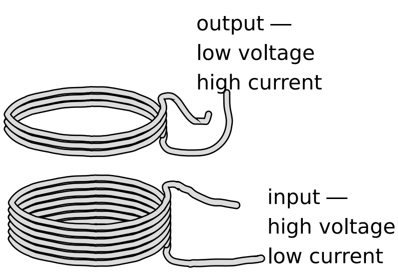
The generator *example 18*

A basic generator, f, consists of a permanent magnet that rotates within a coil of wire. The magnet is turned by a motor or crank, (not shown). As it spins, the nearby magnetic field changes. This changing magnetic field results in an electric field, which has a curly pattern. This electric field pattern creates a current that whips around the coils of wire, and we can tap this current to light the lightbulb.

If the magnet was on a frictionless bearing, could we light the bulb for free indefinitely, thus violating conservation of energy? No. Mechanical work has to be done to crank the magnet, and that's where the energy comes from. If we break the light-bulb circuit, it suddenly gets easier to crank the magnet! This is because the current in the coil sets up its own magnetic field, and that field exerts a torque on the magnet. If we stopped cranking, this torque would quickly make the magnet stop turning.

Self-Check

When you're driving your car, the engine recharges the battery continuously using a device called an alternator, which is really just a generator. Why can't you use the alternator to start the engine if your car's battery is dead? ▷ Answer, p. 710



g / A transformer.

The transformer *example 19*

In example 17 on page 438, we discussed the advantages of transmitting power over electrical lines using high voltages and low currents. However, we don't want our wall sockets to operate at 10000 volts! For this reason, the electric company uses a device called a transformer, g, to convert everything to lower voltages and higher currents inside your house. The coil on the input side creates a magnetic field. Transformers work with alternating current, so the magnetic field surrounding the input coil is always changing. This induces an electric field, which drives a current around the output coil.

Since the electric field is curly, an electron can keep gaining more and more energy by circling through it again and again. Thus the output voltage can be controlled by changing the number of coils of wire on the output side. Changing the number of coils on the input side also has an effect (homework problem 33).

In any case, conservation of energy guarantees that the amount of power on the output side must equal the amount put in originally, $I_{in}V_{in} = I_{out}V_{out}$, so no matter what factor the voltage is reduced by, the current is increased by the same factor.

Discussion Questions

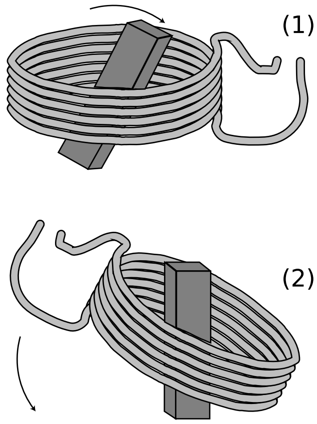
A Suppose the bar magnet in figure f on page 578 has a magnetic field pattern that emerges from its top, circling around and coming back in the bottom. This field is created by electrons orbiting atoms inside the magnet. Are these atomic currents clockwise or counterclockwise as seen from above? In what direction is the current flowing in the circuit?

We have a circling atomic current inside the circling current in the wires. When we have two circling currents like this, they will make torques on each other that will tend to align them in a certain way. Since currents in the same direction attract one another, which way is the torque made by the wires on the bar magnet? Verify that due to this torque, mechanical work has to be done in order to crank the generator.

11.5.2 Why induction?

Faraday's results leave us in the dark about several things:

- They don't explain *why* induction effects occur.
- The relationship $\Gamma_E \propto -\partial B/\partial t$ tells us that a changing magnetic field creates an electric field in the surrounding region of space, but the phrase "surrounding region of space" is vague, and needs to be made mathematical.
- Suppose that we can make the "surrounding region of space" idea more well defined. We would then want to know the proportionality constant that has been hidden by the $\propto$ symbol. Although experiments like Faraday's could be used to find a numerical value for this constant, we would like to know why it should have that particular value.



h / It doesn't matter whether it's the coil or the permanent magnet that spins. Either way, we get a functioning generator.

We can get some guidance from the example of a car's alternator (which just means generator), referred to in the self-check on page 578. To keep things conceptually simple, I carefully avoided mentioning that in a real car's alternator, it isn't actually the permanent magnet that spins. The coil is what spins. The choice of design h/1 or h/2 is merely a matter of engineering convenience, not physics. All that matters is the relative motion of the two objects.

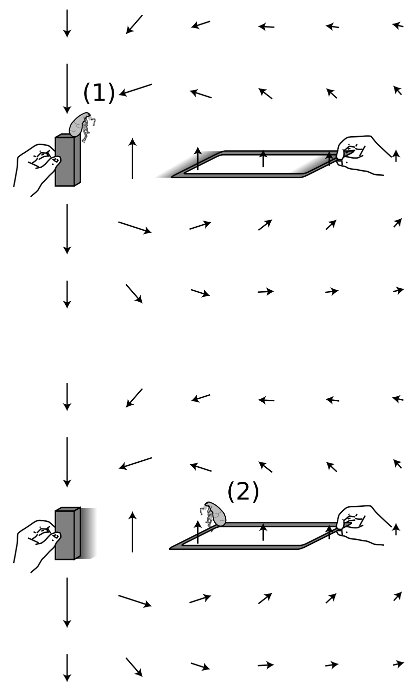
This is highly suggestive. As discussed at the beginning of this chapter, magnetism is a relativistic effect. From arguments about relative motion, we concluded that moving electric charges create magnetic fields. Now perhaps we can use reasoning with the same flavor to show that changing magnetic fields produce curly electric fields. Note that figure h/2 doesn't even require induction. The protons and electrons in the coil are moving through a magnetic field, so they experience forces. The protons can't flow, because the coil is a solid substance, but the electrons can, so a current is induced.[7]

Now if we're convinced that figure h/2 produces a current in the coil, then it seems very plausible that the same will happen in figure h/1, which implies the existence of induction effects. But this example involves circular motion, so it doesn't quite work as a way of proving that induction exists. When we say that motion is relative, we only mean straight-line motion, not circular motion.

A more ironclad relativistic argument comes from the arrangement shown in figure i. This is also a generator — one that is impractical, but much easier to understand.

[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.

Flea 1 doesn't believe in this modern foolishness about induction. She's sitting on the bar magnet, which to her is obviously at rest. As the square wire loop is dragged away from her and the magnet, its protons experience a force out of the page, because the cross product $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$ is out of the page. The electrons, which are negatively charged, feel a force into the page. The conduction electrons are free to move, but the protons aren't. In the front and back sides of the loop, this force is perpendicular to the wire. In the right and left sides, however, the electrons are free to respond to the force. Note that the magnetic field is weaker on the right side. It's as though we had two pumps in a loop of pipe, with the weaker pump trying to push in the opposite direction; the weaker pump loses the argument.[8] We get a current that circulates around the loop.[9] There is no induction going on in this frame of reference; the forces that cause the current are just the ordinary magnetic forces experienced by any charged particle moving through a magnetic field.



i / A generator that works with linear motion.

Flea 2 is sitting on the loop, which she considers to be at rest. In her frame of reference, it's the bar magnet that is moving. Like flea 1, she observes a current circulating around the loop, but unlike flea 1, she cannot use magnetic forces to explain this current. As far as she is concerned, the electrons were initially at rest. Magnetic forces are forces between moving charges and other moving charges, so a magnetic field can never accelerate a charged particle starting from rest. A force that accelerates a charge from rest can only be an *electric* force, so she is forced to conclude that there is an electric field in her region of space. This field drives electrons around and around in circles, so it is apparently violating the loop rule — it is a curly field. What reason can flea 2 offer for the existence of this electric field pattern? Well, she's been noticing that the magnetic field in her region of space has been changing, possibly because that bar magnet over there has been getting further away. She observes that a changing magnetic field creates a curly electric field.

We therefore conclude that induction effects *must* exist based on the fact that motion is relative. If we didn't want to admit induction effects, we would have to outlaw flea 2's frame of reference, but the whole idea of relative motion is that all frames of reference are created equal, and there is no way to determine which one is really

[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.

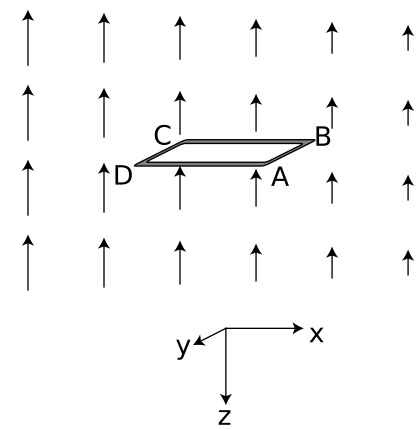
at rest.

This whole line of reasoning was not available to Faraday and his contemporaries, since they thought the relative nature of motion only applied to matter, not to electric and magnetic fields.[10] But with the advantage of modern hindsight, we can understand in fundamental terms the facts that Faraday had to take simply as mysterious experimental observations. For example, the geometric relationship shown in figure d follows directly from the direction of the current we deduced in the story of the two fleas.

[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.

11.5.3 Faraday's law

We can also answer the other questions posed on page 580. The divide-and-conquer approach should be familiar by now. We first determine the circulation Γ_E in the case where the wire loop is very tiny, j. Then we can break down any big loop into a grid of small ones; we've already seen that when we make this kind of grid, the circulations add together. Although we'll continue to talk about a physical loop of wire, as in figure i, the tiny loop can really be just like the edges of an Ampèrian surface: a mathematical construct that doesn't necessarily correspond to a real object.



j / A new version of figure i with a tiny loop. The point of view is above the plane of the loop. In the frame of reference where the magnetic field is constant, the loop is moving to the right.

In the close-up view shown in figure j, the field looks simpler. Just as a tiny part of a curve looks straight, a tiny part of this magnetic field looks like the field vectors are just getting shorter by the same amount with each step to the right. Writing $\mathrm{d}x$ for the width of the loop, we therefore have

$$B(x+\mathrm{d}x) - B(x) = \frac{\partial B}{\partial x}\,\mathrm{d}x$$

for the difference in the strength of the field between the left and right sides. In the frame of reference where the loop is moving, a charge q moving along with the loop at velocity v will experience a magnetic force $\mathbf{F}_B = qvB\hat{\mathbf{y}}$. In the frame moving along with the loop, this is interpreted as an electrical force, $\mathbf{F}_E = qE\hat{\mathbf{y}}$. Observers in the two frames agree on how much force there is, so in the loop's frame, we have an electric field $\mathbf{E} = vB\hat{\mathbf{y}}$. This field is perpendicular to the front and back sides of the loop, BC and DA, so there is no contribution to the circulation along these sides, but there is a counterclockwise contribution to the circulation on CD, and smaller clockwise one on AB. The result is a circulation that is counterclockwise, and has an absolute value

$$\begin{aligned}
|\Gamma_E| &= |E(x)\mathrm{d}y - E(x+\mathrm{d}x)\mathrm{d}y| \\
&= |v[B(x) - B(x+\mathrm{d}x)]|\mathrm{d}y \\
&= |v\,\frac{\partial B}{\partial x}|\,\mathrm{d}x\mathrm{d}y \\
&= |\frac{\mathrm{d}x}{\mathrm{d}t}\,\frac{\partial B}{\partial x}|\,\mathrm{d}x\mathrm{d}y \\
&= |\frac{\partial B}{\partial t}|\,\mathrm{d}A \qquad .
\end{aligned}$$

Using a right-hand rule, the counterclockwise circulation is represented by pointing one's thumb up, but the vector $\partial\mathbf{B}/\partial t$ is down. This is just a rephrasing of the geometric relationship shown in figure d on page 577. We can represent the opposing directions using a minus sign,

$$\Gamma_E = -\frac{\partial B}{\partial t}\,\mathrm{d}A \qquad .$$

Although this derivation was carried out with everything aligned in a specific way along the coordinate axes, it turns out that this relationship can be generalized as a vector dot product,

$$\Gamma_E = -\frac{\partial \mathbf{B}}{\partial t} \cdot \mathrm{d}\mathbf{A} \qquad .$$

Finally, we can take a finite-sized loop and break down the circulation around its edges into a grid of tiny loops. The circulations add, so we have

$$\Gamma_E = -\sum \frac{\partial \mathbf{B}_j}{\partial t} \cdot \mathrm{d}\mathbf{A}_j \qquad .$$

This is known as Faraday's law. (I don't recommend memorizing all these names.) Mathematically, Faraday's law is very similar to the structure of Ampère's law: the circulation of a field around the edges of a surface is equal to the sum of something that points through the

If the loop itself isn't moving, twisting, or changing shape, then the area vectors don't change over time, and we can move the derivative outside the sum, and rewrite Faraday's law in a slightly more transparent form:

$$\begin{aligned}\Gamma_E &= -\frac{\partial}{\partial t} \sum \mathbf{B}_j \cdot \mathrm{d}\mathbf{A}_j \\ &= -\frac{\partial \Phi_B}{\partial t}\end{aligned}$$

A changing magnetic flux makes a curly electric field. You might think based on Gauss' law for magnetic fields that Φ_B would be identically zero. However, Gauss' law only applies to surfaces that are closed, i.e. have no edges.

Self-Check

Check that the units in Faraday's law work out. An easy way to approach this is to use the fact that vB has the same units as E, which can be seen by comparing the equations for magnetic and electric forces used above. ▷ Answer, p. 710

A pathetic generator *example 20*

▷ The horizontal component of the earth's magnetic field varies from zero, at a magnetic pole, to about 10^{-4} T near the equator. Since the distance from the equator to a pole is about 10^7 m, we can estimate, very roughly, that the horizontal component of the earth's magnetic field typically varies by about 10^{-11} T/m as you go north or south. Suppose you connect the terminals of a one-ohm lightbulb to each other with a loop of wire having an area of 1 m^2. Holding the loop so that it lies in the east-west-up-down plane, you run straight north at a speed of 10 m/s, how much current will flow? Next, repeat the same calculation for the surface of a neutron star. The magnetic field on a neutron star is typically 10^9 T, and the radius of an average neutron star is about 10^4 m.

▹ Let's work in the frame of reference of the running person. In this frame of reference, the earth is moving, and therefore the local magnetic field is changing in strength by 10^{-9} T/s. This rate of change is almost exactly the same throughout the interior of the loop, so we can dispense with the summation, and simply write Faraday's law as

$$\Gamma_E = -\frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{A} \qquad .$$

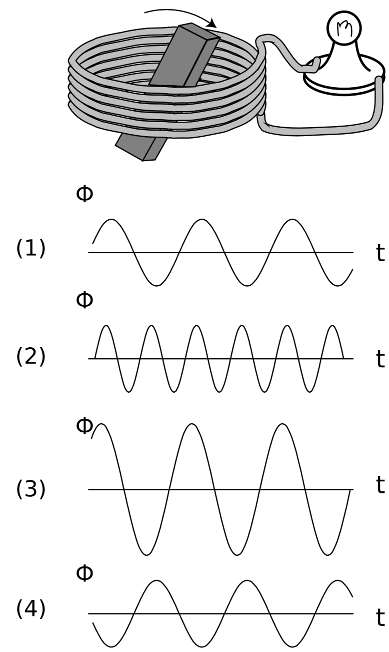
Since what we estimated was the rate of change of the horizontal component, and the vector $\mathbf{A}$ is horizontal (perpendicular to the loop), we can find this dot product simply by multiplying the two numbers:

$$\begin{aligned}\Gamma_E &= (10^{-9}\ \mathrm{T/s})(1\ \mathrm{m}^2)\\ &= 10^{-9}\ \mathrm{T{\cdot}m^2/s}\\ &= 10^{-9}\ \mathrm{V}\end{aligned}$$

This is certainly not enough to light the bulb, and would not even be easy to measure using the most sensitive laboratory instruments.

Now what about the neutron star? We'll pretend you're tough enough that its gravity doesn't instantly crush you. The spatial variation of the magnetic field is on the order of $(10^9\ \mathrm{T}/10^4\ \mathrm{m}) = 10^5\ \mathrm{T/m}$. If you can run north at the same speed of 10 m/s, then in your frame of reference there is a temporal (time) variation of about 10^6 T/s, and a calculation similar to the previous one results in an emf of 10^6 V! This isn't just strong enough to light the bulb, it's sufficient to evaporate it, and kill you as well!

It might seem as though having access to a region of rapidly changing magnetic field would therefore give us an infinite supply of free energy. However, the energy that lights the bulb is actually coming from the mechanical work you do by running through the field. A tremendous force would be required to make the wire loop move through the neutron star's field at any significant speed.



k / Example 21.

Speed and power in a generator *example 21*

▹ Figure k shows three graphs of the magnetic flux through a generator's coils as a function of time. In graph 2, the generator is being cranked at twice the frequency. In 3, a permanent magnet with double the strength has been used. In 4, the generator is being cranked in the opposite direction. Compare the power generated in figures 2-4 with the the original case, 1.

▹ If the flux varies as $\Phi = A \sin \omega t$, then the time derivative occurring in Faraday's law is $\partial\Phi/\partial t = A\omega \cos \omega t$. The absolute value of this is the same as the absolute value of the emf, Γ_E. The current through the lightbulb is proportional to this emf, and the power dissipated depends on the square of the current ($P = I^2R$), so $P \propto A^2\omega^2$. Figures 2 and 3 both give four times the output power (and require four times the input power). Figure 4 gives the same result as figure 1; we can think of this as a negative amplitude, which gives the same result when squared.

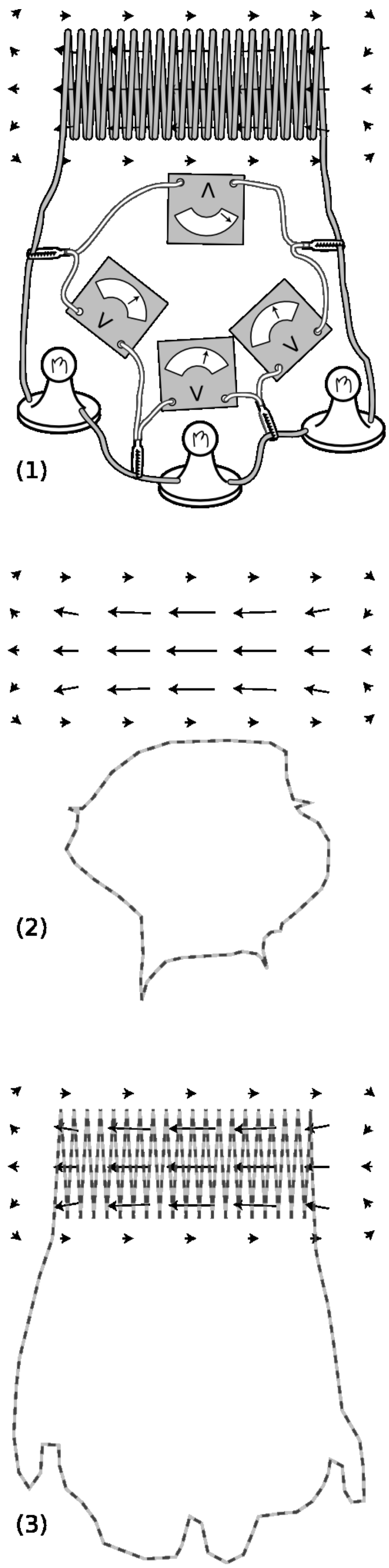
An approximate loop rule *example 22*

Figure l/1 shows a simple RL circuit of the type discussed in the last chapter. A current has already been established in the coil, let's say by a battery. The battery was then unclipped from the coil, and we now see the circuit as the magnetic field in and around the inductor is beginning to collapse. I've already cautioned you that the loop rule doesn't apply in nonstatic situations, so we can't assume that the readings on the four voltmeters add up to zero. The interesting thing is that although they don't add up to exactly zero in this circuit, they very nearly do. Why is the loop rule even approximately valid in this situation?

The reason is that the voltmeters are measuring the emf Γ_E around the path shown in figure l/2, and the stray field of the solenoid is extremely weak out there. In the region where the meters are, the arrows representing the magnetic field would be too small to allow me to draw them to scale, so I have simply omitted them. Since the field is so weak in this region, the flux through the loop is nearly zero, and the rate of change of the flux, $\partial\Phi_B/\partial t$, is also nearly zero. By Faraday's law, then, the emf around this loop is nearly zero.

Now consider figure l/3. The flux through the interior of this path is not zero, because the strong part of the field passes through it, and not just once but many times. To visualize this, imagine that we make a wire frame in this shape, dip it in a tank of soapy water, and pull it out, so that there is a soap-bubble film spanning its interior. Faraday's law refers to the rate of change of the flux through a surface such as this one. (The soap film tends to assume a certain special shape which results in the minimum possible surface area, but Faraday's law would be true for any surface that filled in the loop.) In the coiled part of the wire, the soap makes a three-dimensional screw shape, like the shape you would get if you took the steps of a spiral staircase and smoothed them into a ramp. The loop rule is going to be strongly violated for this path.

We can interpret this as follows. Since the wire in the solenoid has a very low resistance compared to the resistances of the light bulbs, we can expect that the electric field along the corkscrew part of loop l/3 will be very small. As an electron passes through the coil, the work done on it is therefore essentially zero, and the true emf along the coil is zero. In figure l/1, the meter on top is therefore not telling us the actual emf experienced by an electron that passes through the coil. It is telling us the emf experienced by an electron that passes through the meter itself, which is a different quantity entirely. The other three meters, however, really do tell us the emf through the bulbs, since there are no magnetic fields where they are, and therefore no funny induction effects.



l / Example 22.

11.6 Maxwell's Equations

11.6.1 Induced magnetic fields

We have almost, but not quite, done figuring out the complete set of physical laws, called Maxwell's equations, governing electricity and magnetism. We are only missing one more term. For clarity, I'll state Maxwell's equations with the missing part included, and then discuss the physical motivation and experimental evidence for sticking it in:

a / James Clerk Maxwell (1831-1879)

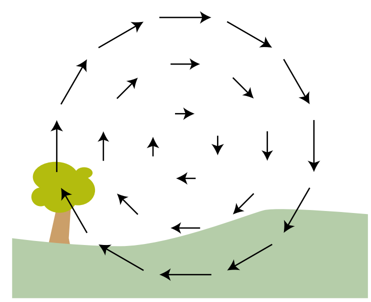
Maxwell's equations

For any closed surface, the fluxes through the surface are

$$\Phi_E = 4\pi k q_{in} \qquad \text{and}$$
$$\Phi_B = 0 \qquad .$$

For any surface that is not closed, the circulations around the edges of the surface are given by

$$\Gamma_E = -\frac{\partial \Phi_B}{\partial t} \qquad \text{and}$$
$$c^2\Gamma_B = \frac{\partial \Phi_E}{\partial t} + 4\pi k I_{through} \qquad .$$

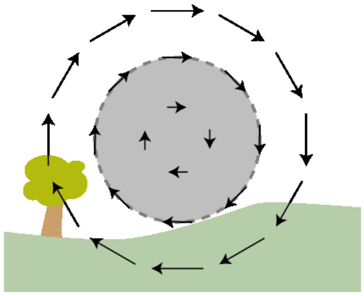


b / Where is the moving charge responsible for this magnetic field?

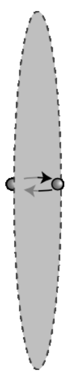
The Φ_E equation is Gauss' law: charges make diverging electric fields. The corresponding equation for Φ_B tells us that magnetic "charges" (monopoles) don't exist, so magnetic fields never diverge. The third equation says that changing magnetic fields induce curly electric fields, whose curliness we can measure using the emf, Γ_E, around a closed loop. The final equation, for Γ_B, is the only one where anything new has been added. Without the new time derivative term, this equation would simply be Ampère's law. (I've chosen to move the c^2 over to the left because it simplifies the writing, and also because it more clearly demonstrates the analogous roles played by charges and currents in the Φ_E and Γ_B equations.)

This new $\partial\Phi_E/\partial t$ term says that just as a changing magnetic field can induce a curly electric field, a changing electric field can induce a curly magnetic field. Why should this be so? The following examples show that Maxwell's equations would not make sense in general without it.

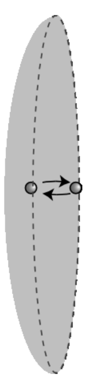
Figure b shows a mysterious curly magnetic field. Magnetic fields are supposed to be made by moving charges, but there don't seem to be any moving charges in this landscape. Where are they? One reasonable guess would be that they're behind your head, where you can't see them. Suppose there's a positively charged particle about to hit you in the back of the head. This particle is like a current going into the page. We're used to dealing with currents made by



c / An Ampèrian surface superimposed on the landscape.



d / An electron jumps through a hoop.



e / An alternative Ampèrian surface.

many charged particles, but logically we can't have some minimum number that would qualify as a current. This is not a static current, however, because the current at a given point in space is not staying the same over time. If the particle is pointlike, then it takes zero time to pass any particular location, and the current is then infinite at that point in space. A moment later, when the particle is passing by some other location, there will be an infinite current there, and zero current in the previous location. If this single particle qualifies as a current, then it should be surrounded by a curly magnetic field, just like any other current.[11]

This explanation is simple and reasonable, but how do we know it's correct? Well, it makes another prediction, which is that the positively charged particle should be making an electric field as well. Not only that, but if it's headed for the back of your head, then it's getting closer and closer, so the electric field should be getting stronger over time. But this is exactly what Maxwell's equations require. There is no current $I_{through}$ piercing the Ampèrian surface shown in figure c, so Maxwell's equation for Γ_B becomes $c^2\Gamma_B = \partial\Phi_E/\partial t$. The only reason for an electric field to change is if there are charged particles making it, and those charged particles are moving. When charged particles are moving, they make magnetic fields as well.

Note that the above example is also sufficient to prove the positive sign of the $\partial\Phi_E/\partial t$ term in Maxwell's equations, which is different from the negative sign of Faraday's $-\partial\Phi_B/\partial t$ term.

The addition of the $\partial\Phi_E/\partial t$ term has an even deeper and more important physical meaning. With the inclusion of this term, Maxwell's equations can describe correctly the way in which disturbances in the electric and magnetic fields ripple outwards at the speed of light. Indeed, Maxwell was the first human to understand that light was in fact an electromagnetic wave. Legend has it that it was on a starry night that he first realized this implication of his equations. He went for a walk with his wife, and told her she was the only other person in the world who really knew what starlight was.

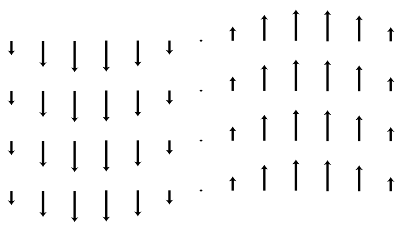
To see how the $\partial\Phi_E/\partial t$ term relates to electromagnetic waves, let's look at an example where we would get nonsense without it. Figure d shows an electron that sits just on one side of an imaginary Ampèrian surface, and then hops through it at some randomly chosen moment. Unadorned with the $\partial\Phi_E/\partial t$ term, Maxwell's equation for Γ_B reads as $c^2\Gamma_B = 4\pi k I_{through}$, which is Ampère's law. If

[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.

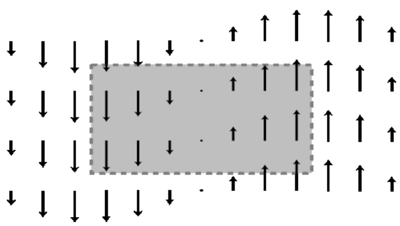
the electron is a pointlike particle, then we have an infinite current $I_{through}$ at the moment when it pierces the imaginary surface, and zero current at all other times. An infinite magnetic circulation Γ_B can only be produced by an infinite magnetic field, so without the $\partial\Phi_E/\partial t$ term, Maxwell's equations predict nonsense: the edge of the surface would experience an infinite magnetic field at one instant, and zero magnetic field at all other times. Even if the infinity didn't upset us, it doesn't make sense that anything special would happen at the moment the electron passed through the surface, because the surface is an imaginary mathematical construct. We could just as well have chosen the curved surface shown in figure e, which the electron never crosses at all. We are already clearly getting nonsensical results by omitting the $\partial\Phi_E/\partial t$ term, and this shouldn't surprise us because Ampère's law only applies to statics. More to the point, Ampère's law doesn't have time in it, so it predicts that this effect is instantaneous. According to Ampère's law, we could send Morse code signals by wiggling the electron back and forth, and these signals would be received at distant locations instantly, without any time delay at all. This contradicts the theory of relativity, one of whose predictions is that information cannot be transmitted at speeds greater than the speed of light.

Discussion Questions

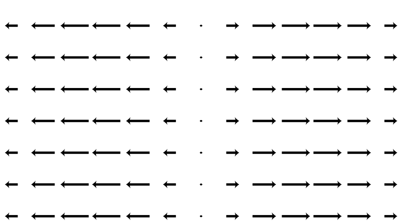
A Induced magnetic fields were introduced in the text via the imaginary landscape shown in figure b on page 587, and I argued that the magnetic field could have been produced by a positive charge coming from behind your head. This is a specific assumption about the *number* of charges (one), the *direction* of motion, and the *sign* of the charge. What are some other scenarios that could explain this field?



f / A magnetic field in the form of a sine wave.



g / The wave pattern is curly. For example, the circulation around this reactangle is nonzero and counterclockwise.



i / An impossible wave pattern.

11.6.2 Light waves

We could indeed send signals using this scheme, and the signals would *be* a form of light. A radio transmitting antenna, for instance, is simply a device for whipping electrons back and forth at megahertz frequencies. Radio waves are just like visible light, but with a lower frequency. With the addition of the $\partial\Phi_E/\partial t$ term, Maxwell's equations are capable of describing electromagnetic waves. It would be possible to use Maxwell's equations to calculate the pattern of the electric and magnetic fields rippling outward from a single electron that fidgets at irregular intervals, but let's pick a simpler example to analyze.

The simplest wave pattern is a sine wave like the one shown in figure f. Let's assume a magnetic field of this form, and see what Maxwell's equations tell us about it. If the wave is traveling through empty space, then there are no charges or currents present, and Maxwell's equations become

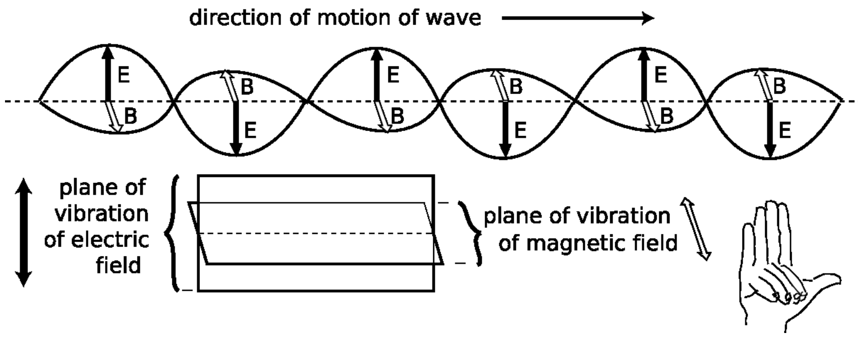
$$\begin{aligned}\Phi_E &= 0\\ \Phi_B &= 0\\ \Gamma_E &= -\frac{\partial\Phi_B}{\partial t}\\ c^2\Gamma_B &= \frac{\partial\Phi_E}{\partial t} \qquad .\end{aligned}$$

The equation $\Phi = 0$ has already been verified for this type of wave pattern in example 34 on page 521. Even if you haven't learned the techniques from that section, it should be visually plausible that this field pattern doesn't diverge or converge on any particular point.

Geometry of the electric and magnetic fields

The equation $c^2\Gamma_B = \partial\Phi_E/\partial t$ tells us that there can be no such thing as a purely magnetic wave. The wave pattern clearly does have a nonvanishing circulation around the edge of the surface suggested in figure g, so there must be an electric flux through the surface. This magnetic field pattern must be intertwined with an electric field pattern that fills the same space. There is also no way that the two sides of the equation could stay synchronized with each other unless the electric field pattern is also a sine wave, and one that has the same wavelength, frequency, and velocity. Since the electric field is making a flux through the indicated surface, it's plausible that the electric field vectors lie in a plane perpendicular to that of the magnetic field vectors. The resulting geometry is shown in figure h. Further justification for this geometry is given later in this subsection.

One feature of figure h that is easily justified is that the electric and magnetic fields are perpendicular not only to each other, but



h / The geometry of an electromagnetic wave.

also to the direction of propagation of the wave. In other words, the vibration is sideways, like people in a stadium "doing the wave," not lengthwise, like the accordion pattern in figure i. (In standard wave terminology, we say that the wave is transverse, not longitudinal.) The wave pattern in figure i is impossible, because it diverges from the middle. For virtually any choice of Gaussian surface, the magnetic and electric fluxes would be nonzero, contradicting the equations $\Phi_B = 0$ and $\Phi_E = 0$.[12]

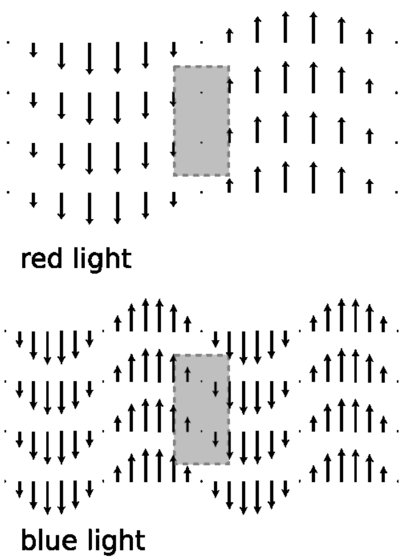
Polarization

Two electromagnetic waves traveling in the same direction through space can differ by having their electric and magnetic fields in different directions, a property of the wave called its polarization.

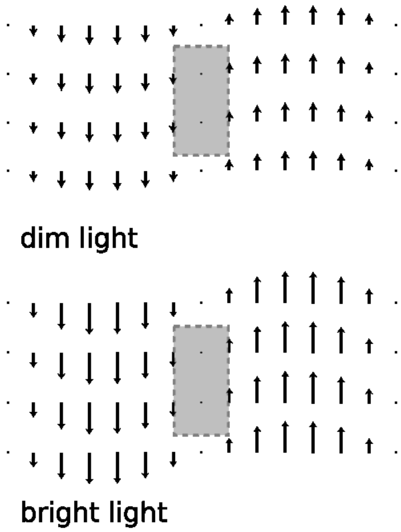
The speed of light

What is the velocity of the waves described by Maxwell's equations? Maxwell convinced himself that light was an electromagnetic wave partly because his equations predicted waves moving at the velocity of light, c. The only velocity that appears in the equations is c, so this is fairly plausible, although a real calculation is required in order to prove that the velocity of the waves isn't something like $2c$ or c/π — or zero, which is also c multiplied by a constant! The following discussion, leading up to a proof that electromagnetic waves travel at c, is meant to be understandable even if you're reading this book out of order, and haven't yet learned much about waves. As always with proofs in this book, the reason to read it isn't to convince yourself that it's true, but rather to build your intuition. The style will be visual. In all the following figures, the wave patterns are moving across the page (let's say to the right), and it usually doesn't matter whether you imagine them as representing the wave's magnetic field or its electric field, because Maxwell's equations in

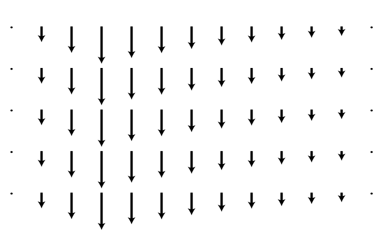
[12]Even if the fields can't be parallel to the direction of propagation, one might wonder whether they could form some angle other than 90 degrees with it. No. One proof is given on page 595. A alternative argument, which is simpler but more esoteric, is that if there was such a pattern, then there would be some other frame of reference in which it would look like figure i.



j / Red and blue light travel at the same speed.



k / Bright and dim light travel at the same speed.



l / A nonsinusoidal wave.

a vacuum have the same form for both fields. Whichever field we imagine the figures as representing, the other field is coming in and out of the page.

The velocity of the waves is not zero. If the wave pattern was standing still in space, then the right sides of the Γ equations would be zero, because there would be no change in the field over time at a particular point. But the left sides are not zero, so this is impossible.[13]

The velocity of the waves is a fixed number for a given wave pattern. Consider a typical sinusoidal wave of visible light, with a distance of half a micrometer from one peak to the next peak. Suppose this wave pattern provides a valid solution to Maxwell's equations when it is moving with a certain velocity. We then know, for instance, that there *cannot* be a valid solution to Maxwell's equations in which the same wave pattern moves with double that velocity. The time derivatives on the right sides of Maxwell's equations for Γ_E and Γ_B would be twice as big, since an observer at a certain point in space would see the wave pattern sweeping past at twice the rate. But the left sides would be the same, so the equations wouldn't equate.

The velocity is the same for all wave patterns. In other words, it isn't $0.878c$ for one wave pattern, and $1.067c$ for some other pattern. This is surprising, since, for example, water waves with different shapes do travel at different speeds. Similarly, even though we speak of "the speed of sound," sound waves do travel at slightly different speeds depending on their pitch and loudness, although the differences are small unless you're talking about cannon blasts or extremely high frequency ultrasound. To see how Maxwell's equations give a consistent velocity, consider figure j. Along the right and left edges of the same Ampèrian surface, the more compressed wave pattern of blue light has twice as strong a field, so the circulations on the left sides of Maxwell's equations are twice as large.[14] To satisfy Maxwell's equations, the time derivatives of the fields must also be twice as large for the blue light. But this is true only if the blue light's wave pattern is moving to the right at the *same* speed as the red light's: if the blue light pattern is sweeping over an observer with a given velocity, then the time between peaks is half as much, like the clicking of the wheels on a train whose cars are half the length.[15]

---

[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

We can also check that bright and dim light, as shown in figure k, have the same velocity. If you haven't yet learned much about waves, then this might be surprising. A material object with more energy goes faster, but that's not the case for waves. The circulation around the edge of the Ampèrian surface shown in the figure is twice as strong for the light whose fields are doubled in strength, so the left sides of Maxwell's Γ equations are doubled. The right sides are also doubled, because the derivative of twice a function is twice the derivative of the original function. Thus if dim light moving with a particular velocity is a solution, then so is bright light, provided that it has the same velocity.

We can now see that all sinusoidal waves have the same velocity. What about nonsinusoidal waves like the one in figure l? There is a mathematical theorem, due to Fourier, that says any function can be made by adding together sinusoidal functions. For instance, $3\sin x - 7\cos 3x$ can be made by adding together the functions $3\sin x$ and $-7\cos 3x$, but Fourier proved that this can be done even for functions, like figure l, that aren't obviously built out of sines and cosines in the first place. Therefore our proof that sinusoidal waves all have the same velocity is sufficient to demonstrate that other waves also have this same velocity.

We're now ready to prove that this universal speed for all electromagnetic waves is indeed c. Since we've already convinced ourselves that all such waves travel at the same speed, it's sufficient to find the velocity of one wave in particular. Let's pick the wave whose fields have magnitudes

$$\begin{aligned} E &= \tilde{E}\sin(x+vt) \qquad \text{and} \\ B &= \tilde{B}\sin(x+vt) \qquad , \end{aligned}$$

which is about as simple as we can get. The peak electric field of this wave has a strength $\tilde{E}$, and the peak magnetic field is $\tilde{B}$. The sine functions go through one complete cycle as x increases by $2\pi = 6.28\ldots$, so the distance from one peak of this wave to the next — its wavelength — is $6.28\ldots$ meters. This means that it is not a wave of visible light but rather a radio wave (its wavelength is on the same order of magnitude as the size of a radio antenna). That's OK. What was glorious about Maxwell's work was that it unified the whole electromagnetic spectrum. Light is simple. Radio waves aren't fundamentally any different than light waves, x-rays, or gamma rays. [16]

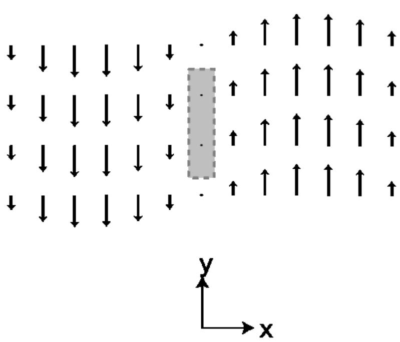
The justification for putting $x + vt$ inside the sine functions is as follows. As the wave travels through space, the whole pattern just

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.

shifts over. The fields are zero at $x=0$, $t=0$, since the sine of zero is zero. This zero-point of the wave pattern shifts over as time goes by; at any time t its location is given by $x+vt=0$. After one second, the zero-point is located at $x=-(1\ \text{s})v$. The distance it travels in one second is therefore numerically equal to v, and this is exactly the concept of velocity: how far something goes per unit time.

The wave has to satisfy Maxwell's equations for Γ_E and Γ_B regardless of what Ampèrian surfaces we pick, and by applying them to any surface, we could determine the speed of the wave. The surface shown in figure m turns out to result in an easy calculation: a narrow strip of width 2ℓ and height h, coinciding with the position of the zero-point of the field at $t=0$.



m / The magnetic field of the wave. The electric field, not shown, is perpendicular to the page.

Now let's apply the equation $c^2\Gamma_B = \partial\Phi_E/\partial t$ at $t=0$. Since the strip is narrow, we can approximate the magnetic field using $\sin x \approx x$, which is valid for small x. The magnetic field on the right edge of the strip, at $x=\ell$, is then $\tilde{B}\ell$, so the right edge of the strip contributes $\tilde{B}\ell h$ to the circulation. The left edge contributes the same amount, so the left side of Maxwell's equation is

$$c^2\Gamma_B = c^2 \cdot 2\tilde{B}\ell h \qquad .$$

The other side of the equation is

$$\begin{aligned}\frac{\partial\Phi_E}{\partial t} &= \frac{\partial}{\partial t}(EA)\\ &= 2\ell h\frac{\partial E}{\partial t} \qquad ,\end{aligned}$$

where we can dispense with the usual sum because the strip is narrow and there is no variation in the field as we go up and down the strip. The derivative equals $v\tilde{E}\cos(x+vt)$, and evaluating the cosine at $x=0$, $t=0$ gives

$$\frac{\partial\Phi_E}{\partial t} = 2v\tilde{E}\ell h$$

Maxwell's equation for Γ_B therefore results in

$$\begin{aligned}2c^2\tilde{B}\ell h &= 2\tilde{E}\ell h v\\ c^2\tilde{B} &= v\tilde{E} \qquad .\end{aligned}$$

An application of $\Gamma_E = -\partial\Phi_B/\partial t$ gives a similar result, except that there is no factor of c^2

$$\tilde{E} = v\tilde{B} \qquad .$$

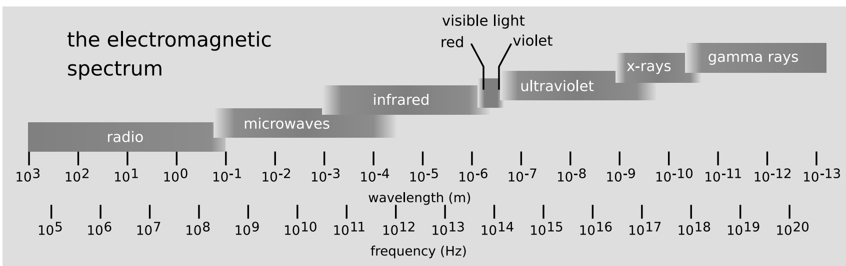
(The minus sign simply represents the right-handed relationship of the fields relative to their direction of propagation.)

Multiplying these last two equations by each other, we get

$$\begin{aligned} c^2\tilde{B}\tilde{E} &= v^2\tilde{E}\tilde{B} \\ c^2 &= v^2 \\ v &= \pm c \qquad . \end{aligned}$$

This is the desired result. (The plus or minus sign shows that the wave can travel in either direction.)

As a byproduct of this calculation, we can find the relationship between the strengths of the electric and magnetic fields in an electromagnetic wave. If, instead of multiplying the equations $c^2\tilde{B} = v\tilde{E}$ and $\tilde{E} = v\tilde{B}$, we divide them, we can easily show that $\tilde{E} = c\tilde{B}$.



n / The electromagnetic spectrum.

Figure n shows the complete spectrum of light waves. The wavelength λ (number of meters per cycle) and frequency f (number of cycles per second) are related by the equation $c = f\lambda$. Maxwell's equations predict that all light waves have the same structure, regardless of wavelength and frequency, so even though radio and x-rays, for example, hadn't been discovered, Maxwell predicted that such waves would have to exist. Maxwell's 1865 prediction passed an important test in 1888, when Heinrich Hertz published the results of experiments in which he showed that radio waves could be manipulated in the same ways as visible light waves. Hertz showed, for example, that radio waves could be reflected from a flat surface, and that the directions of the reflected and incoming waves were related in the same way as with light waves, forming equal angles with the surface. Likewise, light waves can be focused with a curved, dish-shaped mirror, and Hertz demonstrated the same thing with a dish-shaped radio antenna.

Momentum of light waves

A light wave consists of electric and magnetic fields, and fields contain energy. Thus a light wave carries energy with it when it travels from one place to another. If a material object has kinetic energy and moves from one place to another, it must also have momentum, so it is logical to ask whether light waves have momentum as well.

It can be proved based on relativity that it does, and that the momentum and energy are related by the equation $U = p/c$, where p is the magnitude of the momentum vector, and $U = U_e + U_m$ is the sum of the energy of the electric and magnetic fields (see homework problem 11 on page 342). We can now demonstrate this without explicitly referring to relativity, and connect it to the specific structure of a light wave.

The energy density of a light wave is related to the magnitudes of the fields in a specific way — it depends on the squares of their magnitudes, E^2 and B^2, which are the same as the dot products $\mathbf{E}\cdot\mathbf{E}$ and $\mathbf{B}\cdot\mathbf{B}$. We argued on page 474 that since energy is a scalar, the only possible expressions for the energy densities of the fields are dot products like these, multiplied by some constants. This is because the dot product is the only mathematically sensible way of multiplying two vectors to get a scalar result. (Any other way violates the symmetry of space itself.)

How does this relate to momentum? Well, we know that if we double the strengths of the fields in a light beam, it will have four times the energy, because the energy depends on the square of the fields. But we then know that this quadruple-energy light beam must have quadruple the momentum as well. If there wasn't this kind of consistency between the momentum and the energy, then we could violate conservation of momentum by combining light beams or splitting them up. We therefore know that the momentum density of a light beam must depend on a field multiplied by a field. Momentum, however, is a vector, and there is only one physically meaningful way of multiplying two vectors to get a vector result, which is the cross product (see page 694). The momentum density can therefore only depend on the cross products $\mathbf{E}\times\mathbf{E}$, $\mathbf{B}\times\mathbf{B}$, and $\mathbf{E}\times\mathbf{B}$. But the first two of these are zero, since the cross product vanishes when there is a zero angle between the vectors. Thus the momentum per unit volume must equal $\mathbf{E}\times\mathbf{B}$ multiplied by some constant,

$$\mathrm{d}\mathbf{p} = (\text{constant})\mathbf{E}\times\mathbf{B}\,\mathrm{d}v$$

This predicts something specific about the direction of propagation of a light wave: it must be along the line perpendicular to the electric and magnetic fields. We've already seen that this is correct, and also that the electric and magnetic fields are perpendicular to each other. Therefore this cross product has a magnitude

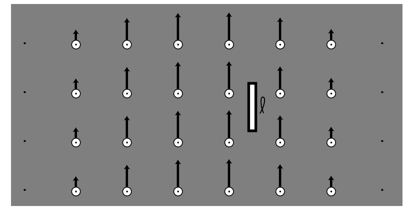
$$\begin{aligned}|\mathbf{E}\times\mathbf{B}| &= |\mathbf{E}||\mathbf{B}|\sin 90^\circ \\ &= |\mathbf{E}||\mathbf{B}| \\ &= \frac{|\mathbf{E}|^2}{c} = c|\mathbf{B}|^2 \qquad ,\end{aligned}$$

where in the last step the relation $|\mathbf{E}| = c|\mathbf{B}|$ has been used.

We now only need to find one physical example in order to fix the constant of proportionality. Indeed, if we didn't know relativity, it would be possible to believe that the constant of proportionality was zero! The simplest example of which I know is as follows. Suppose a piece of wire of length ℓ is bathed in electromagnetic waves coming in sideways, and let's say for convenience that this is a radio wave, with a wavelength that is large compared to ℓ, so that the fields don't change significantly across the length of the wire. Let's say the electric field of the wave happens to be aligned with the wire. Then there is an emf between the ends of the wire which equals $E\ell$, and since the wire is small compared to the wavelength, we can pretend that the field is uniform, not curly, in which case voltage is a well-defined concept, and this is equivalent to a voltage difference $\Delta V = E\ell$ between the ends of the wire. The wire obeys Ohm's law, and a current flows in response to the wave.[17] Equating the expressions $\mathrm{d}U/\mathrm{d}t$ and $I\Delta V$ for the power dissipated by ohmic heating, we have

$$\mathrm{d}U = IE\ell\mathrm{d}t$$

for the energy the wave transfers to the wire in a time interval $\mathrm{d}t$.



o / A classical calculation of the momentum of a light wave. An antenna of length ℓ is bathed in an electromagnetic wave. The black arrows represent the electric field, the white circles the magnetic field coming out of the page. The wave is traveling to the right.

Note that although some electrons have been set in motion in the wire, we haven't yet seen any momentum transfer, since the protons are experiencing the same amount of electric force in the opposite direction. However, the electromagnetic wave also has a magnetic field, and a magnetic field transfers momentum to (exerts a force on) a current. This is only a force on the electrons, because they're what make the current. The magnitude of this force equals ℓIB (homework problem 10), and using the definition of force, $\mathrm{d}\mathbf{p}/\mathrm{d}t$, we find for the magnitude of the momentum transferred:

$$\mathrm{d}p = \ell IB\mathrm{d}t$$

We now know both the amount of energy and the amount of momentum that the wave has lost by interacting with the wire. Dividing these two equations, we find

$$\begin{aligned}\frac{\mathrm{d}p}{\mathrm{d}U} &= \frac{B}{E}\\ &= \frac{1}{c} \qquad ,\end{aligned}$$

which is what we expected based on relativity. This can now be restated in the form $\mathrm{d}\mathbf{p} = (\text{constant})\mathbf{E}\times\mathbf{B}\,\mathrm{d}v$ (homework problem 37).

Note that although the equations $p = U/c$ and $\mathrm{d}\mathbf{p} = (\text{constant})\mathbf{E}\times\mathbf{B}\,\mathrm{d}v$ are consistent with each other for a sine wave, they are not consistent with each other in general. The relativistic argument leading

[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!

up to $p = U/c$ assumed that we were only talking about a single thing traveling in a single direction, whereas no such assumption was made in arguing for the $\mathbf{E} \times \mathbf{B}$ form. For instance, if two light beams of equal strength are traveling through one another, going in opposite directions, their total momentum is zero, which is consistent with the $\mathbf{E} \times \mathbf{B}$ form, but not with U/c.

Some examples were given in chapter 3 of situations where it actually matters that light has momentum.

Angular momentum of light waves

For completeness, we note that since light carries momentum, it must also be possible for it to have angular momentum. If you've studied chemistry, here's an example of why this can be important. You know that electrons in atoms can exist in states labeled s, p, d, f, and so on. What you might not have realized is that these are angular momentum labels. The s state, for example, has zero angular momentum. If light didn't have angular momentum, then, for example, it wouldn't be possible for a hydrogen atom in a p state to change to the lower-energy s state by emitting light. Conservation of angular momentum requires that the light wave carry away all the angular momentum originally possessed by the electron in the p state, since in the s state it has none.

This chapter is summarized on page 737. Notation and terminology are tabulated on pages 718-719.

Problems

The symbols √, ■, etc. are explained on page 610.

1 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:

$\mathbf{v}$	along the fingers, with the hand flat
$\mathbf{B}$	along the fingers, with the knuckles bent
$\mathbf{F}$	along the thumb

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? ■

2 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. √ ■

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 The following data give the results of two experiments in which charged particles were released from the same point in space, and the forces on them were measured:

$$\begin{aligned} &q_1 = 1\ \mu\text{C}\,, && \mathbf{v}_1 = (1\ \text{m/s})\hat{\mathbf{x}}\,, && \mathbf{F}_1 = (-1\ \text{mN})\hat{\mathbf{y}} \\ &q_2 = -2\ \mu\text{C}\,, && \mathbf{v}_2 = (-1\ \text{m/s})\hat{\mathbf{x}}\,, && \mathbf{F}_2 = (-2\ \text{mN})\hat{\mathbf{y}} \end{aligned}$$

The data are insufficient to determine the magnetic field vector; demonstrate this by giving two different magnetic field vectors, both of which are consistent with the data.

6 The following data give the results of two experiments in which charged particles were released from the same point in space, and the forces on them were measured:

$$\begin{aligned} &q_1 = 1\ \text{nC}\,, && \mathbf{v}_1 = (1\ \text{m/s})\hat{\mathbf{z}}\,, && \mathbf{F}_1 = (5\ \text{pN})\hat{\mathbf{x}} + (2\ \text{pN})\hat{\mathbf{y}} \\ &q_2 = 1\ \text{nC}\,, && \mathbf{v}_2 = (3\ \text{m/s})\hat{\mathbf{z}}\,, && \mathbf{F}_2 = (10\ \text{pN})\hat{\mathbf{x}} + (4\ \text{pN})\hat{\mathbf{y}} \end{aligned}$$

Is there a nonzero electric field at this point? A nonzero magnetic field?

7 The following data give the results of three experiments in which charged particles were released from the same point in space, and the forces on them were measured:

$$\begin{aligned} &q_1 = 1\ \text{C}\,, && \mathbf{v}_1 = 0\,, && \mathbf{F}_1 = (1\ \text{N})\hat{\mathbf{y}} \\ &q_2 = 1\ \text{C}\,, && \mathbf{v}_2 = (1\ \text{m/s})\hat{\mathbf{x}}\,, && \mathbf{F}_2 = (1\ \text{N})\hat{\mathbf{y}} \\ &q_3 = 1\ \text{C}\,, && \mathbf{v}_3 = (1\ \text{m/s})\hat{\mathbf{z}}\,, && \mathbf{F}_3 = 0 \end{aligned}$$

Determine the electric and magnetic fields. ✓

8 In problem 7, 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 **E** and **B**? If so, design it. If not, why not?

9 (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? ▷ Answer, p. 712
(b) Show that the units of your answer in part a work out correctly.

Remark: 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.

10 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. √

(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 9a. Now find the magnetic force wire 1 makes on wire 2, in terms of I_1, I_2, L, and R. ▷ Answer, p. 712

(c) Show that the units of the answer to part b work out to be newtons.

11 This problem is a continuation of problem 10. Note that the answer to problem 10b is given on page 712.

(a) Interchanging the 1's and 2's in the answer to problem 10b, 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?

12 (a) In the photo of the vacuum tube apparatus in figure n on page 547, 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 11.

13 A charged particle of mass m and charge q moves in a circle due to a uniform magnetic field of magnitude B, which points perpendicular to the plane of the circle.

(a) Assume the particle is positively charged. Make a sketch showing the direction of motion and the direction of the field, and show that the resulting force is in the right direction to produce circular motion.

(b) Find the radius, r, of the circle, in terms of m, q, v, and B. √

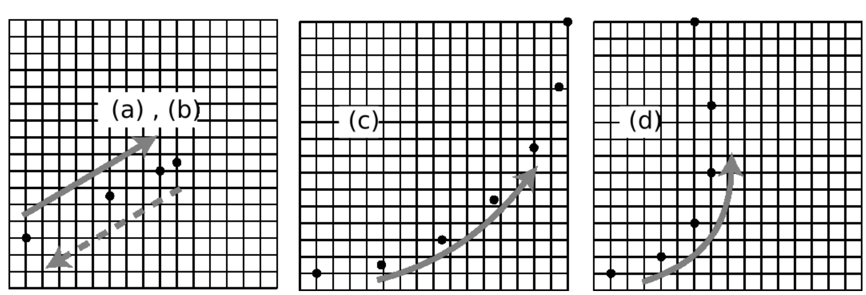
(c) Show that your result from part b has the right units.

(d) Discuss all four variables occurring on the right-hand side of your answer from part b. Do they make sense? For instance, what should happen to the radius when the magnetic field is made stronger? Does your equation behave this way?
(e) Restate your result so that it gives the particle's angular frequency, ω, in terms of the other variables, and show that v drops out. $\checkmark$

Remark: A charged particle can be accelerated in a circular device called a cyclotron, in which a magnetic field is what keeps them from going off straight. This frequency is therefore known as the cyclotron frequency. The particles are accelerated by other forces (electric forces), which are AC. As long as the electric field is operated at the correct cyclotron frequency for the type of particles being manipulated, it will stay in sync with the particles, giving them a shove in the right direction each time they pass by. The particles are speeding up, so this only works because the cyclotron frequency is independent of velocity.

14 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.)

15 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.



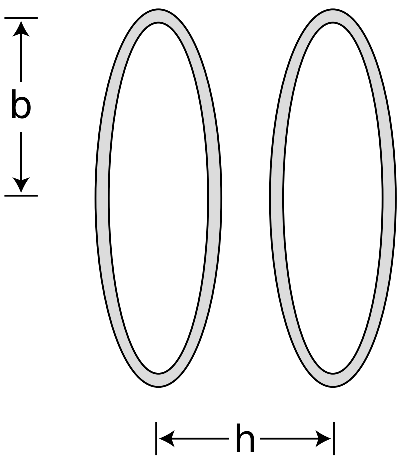
Problem 15.

16 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 549.

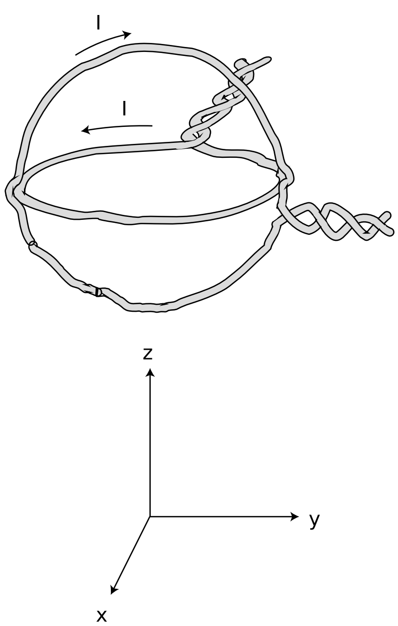
17 (a) Modify the calculation on page 554 to determine the component of the magnetic field of a sheet of charge that is perpendicular to the sheet. √
(b) Show that your answer has the right units.
(c) Show that your answer approaches zero as z approaches infinity.
(d) What happens to your answer in the case of $a = b$? Explain why this makes sense.

18 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. √
(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. √
(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.)

19 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 $w \gg h$. √
(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.



Problem 21.



Problem 24.

20 If you put four times more current through a solenoid, how many times more energy is stored in its magnetic field? √

21 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 n on page 547.

(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. √

(b) What value of h (not equal to b) would make this percentage difference equal to zero? √

22 The equation $B_z = \beta kIA/c^2r^3$ was found on page 557 for the distant field of a dipole. Show, as asserted there, that the constant β must be unitless.

23 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? ▷ Hint, p. 705

24 The figure shows a nested pair of circular wire loops used to create magnetic fields. (The twisting of the leads is a practical trick for reducing the magnetic fields they contribute, so the fields are very nearly what we would expect for an ideal circular current loop.) The coordinate system below is to make it easier to discuss directions in space. One loop is in the $y-z$ plane, the other in the $x-y$ plane. Each of the loops has a radius of 1.0 cm, and carries 1.0 A in the direction indicated by the arrow.

(a) Calculate the magnetic field that would be produced by *one* such loop, at its center. √

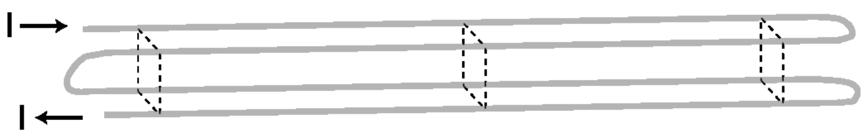
(b) Describe the direction of the magnetic field that would be produced, at its center, by the loop in the $x-y$ plane alone.

(c) Do the same for the other loop.

(d) Calculate the magnitude of the magnetic field produced by the two loops in combination, at their common center. Describe its direction. √

25 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.

26 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. $\surd$
(b) Based on your answer to part a, find the force acting between the solenoids. $\surd$



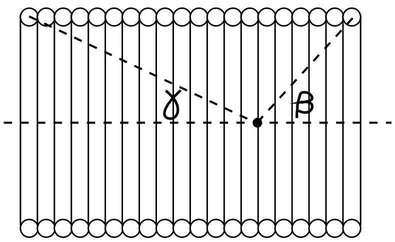
Problem 27.

27 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. $\surd$

28 (a) For the geometry described in example 8 on page 552, 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 557. 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. √

29 (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. √
(b) Show that your answer has the right units to be an inductance.



Problem 30.

30 (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 561 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 565 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?

31 The first step in the proof of Ampère's law on page 566 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.

32 On page 571, 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.

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.

35 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.)

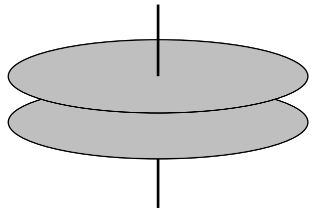
36 (a) For each term appearing on the right side of Maxwell's equations, give an example of an everyday situation it describes.
(b) Most people doing calculations in the SI system of units don't use k and k/c^2. Instead, they express everything in terms of the constants

$$\epsilon_o = \frac{1}{4\pi k} \qquad \text{and}$$
$$\mu_o = \frac{4\pi k}{c^2} \qquad .$$

Rewrite Maxwell's equations in terms of these constants, eliminating k and c everywhere.

37 (a) Prove that in an electromagnetic 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 $\mathrm{d}\mathbf{p} \propto \mathbf{E} \times \mathbf{B}\, \mathrm{d}v$, where $\mathrm{d}\mathbf{p}$ is the momentum of the part of a plane light wave contained in the volume $\mathrm{d}v$. 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.) ✓

38 (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 37a. A real beam of light usually consists of many short wavetrains, not one big sine wave, but don't worry about that. ✓ ▷ Hint, p. 705
(c) A beam of sunlight has an intensity of $P/A = 1.35 \times 10^3$ W/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: $E \sim 10^6$ V/m in a thunderstorm, and $B \sim 10^{-3}$ T for the Earth's magnetic field. ✓



Problem 39.

39 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. ▷ Hint, p. 705 ✓

40 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, p. 717

(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, p. 717

(c) Discuss how currents and magnetic fields would behave under time reversal. ▷ Hint, p. 705

(d) Similarly, show that the equation $\mathrm{d}\mathbf{p} \propto \mathbf{E} \times \mathbf{B}$ is still valid under time reversal.

41 This problem is a more advanced exploration of the time-reversal ideas introduced in problem 40.

(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 $\mathrm{d}\mathbf{p} \propto \mathbf{E} \times \mathbf{B}$ 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 $q \to -q$.

42 The purpose of this problem is to prove that the constant of proportionality a in the equation $\mathrm{d}U_m = aB^2\mathrm{d}v$, for the energy density of the magnetic field, is given by $a = c^2/8\pi k$ as asserted on page 555. 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 555, 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.) $\checkmark$

(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. $\checkmark$

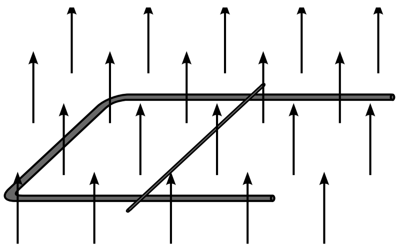
(d) The power supplied by the battery is $P = I\Gamma_E$ (like $P = I\Delta V$, but with an emf instead of a voltage difference), and the circulation is given by $\Gamma = -\mathrm{d}\Phi_B/\mathrm{d}t$. The negative sign indicates that the battery is being drained. Calculate the energy supplied by the battery, and, as in part c, show that the result can be stated in terms of η, v, and universal constants. $\checkmark$

(e) Find the increase in the magnetic-field energy, in terms of η, v, and the unknown constant a. $\checkmark$

(f) Use conservation of energy to relate your answers from parts c, d, and e, and solve for a. $\checkmark$

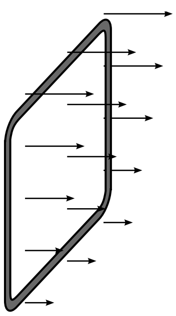
43 A positively charged particle is released from rest at the origin at $t = 0$, in a region of vacuum through which an electromagnetic wave is passing. The particle accelerates in response to the wave. In this region of space, the wave varies as $\mathbf{E} = \hat{\mathbf{x}}\tilde{E}\sin\omega t$, $\mathbf{B} = \hat{\mathbf{y}}\tilde{B}\sin\omega t$, 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. ✓
(b) Show that the motion is confined to $-z_{max} \le z \le z_{max}$, where $z_{max} = 1.101\left(q^2\tilde{E}\tilde{B}/m^2\omega^3\right)$.



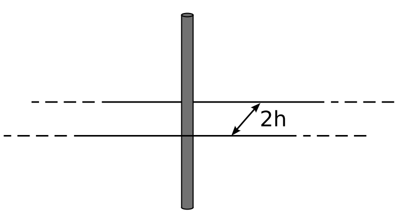
Problem 44.

44 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).



Problem 45.

45 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.



Problem 46.

46 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.

47 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.

Key to symbols:
easy typical challenging difficult very difficult
✓ An answer check is available at www.lightandmatter.com.

Exercises

Exercise 11B: Polarization

Apparatus:

calcite (Iceland spar) crystal

polaroid film

1. Lay the crystal on a piece of paper that has print on it. You will observe a double image. See what happens if you rotate the crystal.

Evidently the crystal does something to the light that passes through it on the way from the page to your eye. One beam of light enters the crystal from underneath, but two emerge from the top; by conservation of energy the energy of the original beam must be shared between them. Consider the following three possible interpretations of what you have observed:

(a) The two new beams differ from each other, and from the original beam, only in energy. Their other properties are the same.

(b) The crystal adds to the light some mysterious new property (not energy), which comes in two flavors, X and Y. Ordinary light doesn't have any of either. One beam that emerges from the crystal has some X added to it, and the other beam has Y.

(c) There is some mysterious new property that is possessed by all light. It comes in two flavors, X and Y, and most ordinary light sources make an equal mixture of type X and type Y light. The original beam is an even mixture of both types, and this mixture is then split up by the crystal into the two purified forms.

In parts 2 and 3 you'll make observations that will allow you to figure out which of these is correct.

2. Now place a polaroid film over the crystal and see what you observe. What happens when you rotate the film in the horizontal plane? Does this observation allow you to rule out any of the three interpretations?

3. Now put the polaroid film under the crystal and try the same thing. Putting together all your observations, which interpretation do you think is correct?

4. Look at an overhead light fixture through the polaroid, and try

rotating it. What do you observe? What does this tell you about the light emitted by the lightbulb?

5. Now position yourself with your head under a light fixture and directly over a shiny surface, such as a glossy tabletop. You'll see the lamp's reflection, and the light coming from the lamp to your eye will have undergone a reflection through roughly a 180-degree angle (i.e. it very nearly reversed its direction). Observe this reflection through the polaroid, and try rotating it. Finally, position yourself so that you are seeing glancing reflections, and try the same thing. Summarize what happens to light with properties X and Y when it is reflected. (This is the principle behind polarizing sunglasses.)

Chapter 12
Quantum Physics

12.1 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.

a / In 1980, the continental U.S. got its first taste of active volcanism in recent memory with the eruption of Mount St. Helens.

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]

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.

Discussion Question

A 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?

[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.

12.1.1 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.

Discussion Question

A 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?

12.1.2 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.

Statistical independence

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.

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.

Addition of probabilities

The law of independent probabilities tells us to use multiplication to calculate the probability that both A and B will happen, assuming the probabilities are independent. What about the probability of an "or" rather than an "and"? If two events A and B 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 $70\% + 30\% = 100\%$. Water and land are mutually exclusive, and there are no other possibilities, so the probabilities had to add up to 100%. It works the same if there are more than two possibilities — if you can classify all possible outcomes into a list of mutually exclusive results, then all the probabilities have to add up to 1, or 100%. This property of probabilities is known as normalization.

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 game mentioned above, 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.

Self-Check
Which of the following things *must* have 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, p. 711

Discussion Questions

A 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?

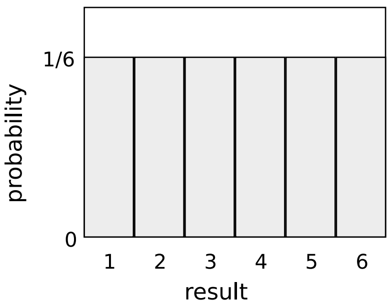
B Why isn't it valid to define randomness by saying that randomness is when all the outcomes are equally likely?

C The sequence of digits 121212121212121212 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?

b / Why are dice random?

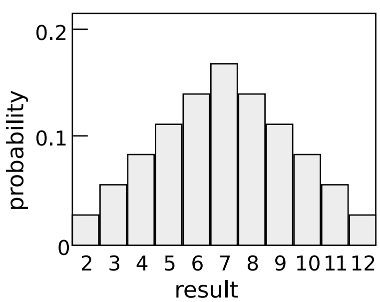
12.1.3 Probability distributions

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.



c / Probability distribution for the result of rolling a single die.

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, c. 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.



d / Rolling two dice and adding them up.

Figure d 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: 2.999999999999999996876876587658465436, 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.

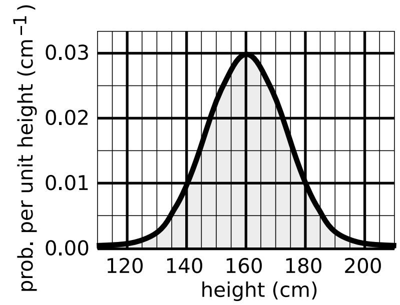
But the y axis can no longer be a unitless probability scale. In radioactive decay, for example, we want the x axis to have units of time, and we want areas under the curve to be unitless probabilities.

The area of a single square on the graph paper is then

$$
\begin{aligned}
&(\text{unitless area of a square})\\
&= (\text{width of square with time units})\\
&\quad\times(\text{height of square}) \qquad .
\end{aligned}
$$

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 e shows another example, a probability distribution for people's height. This kind of bell-shaped curve is quite common.



e / A probability distribution for height of human adults (not real data).

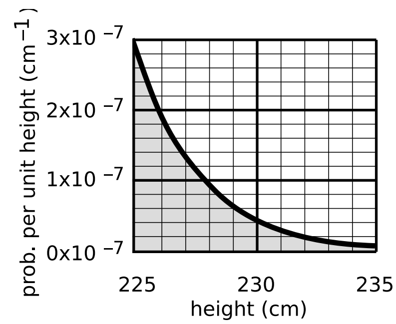
Self-Check

Compare the number of people with heights in the range of 130-135 cm to the number in the range 135-140. ▷ Answer, p. 711

Looking for tall basketball players — *example 1*

▷ A certain country with a large population wants to find very tall people to be on its Olympic basketball team and strike a blow against western imperialism. Out of a pool of 10^8 people who are the right age and gender, how many are they likely to find who are over 225 cm (7 feet 4 inches) in height? Figure (d) gives a close-up of the "tails" of the distribution shown previously.

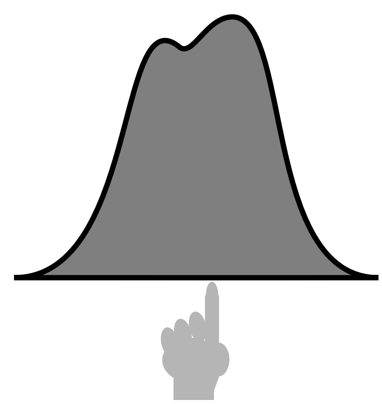
▷ The shaded area under the curve represents the probability that a given person is tall enough. Each rectangle represents a probability of 0.2×10^{-7} cm^{-1} $\times$ 1 cm = 2×10^{-9}. There are about 35 rectangles covered by the shaded area, so the probability of having a height greater than 230 cm is 7×10^{-8} , or just under one in ten million. Using the rule for calculating averages, the average, or expected number of people this tall is $(10^8) \times (7 \times 10^{-8})$ = 7.



f / Example 1.

Average and width of a probability distribution

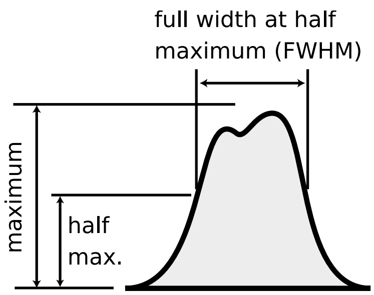
If the next Martian you meet asks you, "How tall is an adult human?," you will probably reply with a statement about the average human height, such as "Oh, about 5 feet 6 inches." If you wanted to explain a little more, you could say, "But that's only an average. Most people are somewhere between 5 feet 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.



g / 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 h.

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



h / The full width at half maximum (FWHM) of a probability distribution.

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).

12.1.4 Exponential decay and half-life

Half-life

Most people know that radioactivity "lasts a certain amount of time," but that simple statement leaves out a lot. As an example, consider the following medical procedure used to diagnose thyroid function. A very small quantity of the isotope ^{131}I, produced in a nuclear reactor, is fed to or injected into the patient. The body's biochemical systems treat this artificial, radioactive isotope exactly the same as ^{127}I, which is the only naturally occurring type. (Nutritionally, iodine is a necessary trace element. Iodine taken into the body is partly excreted, but the rest becomes concentrated in the thyroid gland. Iodized salt has had iodine added to it to prevent the nutritional deficiency known as goiters, in which the iodine-starved thyroid becomes swollen.) As the ^{131}I undergoes beta decay, it emits electrons, neutrinos, and gamma rays. The gamma rays can be measured by a detector passed over the patient's body. As the radioactive iodine becomes concentrated in the thyroid, the amount of gamma radiation coming from the thyroid becomes greater, and that emitted by the rest of the body is reduced. The rate at which the iodine concentrates in the thyroid tells the doctor about the health of the thyroid.

If you ever undergo this procedure, someone will presumably explain a little about radioactivity to you, to allay your fears that you will turn into the Incredible Hulk, or that your next child will have an unusual number of limbs. Since iodine stays in your thyroid for a long time once it gets there, one thing you'll want to know is whether your thyroid is going to become radioactive forever. They may just tell you that the radioactivity "only lasts a certain amount of time," but we can now carry out a quantitative derivation of how the radioactivity really will die out.

Let $P_{surv}(t)$ 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}I atoms decay in 8 hours, so we have

$$P_{surv}(8 \text{ hr}) = 0.5 \qquad .$$

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,

$$\begin{aligned} P_{surv}(16 \text{ hr}) &= 0.50 \times 0.50 \\ &= 0.25 \qquad . \end{aligned}$$

Similarly we have

$$\begin{aligned} P_{surv}(24 \text{ hr}) &= 0.50 \times 0.5 \times 0.5 \\ &= 0.125 \qquad . \end{aligned}$$

Generalizing from this pattern, the probability of surviving for any time t that is a multiple of 8 hours is

$$P_{surv}(t) = 0.5^{t/8\text{ hr}} \qquad .$$

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

$$P_{surv}(8\text{ hr}) = P_{surv}(4\text{ hr}) \times P_{surv}(4\text{ hr}) \qquad ,$$

which can be rearranged to give

$$\begin{aligned} P_{surv}(4\text{ hr}) &= \sqrt{P_{surv}(8\text{ hr})} \\ &= \sqrt{0.5} \\ &= 0.707 \qquad . \end{aligned}$$

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:

Exponential Decay Equation

$$P_{surv}(t) = 0.5^{t/t_{1/2}}$$

Using the rule for calculating averages, we can also find the number of atoms, $N(t)$, remaining in a sample at time t:

$$N(t) = N(0) \times 0.5^{t/t_{1/2}}$$

Both of these equations have graphs that look like dying-out exponentials, as in the example below.

Radioactive contamination at Chernobyl *example 2*

▷ 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?

▷ (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,

$$P_{surv} = 0.5^{t/t_{1/2}} \qquad ,$$

we want to solve for t. Taking natural logarithms of both sides,

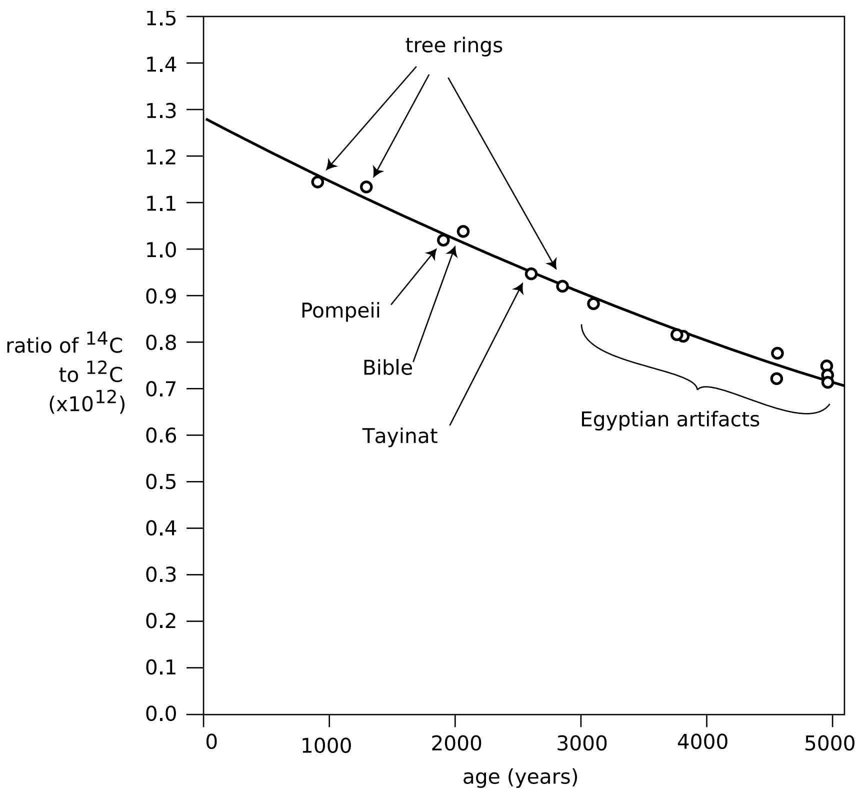
$$\ln P = \frac{t}{t_{1/2}} \ln 0.5 \qquad ,$$

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.



i / Calibration of the ^{14}C dating method using tree rings and artifacts whose ages were known from other methods. Redrawn from Emilio Segrè, **Nuclei and Particles**, 1965.

^{14}C *Dating* *example 3*

Almost all the carbon on Earth is ^{12}C, but not quite. The isotope ^{14}C, with a half-life of 5600 years, is produced by cosmic rays in the atmosphere. It decays naturally, but is replenished at such a rate that the fraction of ^{14}C in the atmosphere remains constant, at 1.3×10^{-12}. Living plants and animals take in both ^{12}C and ^{14}C from the atmosphere and incorporate both into their bodies. Once the living organism dies, it no longer takes in C atoms from the atmosphere, and the proportion of ^{14}C gradually falls off as it undergoes radioactive decay. This effect can be used to find the age of dead organisms, or human artifacts made from plants or animals. Figure i on page 625 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.

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:

$$\begin{aligned}
&(\text{number of decays between } t \text{ and } t+\Delta t)\\
&\qquad = N(t) - N(t+\Delta t)\\
&\qquad = N(0)\left[P_{surv}(t) - P_{surv}(t+\Delta t)\right]\\
&\qquad = N(0)\left[0.5^{t/t_{1/2}} - 0.5^{(t+\Delta t)/t_{1/2}}\right]\\
&\qquad = N(0)\left[1 - 0.5^{\Delta t/t_{1/2}}\right]0.5^{t/t_{1/2}}
\end{aligned}$$

A problem arises when Δt is small compared to $t_{1/2}$. For instance, suppose you have a hunk of 10^{22} atoms of $^{235}\mathrm{U}$, with a half-life of 700 million years, which is 2.2×10^{16} s. You want to know how many decays will occur in $\Delta t = 1$ s. Since we're specifying the current number of atoms, $t = 0$. As you plug in to the formula above on your calculator, the quantity $0.5^{\Delta t/t_{1/2}}$ comes out on your calculator to equal one, so the final result is zero. That's incorrect, though. In reality, $0.5^{\Delta t/t_{1/2}}$ should equal 0.999999999999999968, but your calculator only gives eight digits of precision, so it rounded it off to one. In other words, the probability that a $^{235}\mathrm{U}$ atom will survive for 1 s is very close to one, but not equal to one. The number of decays in one second is therefore 3.2×10^5, not zero.

Well, my calculator only does eight digits of precision, just like yours, so how did I know the right answer? The way to do it is to use the following approximation:

$$a^b \approx 1 + b\,\ln a, \quad \text{if } b \ll 1$$

(The symbol $\ll$ means "is much less than.") Using it, we can find the following approximation:

$$\begin{aligned}
&(\text{number of decays between } t \text{ and } t+\Delta t)\\
&\qquad = N(0)\left[1 - 0.5^{\Delta t/t_{1/2}}\right]0.5^{t/t_{1/2}}\\
&\qquad \approx N(0)\left[1 - \left(1 + \frac{\Delta t}{t_{1/2}}\ln 0.5\right)\right]0.5^{t/t_{1/2}}\\
&\qquad \approx (\ln 2)N(0)0.5^{t/t_{1/2}}\frac{\Delta t}{t_{1/2}}
\end{aligned}$$

This also gives us a way to calculate the rate of decay, i.e. the number of decays per unit time. Dividing by Δt on both sides, we have

$$(\text{decays per unit time}) \approx \frac{(\ln 2)N(0)}{t_{1/2}}0.5^{t/t_{1/2}} \quad , \quad \text{if } \Delta t \ll t_{1/2}$$

The hot potato *example 4*

▷ 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?

▷ 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 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.

If the sheer number of equations is starting to seem formidable, let's pause and think for a second. The simple equation for P_{surv} is something you can derive easily from the law of independent probabilities any time you need it. From that, you can quickly find the exact equation for the rate of decay. The derivation of the approximate equations for $\Delta t \ll t$ is a little hairier, but note that except for the factors of $\ln 2$, everything in these equations can be found simply from considerations of logic and units. For instance, a longer half-life will obviously lead to a slower rate of decays, so it makes sense that we divide by it. As for the $\ln 2$ factors, they are exactly the kind of thing that one looks up in a book when one needs to know them.

Discussion Questions

A 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?

B For 1 s, Fred holds in his hands 1 kg of radioactive stuff with a half-life of 1000 years. Ginger holds 1 kg of a different substance, with a half-life of 1 min, for the same amount of time. Did they place themselves in equal danger, or not?

C How would you interpret it if you calculated $N(t)$, and found it was less than one?

D 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?

12.1.5 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

$$(\text{probability of } a \le x \le b) = \int_a^b D(x)\mathrm{d}x \qquad .$$

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 $x_1P_1 + x_2P_2 + \ldots$ 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

$$(\text{average value of } x) = \int xD(x)\mathrm{d}x \qquad ,$$

where the integral is over all possible values of x.

Probability distribution for radioactive decay — *example 5*

Here is a rigorous justification for the statement in subsection 12.1.4 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

$$D(t) = k0.5^{t/t_{1/2}} \qquad ,$$

where k is a constant that we need to determine. The atom is guaranteed to decay eventually, so normalization gives us

$$\begin{aligned}(\text{probability of } 0 \le t < \infty) &= 1 \\ &= \int_0^\infty D(t)\mathrm{d}t \qquad .\end{aligned}$$

The integral is most easily evaluated by converting the function into an exponential with e as the base

$$\begin{aligned}D(t) &= k\exp\left[\ln\left(0.5^{t/t_{1/2}}\right)\right] \\ &= k\exp\left[\frac{t}{t_{1/2}}\ln 0.5\right] \\ &= k\exp\left(-\frac{\ln 2}{t_{1/2}}t\right) \qquad ,\end{aligned}$$

which gives an integral of the familiar form $\int e^{cx}\mathrm{d}x = (1/c)e^{cx}$. We thus have

$$1 = -\frac{kt_{1/2}}{\ln 2}\exp\left(-\frac{\ln 2}{t_{1/2}}t\right)\Bigg]_0^\infty \qquad ,$$

which gives the desired result:

$$k = \frac{\ln 2}{t_{1/2}} \qquad .$$

Average lifetime *example 6*

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:

$$\text{(average lifetime)} = \int_0^\infty t\, D(t)\, dt$$

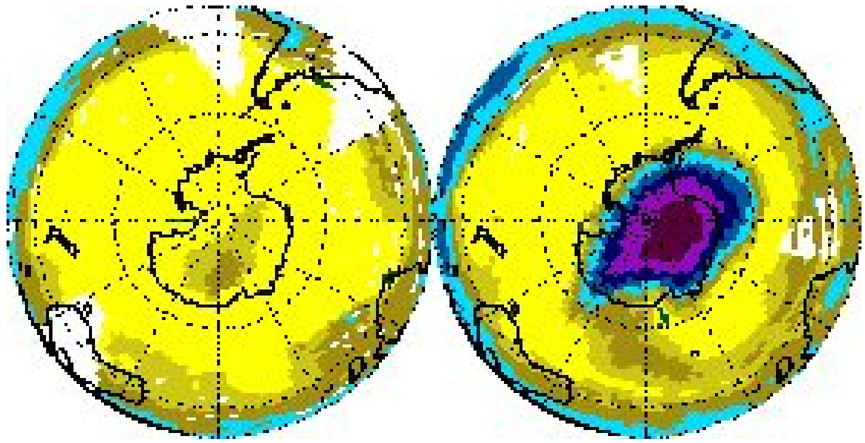
Using the convenient base-e form again, we have

$$\text{(average lifetime)} = \frac{\ln 2}{t_{1/2}} \int_0^\infty t \exp\left(-\frac{\ln 2}{t_{1/2}} t\right) dt \qquad .$$

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

$$\begin{aligned}\text{(average lifetime)} &= \frac{\ln 2}{t_{1/2}} \left(\frac{t_{1/2}}{\ln 2}\right)^2 \\ &= \frac{t_{1/2}}{\ln 2} \\ &= 1.443\, t_{1/2} \qquad ,\end{aligned}$$

which is, as expected, longer than one half-life.



j / In recent decades, a huge hole in the ozone layer has spread out from Antarctica. Left: November 1978. Right: November 1992

12.2 Light as a Particle

The only thing that interferes with my learning is my education.
Albert Einstein

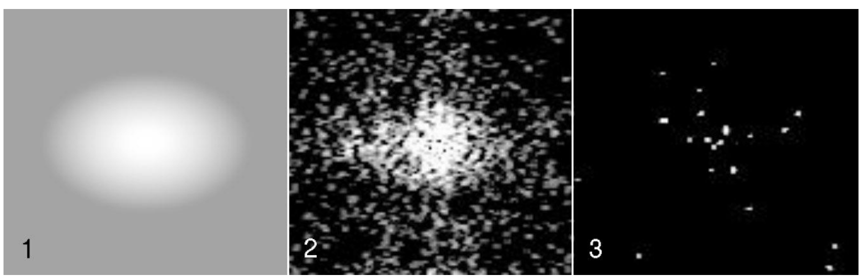
Radioactivity is random, but do the laws of physics exhibit randomness in other contexts besides radioactivity? Yes. Radioactive decay was just a good playpen to get us started with concepts of randomness, because all atoms of a given isotope are identical. By stocking the playpen with an unlimited supply 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!

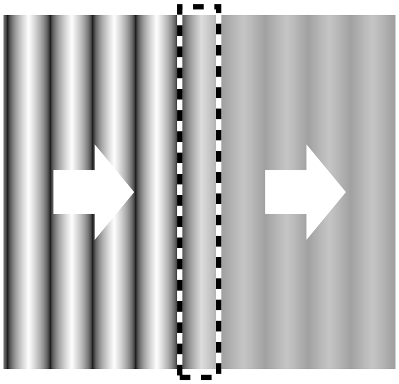
12.2.1 Evidence for light as a particle

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.

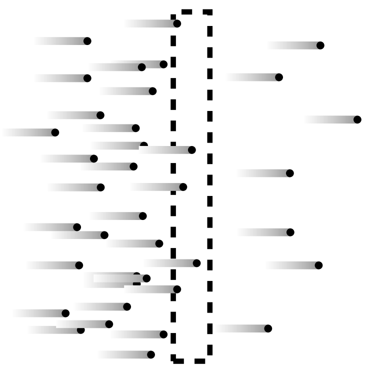
In those days the data were sketchy, the ideas vague, and the experiments difficult to interpret; it took a genius like Einstein to cut through the thicket of confusion and find a simple solution. Today, however, we can get right to the heart of the matter with a piece of ordinary consumer electronics, 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.

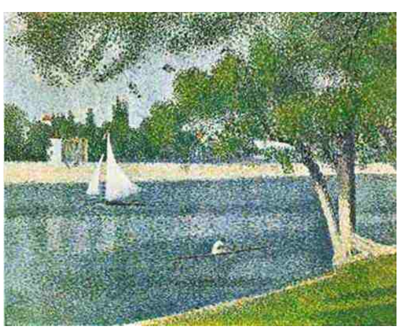
But figures a/2 and a/3 show that some pixels take strong hits while others pick up no energy at all. Instead of the wave picture, the image that is naturally evoked by the data is something more like a hail of bullets from a machine gun, c. Each "bullet" of light apparently carries only a tiny amount of energy, which is why detecting them individually requires a sensitive digital camera rather than an eye or a piece of film.



c / A stream of particles is partially absorbed.

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

A Suppose someone rebuts the digital camera data in figure a, claiming that the random pattern of dots occurs not because of anything fundamental about the nature of light but simply because the camera's pixels are not all exactly the same. How could we test this interpretation?

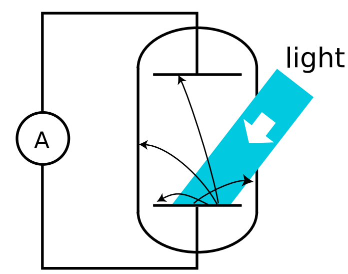
B Discuss how the correspondence principle applies to the observations and concepts discussed in this section.

12.2.2 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.)

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.



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).

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.

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.

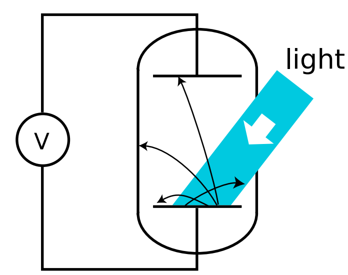
To make any progress, we need to consider the physical process by which a photon would eject an electron from the metal electrode. A metal contains electrons that are free to move around. Ordinarily, in the interior of the metal, such an electron feels attractive forces from atoms in every direction around it. The forces cancel out. But if the electron happens to find itself at the surface of the metal, the attraction from the interior side is not balanced out by any attraction from outside. In popping out through the surface the electron therefore loses some amount of energy E_s, which depends on the type of metal used.

Suppose a photon strikes an electron, annihilating itself and giving up all its energy to the electron. (We now know that this is what always happens in the photoelectric effect, although it had not yet been established in 1905 whether or not the photon was completely annihilated.) The electron will (1) lose kinetic energy through collisions with other electrons as it plows through the metal on its way to the surface; (2) lose an amount of kinetic energy equal to E_s as it emerges through the surface; and (3) lose more energy on its way across the gap between the plates, due to the electric field between the plates. Even if the electron happens to be right at the surface of the metal when it absorbs the photon, and even if the electric field between the plates has not yet built up very much, E_s is the bare minimum amount of energy that it must receive from the photon if it is to contribute to a measurable current. The reason for using very clean electrodes is to minimize E_s and make it have a definite value characteristic of the metal surface, not a mixture of values due to the various types of dirt and crud that are present in tiny amounts on all surfaces in everyday life.

We can now interpret the frequency dependence of the photoelectric effect in a simple way: apparently the amount of energy possessed by a photon is related to its frequency. A low-frequency red or infrared photon has an energy less than E_s, so a beam of them will not produce any current. A high-frequency blue or violet photon, on the other hand, packs enough of a punch to allow an electron to make it to the other plate. At frequencies higher than the minimum, the photoelectric current continues to increase with the frequency of the light because of effects (1) and (3).

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.

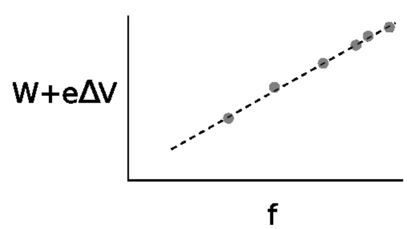


g / A different way of studying the photoelectric effect.

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:

$$E = hf \qquad ,$$

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

[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.

even know E_s in advance, and could only graph $e\Delta V$ versus f? ▷ Answer, p. 711

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.

Number of photons emitted by a lightbulb per second *example 7*

▷ 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

$$\begin{aligned} (100\ \text{J})/E_{photon} &= (100\ \text{J})/(hc/\lambda) \\ &\approx 3 \times 10^{20} \end{aligned}$$

Momentum of a photon *example 8*

▷ According to the theory of relativity, the momentum of a beam of light is given by $p = E/c$ (see homework problem 11 on page 342). 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

$$\begin{aligned} p &= E/c \\ &= \frac{h}{c} f \quad . \end{aligned}$$

To reexpress this in terms of wavelength, we use $c = f\lambda$:

$$\begin{aligned} p &= \frac{h}{c} \cdot \frac{c}{\lambda} \\ &= \frac{h}{\lambda} \end{aligned}$$

The second form turns out to be simpler.

Discussion Questions

A The photoelectric effect only ever ejects a very tiny percentage of the electrons available near the surface of an object. How well does this agree with the wave model of light, and how well with the particle model? Consider the two different distance scales involved: the wavelength of the light, and the size of an atom, which is on the order of 10^{-10} or 10^{-9} m.

B What is the significance of the fact that Planck's constant is numerically very small? How would our everyday experience of light be different if it was not so small?

C How would the experiments described above be affected if a single electron was likely to get hit by more than one photon?

D Draw some representative trajectories of electrons for $\Delta V = 0$, ΔV less than the maximum value, and ΔV greater than the maximum value.

E Explain based on the photon theory of light why ultraviolet light would be more likely than visible or infrared light to cause cancer by damaging DNA molecules. How does this relate to discussion question C?

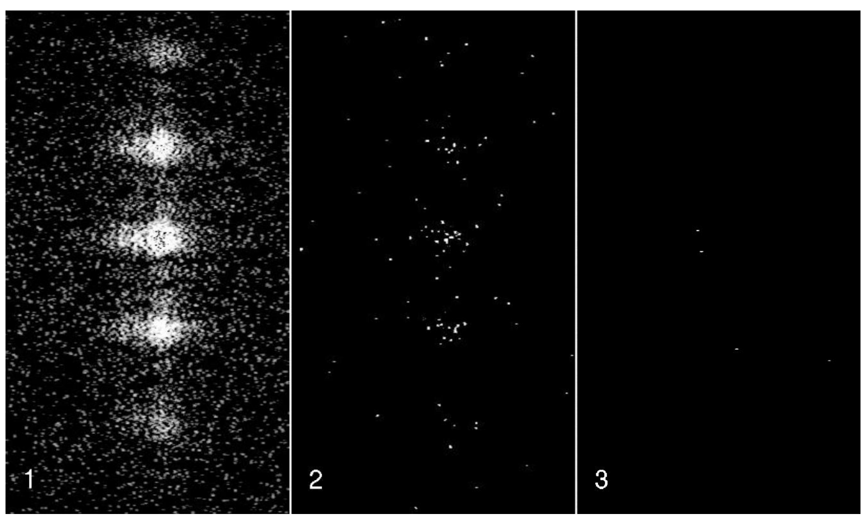
F Does $E = hf$ imply that a photon changes its energy when it passes from one transparent material into another substance with a different index of refraction?

12.2.3 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 page 631. 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 631 are simply cutouts from these figures.)

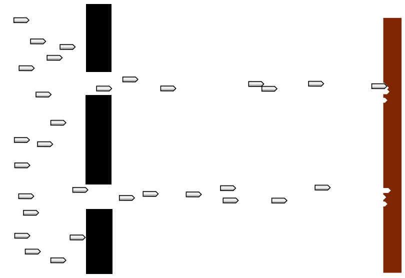


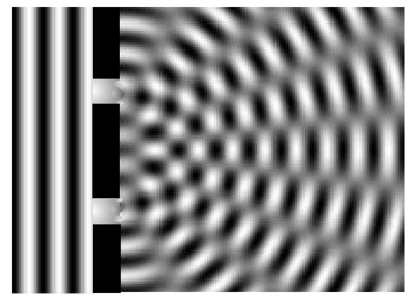
i / Wave interference patterns photographed by Prof. Lyman Page with a digital camera. Laser light with a single well-defined wavelength passed through a series of absorbers to cut down its intensity, then through a set of slits to produce interference, and finally into a digital camera chip. (A triple slit was actually used, but for conceptual simplicity we discuss the results in the main text as if it was a double slit.) In panel 2 the intensity has been reduced relative to 1, and even more so for panel 3.

Neither the pure wave theory nor the pure particle theory can explain the results. If light was only a particle and not a wave, there would be no interference effect. The result of the experiment would be like firing a hail of bullets through a double slit, j. Only two spots directly behind the slits would be hit.

If, on the other hand, light was only a wave and not a particle, we would get the same kind of diffraction pattern that would happen with a water wave, k. There would be no discrete dots in the photo, only a diffraction pattern that shaded smoothly between light and dark.

Applying the definitions to this experiment, light must be both a particle and a wave. It is a wave because it exhibits interference effects. At the same time, the fact that the photographs contain discrete dots is a direct demonstration that light refuses to be split into units of less than a single photon. There can only be whole numbers of photons: four photons in figure i/3, for example.


j / Bullets pass through a double slit.


k / A water wave passes through a double slit.


l / A single photon can go through both slits.

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.

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 12.3.4 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.

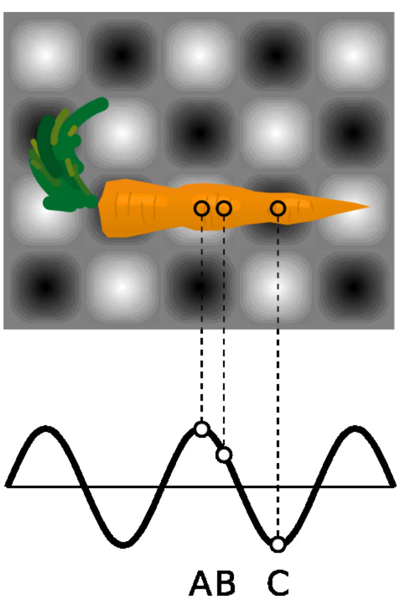
The probability interpretation

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,

$$(\text{probability distribution}) \propto (\text{amplitude})^2 \qquad .$$

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.



m / Example 9.

A microwave oven *example 9*

▷ The figure shows two-dimensional (top) and one-dimensional (bottom) representations of the standing wave inside a microwave oven. Gray represents zero field, and white and black signify the strongest fields, with white being a field that is in the opposite direction compared to black. Compare the probabilities of detecting a microwave photon at points A, B, and C.

▷ A and C are both extremes of the wave, so the probabilities of detecting a photon at A and C are equal. It doesn't matter that we have represented C as negative and A as positive, because it is the square of the amplitude that is relevant. The amplitude at B is about 1/2 as much as the others, so the probability of detecting a photon there is about 1/4 as much.

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 the following chapter, 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.

What is the proportionality constant? *example 10*

▷ What is the proportionality constant that would make an actual equation out of (probability distribution) $\propto$ (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:

$$P = \frac{\text{energy in volume } v}{\text{energy of photon}} = \frac{\text{energy in volume } v}{hf}$$

We assume v is small enough so that the electric and magnetic fields are nearly constant throughout it. We then have

$$P = \frac{\left(\frac{1}{8\pi k}|\mathbf{E}|^2 + \frac{1}{2\mu_o}|\mathbf{B}|^2\right) v}{hf} \qquad .$$

We can simplify this formidable looking expression by recognizing that in an electromagnetic wave, $|\mathbf{E}|$ and $|\mathbf{B}|$ are related by $|\mathbf{E}| = c|\mathbf{B}|$. With some algebra, it turns out that the electric and magnetic fields each contribute half the total energy, so we can simplify this to

$$P = 2\frac{\left(\frac{1}{8\pi k}|\mathbf{E}|^2\right) v}{hf} = \frac{v}{4\pi k hf}|\mathbf{E}|^2 \qquad .$$

As advertised, the probability is proportional to the square of the wave's amplitude.

Discussion Questions

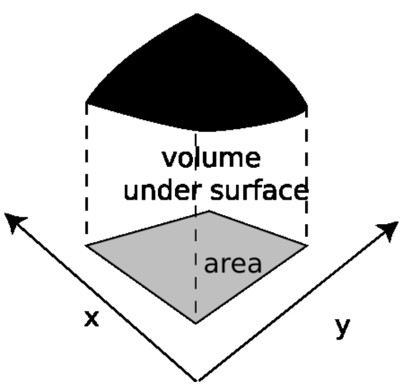
A Referring back to the example of the carrot in the microwave oven, show that it would be nonsensical to have probability be proportional to the field itself, rather than the square of the field.

B Einstein did not try to reconcile the wave and particle theories of light, and did not say much about their apparent inconsistency. Einstein basically visualized a beam of light as a stream of bullets coming from a machine gun. In the photoelectric effect, a photon "bullet" would only hit one atom, just as a real bullet would only hit one person. Suppose someone reading his 1905 paper wanted to interpret it by saying that Einstein's so-called particles of light are simply short wave-trains that only occupy a small region of space. Comparing the wavelength of visible light (a few hundred nm) to the size of an atom (on the order of 0.1 nm), explain why this poses a difficulty for reconciling the particle and wave theories.

C Can a white photon exist?

D In double-slit diffraction of photons, would you get the same pattern of dots on the digital camera image if you covered one slit? Why should it matter whether you give the photon two choices or only one?

12.2.4 Photons in three dimensions



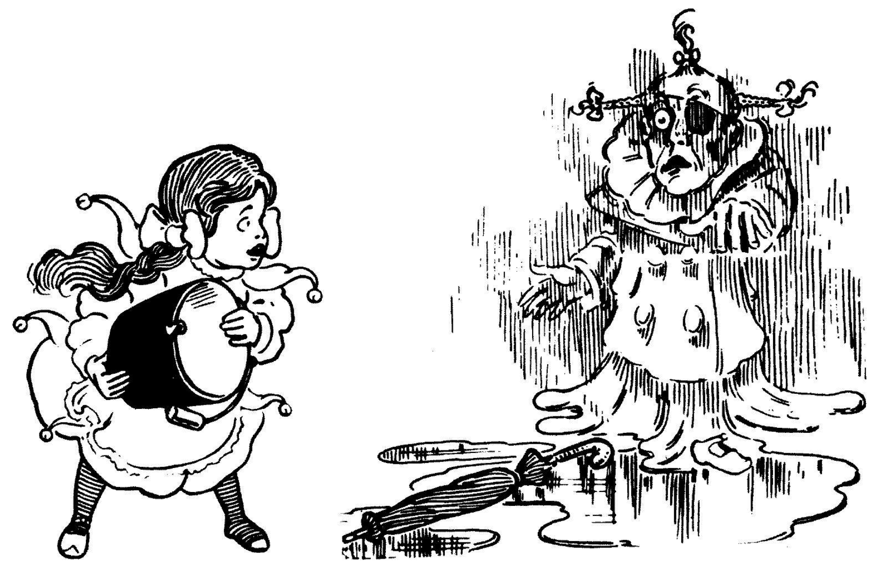
n / Probability is the volume under a surface defined by $D(x, y)$.

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 12.1 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: (area) = (height)(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: (probability) = (depth)(length)(width) = $\mathrm{m}^{-2} \cdot \mathrm{m} \cdot \mathrm{m}$. In terms of calculus, we have $P = \int D \, \mathrm{d}x \, \mathrm{d}y$.

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, $P = \int D \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z$.



12.3 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.

12.3.1 Electrons as waves

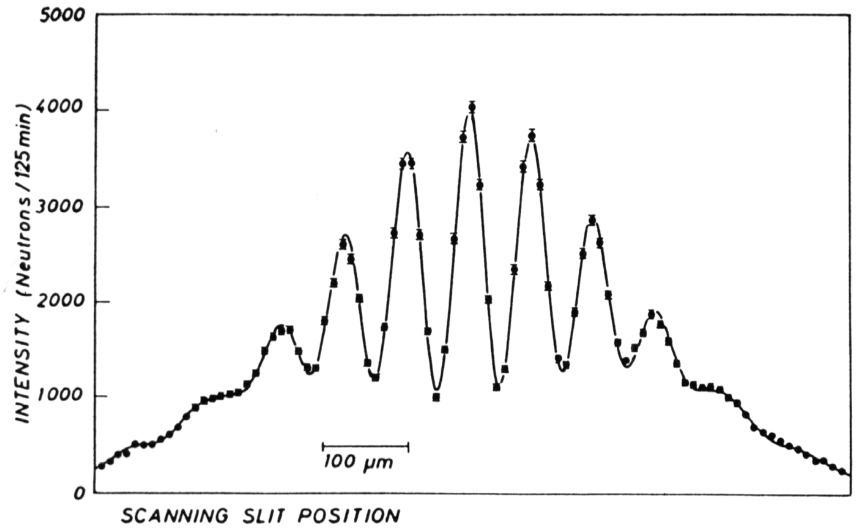
We started our journey into quantum physics by studying the random behavior of *matter* in radioactive decay, and then asked how randomness could be linked to the basic laws of nature governing *light*. The probability interpretation of wave-particle duality was strange and hard to accept, but it provided such a link. It is now natural to ask whether the same explanation could be applied to matter. If the fundamental building block of light, the photon, is a particle as well as a wave, is it possible that the basic units of matter, such as electrons, are waves as well as particles?

A young French aristocrat studying physics, Louis de Broglie (pronounced "broylee"), made exactly this suggestion in his 1923 Ph.D. thesis. His idea had seemed so farfetched that there was serious doubt about whether to grant him the degree. Einstein was asked for his opinion, and with his strong support, de Broglie got his degree.

Only two years later, American physicists C.J. Davisson and L. Germer confirmed de Broglie's idea by accident. They had been studying the scattering of electrons from the surface of a sample of nickel, made of many small crystals. (One can often see such a crystalline pattern on a brass doorknob that has been polished by repeated handling.) An accidental explosion occurred, and when they put their apparatus back together they observed something entirely different: the scattered electrons were now creating an interference pattern! This dramatic proof of the wave nature of matter came about because the nickel sample had been melted by the explosion and then resolidified as a single crystal. The nickel atoms, now nicely arranged in the regular rows and columns of a crystalline lattice, were acting as the lines of a diffraction grating. The new crystal was analogous to the type of ordinary diffraction grating in which the lines are etched on the surface of a mirror (a reflection grating) rather than the kind in which the light passes through the transparent gaps between the lines (a transmission grating).

Although we will concentrate on the wave-particle duality of electrons because it is important in chemistry and the physics of atoms, all the other "particles" of matter you've learned about show wave properties as well. Figure a, for instance, shows a wave interference pattern of neutrons.

It might seem as though all our work was already done for us, and there would be nothing new to understand about electrons: they have the same kind of funny wave-particle duality as photons. That's almost true, but not quite. There are some important ways in which electrons differ significantly from photons:



a / A double-slit interference pattern made with neutrons. (A. Zeilinger, R. Gähler, C.G. Shull, W. Treimer, and W. Mampe, *Reviews of Modern Physics*, Vol. 60, 1988.)

1. Electrons have mass, and photons don't.

2. Photons always move at the speed of light, but electrons can move at any speed less than c.

3. Photons don't have electric charge, but electrons do, so electric forces can act on them. The most important example is the atom, in which the electrons are held by the electric force of the nucleus.

4. Electrons cannot be absorbed or emitted as photons are. Destroying an electron or creating one out of nothing would violate conservation of charge.

(In section 12.4 we will learn of one more fundamental way in which electrons differ from photons, for a total of five.)

Because electrons are different from photons, it is not immediately obvious which of the photon equations from the previous chapter can be applied to electrons as well. A particle property, the energy of one photon, is related to its wave properties via $E = hf$ or, equivalently, $E = hc/\lambda$. The momentum of a photon was given by $p = hf/c$ or $p = h/\lambda$. Ultimately it was a matter of experiment to determine which of these equations, if any, would work for electrons, but we can make a quick and dirty guess simply by noting that some of the equations involve c, the speed of light, and some do not. Since c is irrelevant in the case of an electron, we might guess that the equations of general validity are those that do not have c

in them:

$$E = hf$$
$$p = h/\lambda$$

This is essentially the reasoning that de Broglie went through, and experiments have confirmed these two equations for all the fundamental building blocks of light and matter, not just for photons and electrons.

The second equation, which I soft-pedaled in the previous chapter, takes on a greater important for electrons. This is first of all because the momentum of matter is more likely to be significant than the momentum of light under ordinary conditions, and also because force is the transfer of momentum, and electrons are affected by electrical forces.

The wavelength of an elephant *example 11*

▷ What is the wavelength of a trotting elephant?

▷ One may doubt whether the equation should be applied to an elephant, which is not just a single particle but a rather large collection of them. Throwing caution to the wind, however, we estimate the elephant's mass at 10^3 kg and its trotting speed at 10 m/s. Its wavelength is therefore roughly

$$\begin{aligned} \lambda &= \frac{h}{p} \\ &= \frac{h}{mv} \\ &= \frac{6.63 \times 10^{-34}\ \text{J}\cdot\text{s}}{(10^3\ \text{kg})(10\ \text{m/s})} \\ &\sim 10^{-37}\ \frac{(\text{kg}\cdot\text{m}^2/\text{s}^2)\cdot\text{s}}{\text{kg}\cdot\text{m/s}} \\ &= 10^{-37}\ \text{m} \end{aligned}$$

The wavelength found in this example is so fantastically small that we can be sure we will never observe any measurable wave phenomena with elephants or any other human-scale objects. The result is numerically small because Planck's constant is so small, and as in some examples encountered previously, this smallness is in accord with the correspondence principle.

Although a smaller mass in the equation $\lambda = h/mv$ does result in a longer wavelength, the wavelength is still quite short even for individual electrons under typical conditions, as shown in the following example.

The typical wavelength of an electron *example 12*

▷ Electrons in circuits and in atoms are typically moving through voltage differences on the order of 1 V, so that a typical energy is $(e)(1\ \text{V})$, which is on the order of 10^{-19} J. What is the wavelength of an electron with this amount of kinetic energy?

▹ This energy is nonrelativistic, since it is much less than mc^2. Momentum and energy are therefore related by the nonrelativistic equation $K = p^2/2m$. Solving for p and substituting in to the equation for the wavelength, we find

$$\lambda = \frac{h}{\sqrt{2mK}}$$
$$= 1.6 \times 10^{-9}\ \text{m} \qquad .$$

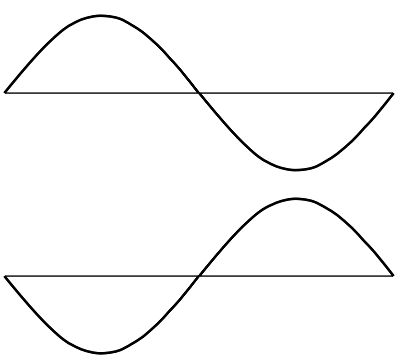
This is on the same order of magnitude as the size of an atom, which is no accident: as we will discuss in the next chapter in more detail, an electron in an atom can be interpreted as a standing wave. The smallness of the wavelength of a typical electron also helps to explain why the wave nature of electrons wasn't discovered until a hundred years after the wave nature of light. To scale the usual wave-optics devices such as diffraction gratings down to the size needed to work with electrons at ordinary energies, we need to make them so small that their parts are comparable in size to individual atoms. This is essentially what Davisson and Germer did with their nickel crystal.

Self-Check

These remarks about the inconvenient smallness of electron wavelengths apply only under the assumption that the electrons have typical energies. What kind of energy would an electron have to have in order to have a longer wavelength that might be more convenient to work with?
▹ Answer, p. 711

What kind of wave is it?

If a sound wave is a vibration of matter, and a photon is a vibration of electric and magnetic fields, what kind of a wave is an electron made of? The disconcerting answer is that there is no experimental "observable," i.e. directly measurable quantity, to correspond to the electron wave itself. In other words, there are devices like microphones that detect the oscillations of air pressure in a sound wave, and devices such as radio receivers that measure the oscillation of the electric and magnetic fields in a light wave, but nobody has ever found any way to measure the electron wave directly.



b / These two electron waves are not distinguishable by any measuring device.

We can of course detect the energy (or momentum) possessed by an electron just as we could detect the energy of a photon using a digital camera. (In fact I'd imagine that an unmodified digital camera chip placed in a vacuum chamber would detect electrons just as handily as photons.) But this only allows us to determine where the wave carries high probability and where it carries low probability. Probability is proportional to the square of the wave's amplitude, but measuring its square is not the same as measuring the wave itself. In particular, we get the same result by squaring either a positive number or its negative, so there is no way to determine the positive or negative sign of an electron wave.

Most physicists tend toward the school of philosophy known as operationalism, which says that a concept is only meaningful if we can define some set of operations for observing, measuring, or testing it. According to a strict operationalist, then, the electron wave itself is a meaningless concept. Nevertheless, it turns out to be one of those concepts like love or humor that is impossible to measure and yet very useful to have around. We therefore give it a symbol, Ψ (the capital Greek letter psi), and a special name, the electron *wavefunction* (because it is a function of the coordinates x, y, and z that specify where you are in space). It would be impossible, for example, to calculate the shape of the electron wave in a hydrogen atom without having some symbol for the wave. But when the calculation produces a result that can be compared directly to experiment, the final algebraic result will turn out to involve only Ψ^2, which is what is observable, not Ψ itself.

Since Ψ, unlike E and B, is not directly measurable, we are free to make the probability equations have a simple form: instead of having the probability density equal to some funny constant multiplied by Ψ^2, we simply define Ψ so that the constant of proportionality is one:

$$\text{(probability distribution)} = \Psi^2 \qquad .$$

Since the probability distribution has units of m^{-3}, the units of Ψ must be $\mathrm{m}^{-3/2}$.

Discussion Question

A Frequency is oscillations per second, whereas wavelength is meters per oscillation. How could the equations $E = hf$ and $p = h/\lambda$ be made to look more alike by using quantities that were more closely analogous? (This more symmetric treatment makes it easier to incorporate relativity into quantum mechanics, since relativity says that space and time are not entirely separate.)

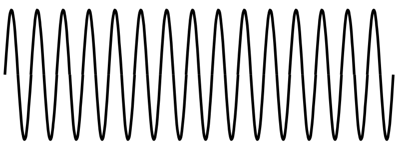
12.3.2 Dispersive waves

A colleague of mine who teaches chemistry loves to tell the story about an exceptionally bright student who, when told of the equation $p = h/\lambda$, protested, "But when I derived it, it had a factor of 2!" The issue that's involved is a real one, albeit one that could be glossed over (and is, in most textbooks) without raising any alarms in the mind of the average student. The present optional section addresses this point; it is intended for the student who wishes to delve a little deeper.

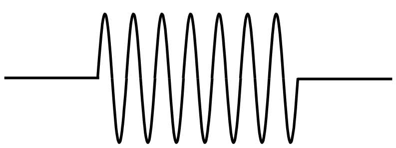
Here's how the now-legendary student was presumably reasoning. We start with the equation $v = f\lambda$, which is valid for any sine wave, whether it's quantum or classical. Let's assume we already know $E = hf$, and are trying to derive the relationship between wavelength and momentum:

$$\begin{aligned}\lambda &= \frac{v}{f} \\ &= \frac{vh}{E} \\ &= \frac{vh}{\frac{1}{2}mv^2} \\ &= \frac{2h}{mv} \\ &= \frac{2h}{p} \qquad .\end{aligned}$$

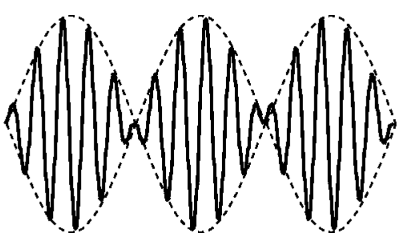
The reasoning seems valid, but the result does contradict the accepted one, which is after all solidly based on experiment.



c / Part of an infinite sine wave.



d / A finite-length sine wave.



e / A beat pattern created by superimposing two sine waves with slightly different wavelengths.

The mistaken assumption is that we can figure everything out in terms of pure sine waves. Mathematically, the only wave that has a perfectly well defined wavelength and frequency is a sine wave, and not just any sine wave but an infinitely long sine wave, c. The unphysical thing about such a wave is that it has no leading or trailing edge, so it can never be said to enter or leave any particular region of space. Our derivation made use of the velocity, v, and if velocity is to be a meaningful concept, it must tell us how quickly stuff (mass, energy, momentum, . . .) is transported from one region of space to another. Since an infinitely long sine wave doesn't remove any stuff from one region and take it to another, the "velocity of its stuff" is not a well defined concept.

Of course the individual wave peaks do travel through space, and one might think that it would make sense to associate their speed with the "speed of stuff," but as we will see, the two velocities are in general unequal when a wave's velocity depends on wavelength. Such a wave is called a *dispersive* wave, because a wave pulse consisting of a superposition of waves of different wavelengths will separate (disperse) into its separate wavelengths as the waves move through space at different speeds. Nearly all the waves we have encountered have been nondispersive. For instance, sound waves and light waves (in a vacuum) have speeds independent of wavelength. A water wave is one good example of a dispersive wave. Long-wavelength water waves travel faster, so a ship at sea that encounters a storm typically sees the long-wavelength parts of the wave first. When dealing with dispersive waves, we need symbols and words to distinguish the two speeds. The speed at which wave peaks move is called the phase velocity, v_p, and the speed at which "stuff" moves is called the group velocity, v_g.

An infinite sine wave can only tell us about the phase velocity, not the group velocity, which is really what we would be talking about when we refer to the speed of an electron. If an infinite sine wave is the simplest possible wave, what's the next best thing?

We might think the runner up in simplicity would be a wave train consisting of a chopped-off segment of a sine wave, d. However, this kind of wave has kinks in it at the end. A simple wave should be one that we can build by superposing a small number of infinite sine waves, but a kink can never be produced by superposing any number of infinitely long sine waves.

Actually the simplest wave that transports stuff from place to place is the pattern shown in figure e. Called a beat pattern, it is formed by superposing two sine waves whose wavelengths are similar but not quite the same. If you have ever heard the pulsating howling sound of musicians in the process of tuning their instruments to each other, you have heard a beat pattern. The beat pattern gets stronger and weaker as the two sine waves go in and out of phase with each other. The beat pattern has more "stuff" (energy, for example) in the areas where constructive interference occurs, and less in the regions of cancellation. As the whole pattern moves through space, stuff is transported from some regions and into other ones.

If the frequency of the two sine waves differs by 10%, for instance, then ten periods will be occur between times when they are in phase. Another way of saying it is that the sinusoidal "envelope" (the dashed lines in figure e) has a frequency equal to the difference in frequency between the two waves. For instance, if the waves had frequencies of 100 Hz and 110 Hz, the frequency of the envelope would be 10 Hz.

To apply similar reasoning to the wavelength, we must define a quantity $z = 1/\lambda$ that relates to wavelength in the same way that frequency relates to period. In terms of this new variable, the z of the envelope equals the difference between the $z's$ of the two sine waves.

The group velocity is the speed at which the envelope moves through space. Let Δf and Δz be the differences between the frequencies and $z's$ of the two sine waves, which means that they equal the frequency and z of the envelope. The group velocity is $v_g = f_{envelope}\lambda_{envelope} = \Delta f/\Delta \text{z}$. If Δf and Δz are sufficiently small, we can approximate this expression as a derivative,

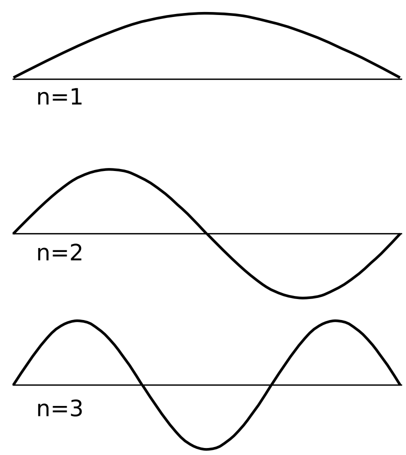
$$v_g = \frac{\mathrm{d}f}{\mathrm{d}z} \qquad .$$

This expression is usually taken as the definition of the group velocity for wave patterns that consist of a superposition of sine waves having a narrow range of frequencies and wavelengths. In quantum mechanics, with $f = E/h$ and $z = p/h$, we have $v_g = \mathrm{d}E/\mathrm{d}p$. In the case of a nonrelativistic electron the relationship between energy and momentum is $E = p^2/2m$, so the group velocity is $\mathrm{d}E/\mathrm{d}p = p/m = v$, exactly what it should be. It is only the phase velocity that differs by a factor of two from what we would have expected, but the phase velocity is not the physically important thing.

12.3.3 Bound states

Electrons are at their most interesting when they're in atoms, that is, when they are bound within a small region of space. We can understand a great deal about atoms and molecules based on simple arguments about such bound states, without going into any of the realistic details of atom. The simplest model of a bound state is known as the particle in a box: like a ball on a pool table, the electron feels zero force while in the interior, but when it reaches an edge it encounters a wall that pushes back inward on it with a large force. In particle language, we would describe the electron as bouncing off of the wall, but this incorrectly assumes that the electron has a certain path through space. It is more correct to describe the electron as a wave that undergoes 100% reflection at the boundaries of the box.

Like a generation of physics students before me, I rolled my eyes when initially introduced to the unrealistic idea of putting a particle in a box. It seemed completely impractical, an artificial textbook invention. Today, however, it has become routine to study electrons in rectangular boxes in actual laboratory experiments. The "box" is actually just an empty cavity within a solid piece of silicon, amounting in volume to a few hundred atoms. The methods for creating these electron-in-a-box setups (known as "quantum dots") were a by-product of the development of technologies for fabricating computer chips.



f / Three possible standing-wave patterns for a particle in a box.

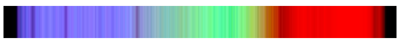
For simplicity let's imagine a one-dimensional electron in a box, i.e. we assume that the electron is only free to move along a line. The resulting standing wave patterns, of which the first three are shown in the figure, are just like some of the patterns we encountered with sound waves in musical instruments. The wave patterns must be zero at the ends of the box, because we are assuming the walls are impenetrable, and there should therefore be zero probability of finding the electron outside the box. Each wave pattern is labeled according to n, the number of peaks and valleys it has. In quantum physics, these wave patterns are referred to as "states" of the particle-in-the-box system.

The following seemingly innocuous observations about the particle in the box lead us directly to the solutions to some of the most vexing failures of classical physics:

The particle's energy is quantized (can only have certain values). Each wavelength corresponds to a certain momentum, and a given momentum implies a definite kinetic energy, $E = p^2/2m$. (This is the second type of energy quantization we have encountered. The type we studied previously had to do with restricting the number of particles to a whole number, while assuming some specific wavelength and energy for each particle. This type of quantization refers to the energies that a single particle can have. Both photons and matter particles demonstrate both types of quantization under the appropriate circumstances.)

The particle has a minimum kinetic energy. Long wavelengths correspond to low momenta and low energies. There can be no state with an energy lower than that of the $n = 1$ state, called the ground state.

The smaller the space in which the particle is confined, the higher its kinetic energy must be. Again, this is because long wavelengths give lower energies.



g / The spectrum of the light from the star Sirius.

Spectra of thin gases *example 13*

A fact that was inexplicable by classical physics was that thin gases absorb and emit light only at certain wavelengths. This was observed both in earthbound laboratories and in the spectra of stars. The figure on the left shows the example of the spectrum of the star Sirius, in which there are "gap teeth" at certain wavelengths. Taking this spectrum as an example, we can give a straightforward explanation using quantum physics.

Energy is released in the dense interior of the star, but the outer layers of the star are thin, so the atoms are far apart and electrons are confined within individual atoms. Although their standing-wave patterns are not as simple as those of the particle in the box, their energies are quantized.

When a photon is on its way out through the outer layers, it can be absorbed by an electron in an atom, but only if the amount of energy it carries happens to be the right amount to kick the electron from one of the allowed energy levels to one of the higher levels. The photon energies that are missing from the spectrum are the ones that equal the difference in energy between two electron energy levels. (The most prominent of the absorption lines in Sirius's spectrum are absorption lines of the hydrogen atom.)

The stability of atoms *example 14*

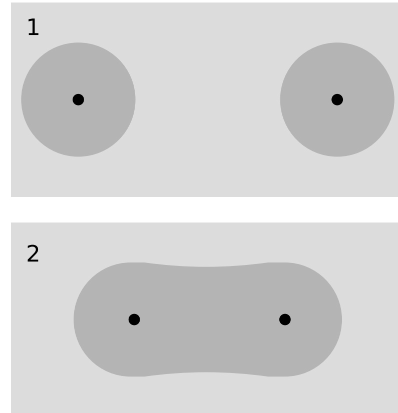
In many Star Trek episodes the Enterprise, in orbit around a planet, suddenly lost engine power and began spiraling down toward the planet's surface. This was utter nonsense, of course, due to conservation of energy: the ship had no way of getting rid of energy, so it did not need the engines to replenish it.

Consider, however, the electron in an atom as it orbits the nucleus. The electron *does* have a way to release energy: it has an acceleration due to its continuously changing direction of motion, and according to classical physics, any accelerating charged particle emits electromagnetic waves. According to classical physics, atoms should collapse!

The solution lies in the observation that a bound state has a minimum energy. An electron in one of the higher-energy atomic states can and does emit photons and hop down step by step in energy. But once it is in the ground state, it cannot emit a photon because there is no lower-energy state for it to go to.

Chemical bonds *example 15*

I began this section with a classical argument that chemical bonds, as in an H_2 molecule, should not exist. Quantum physics explains why this type of bonding does in fact occur. When the atoms are next to each other, the electrons are shared between them. The "box" is about twice as wide, and a larger box allows a smaller energy. Energy is required in order to separate the atoms. (A qualitatively different type of bonding is discussed on page 679.)



h / Two hydrogen atoms bond to form an H_2 molecule. In the molecule, the two electrons' wave patterns overlap , and are about twice as wide.

Discussion Questions

A Neutrons attract each other via the strong nuclear force, so according to classical physics it should be possible to form nuclei out of clusters of two or more neutrons, with no protons at all. Experimental searches, however, have failed to turn up evidence of a stable two-neutron system (dineutron) or larger stable clusters. Explain based on quantum physics why a dineutron might spontaneously fly apart.

B The following table shows the energy gap between the ground state and the first excited state for four nuclei, in units of picojoules. (The nuclei were chosen to be ones that have similar structures, e.g., they are all spherical in shape.)

nucleus	energy gap (picojoules)
^{4}He	3.234
^{16}O	0.968
^{40}Ca	0.536
^{208}Pb	0.418

Explain the trend in the data.

12.3.4 The uncertainty principle and measurement

Eliminating randomness through measurement?

A common reaction to quantum physics, among both early-twentieth-century physicists and modern students, is that we should be able to get rid of randomness through accurate measurement. If I say, for example, that it is meaningless to discuss the path of a photon or an electron, one might suggest that we simply measure the particle's position and velocity many times in a row. This series of snapshots would amount to a description of its path.

A practical objection to this plan is that the process of measurement will have an effect on the thing we are trying to measure. This may not be of much concern, for example, when a traffic cop measure's your car's motion with a radar gun, because the energy and momentum of the radar pulses are insufficient to change the car's motion significantly. But on the subatomic scale it is a very real problem. Making a videotape through a microscope of an electron orbiting a nucleus is not just difficult, it is theoretically impossible. The video camera makes pictures of things using light that has bounced off them and come into the camera. If even a single photon of visible light was to bounce off of the electron we were trying to study, the electron's recoil would be enough to change its behavior significantly.

The Heisenberg uncertainty principle

This insight, that measurement changes the thing being measured, is the kind of idea that clove-cigarette-smoking intellectuals outside of the physical sciences like to claim they knew all along. If only, they say, the physicists had made more of a habit of reading literary journals, they could have saved a lot of work. The anthropologist Margaret Mead has recently been accused of inadvertently encouraging her teenaged Samoan informants to exaggerate the freedom of youthful sexual experimentation in their society. If this is considered a damning critique of her work, it is because she could have done better: other anthropologists claim to have been able to eliminate the observer-as-participant problem and collect untainted data.

The German physicist Werner Heisenberg, however, showed that in quantum physics, *any* measuring technique runs into a brick wall when we try to improve its accuracy beyond a certain point. Heisenberg showed that the limitation is a question of *what there is to be known*, even in principle, about the system itself, not of the ability or inability of a specific measuring device to ferret out information that is knowable but not previously hidden.

Suppose, for example, that we have constructed an electron in a box (quantum dot) setup in our laboratory, and we are able adjust the length L of the box as desired. All the standing wave patterns pretty much fill the box, so our knowledge of the electron's position is of limited accuracy. If we write Δx for the range of uncertainty in our knowledge of its position, then Δx is roughly the same as the length of the box:

$$\Delta x \approx L$$

If we wish to know its position more accurately, we can certainly squeeze it into a smaller space by reducing L, but this has an unintended side-effect. A standing wave is really a superposition of two traveling waves going in opposite directions. The equation $p = h/\lambda$ really only gives the magnitude of the momentum vector, not its direction, so we should really interpret the wave as a 50/50 mixture of a right-going wave with momentum $p = h/\lambda$ and a left-going one with momentum $p = -h/\lambda$. The uncertainty in our knowledge of the electron's momentum is $\Delta p = 2h/\lambda$, covering the range between these two values. Even if we make sure the electron is in the ground state, whose wavelength $\lambda = 2L$ is the longest possible, we have an uncertainty in momentum of $\Delta p = h/L$. In general, we find

$$\Delta p \gtrsim h/L \qquad ,$$

with equality for the ground state and inequality for the higher-energy states. Thus if we reduce L to improve our knowledge of the electron's position, we do so at the cost of knowing less about its momentum. This trade-off is neatly summarized by multiplying the two equations to give

$$\Delta p \Delta x \gtrsim h \qquad .$$

Although we have derived this in the special case of a particle in a box, it is an example of a principle of more general validity:

The Heisenberg uncertainty principle
It is not possible, even in principle, to know the momentum and the position of a particle simultaneously and with perfect accuracy. The uncertainties in these two quantities are always such that $\Delta p \Delta x \gtrsim h$.

(This approximation can be made into a strict inequality, $\Delta p \Delta x > h/4\pi$, but only with more careful definitions, which we will not bother with.)

Note that although I encouraged you to think of this derivation in terms of a specific real-world system, the quantum dot, no reference was ever made to any specific laboratory equipment or procedures. The argument is simply that we cannot *know* the particle's position very accurately unless it *has* a very well defined position, it cannot have a very well defined position unless its wave-pattern covers only a very small amount of space, and its wave-pattern cannot be thus compressed without giving it a short wavelength and a correspondingly uncertain momentum. The uncertainty principle is therefore a restriction on how much there is to know about a particle, not just on what we can know about it with a certain technique.

An estimate for electrons in atoms *example 16*

▷ A typical energy for an electron in an atom is on the order of (1 volt)$\cdot e$, which corresponds to a speed of about 1% of the speed of light. If a typical atom has a size on the order of 0.1 nm, how close are the electrons to the limit imposed by the uncertainty principle?

▷ If we assume the electron moves in all directions with equal probability, the uncertainty in its momentum is roughly twice its typical momentum. This only an order-of-magnitude estimate, so we take Δp to be the same as a typical momentum:

$$\begin{aligned}\Delta p\Delta x &= p_{typical}\Delta x \\ &= (m_{electron})(0.01c)(0.1\times10^{-9}\text{ m}) \\ &= 3\times10^{-34}\text{ J}\cdot\text{s}\end{aligned}$$

This is on the same order of magnitude as Planck's constant, so evidently the electron is "right up against the wall." (The fact that it is somewhat less than h is of no concern since this was only an estimate, and we have not stated the uncertainty principle in its most exact form.)

Self-Check

If we were to apply the uncertainty principle to human-scale objects, what would be the significance of the small numerical value of Planck's constant? ▷ Answer, p. 711

Historical Note

The true nature of Heisenberg's role in the Nazi atomic bomb effort is a fascinating question, and dramatic enough to have inspired a well-received 1998 theatrical play, "Copenhagen." The real story, however, may never be completely unraveled. Heisenberg was the scientific leader of the German bomb program up until its cancellation in 1942, when the German military decided that it was too ambitious a project to undertake in wartime, and too unlikely to produce results.

Some historians believe that Heisenberg intentionally delayed and obstructed the project because he secretly did not want the Nazis to get the bomb. Heisenberg's apologists point out that he never joined the Nazi party, and was not anti-Semitic. He actively resisted the government's *Deutsche-Physik* policy of eliminating supposed Jewish influences from physics, and as a result was denounced by the S.S. as a traitor, escaping punishment only because Himmler personally declared him innocent. One strong piece of evidence for this view is a secret message carried to the U.S. in 1941, by one of the last Jews to escape from Berlin, and eventually delivered to the chairman of the Uranium Committee, which was then studying the feasibility of a bomb. The message states "...that a large number of German physicists are working intensively on the problem of the uranium bomb under the direction of Heisenberg, [and] that Heisenberg himself tries to delay the work as much as possible, fearing the catastrophic results of success. But he cannot help fulfilling the orders given to him, and if the problem can be solved, it will be solved probably in the near future. So he gave the advice to us to hurry up if U.S.A. will not come too late." The message supports the view that Heisenberg intentionally misled his government about the bomb's technical feasibility; Nazi Minister of Armaments Albert Speer wrote that he was convinced to drop the project after a 1942 meeting with Heisenber because "the physicists themselves didn't want to put too much into it." Heisenberg also may have warned Danish physicist Niels Bohr personally in September 1941 about the existence of the Nazi bomb effort.

On the other side of the debate, critics of Heisenberg say that he clearly wanted Germany to win the war, that he visited German-occupied territories in a semi-official role, and that he may simply have not been very good at his job directing the bomb project. On a visit to the occupied Netherlands in 1943, he told a colleague, "Democracy cannot develop sufficient energy to rule Europe. There are, therefore, only two alternatives: Germany and Russia. And then a Europe under German leadership would be the lesser evil." Some historians[3] argue that the real point of Heisenberg's meeting with Bohr was to try to convince the U.S. not to try to build a bomb, so that Germany, possessing a nuclear monopoly, would defeat the Soviets — this was after the June 1941 entry of the U.S.S.R. into the war, but before the December 1941 Pearl Harbor attack brought the U.S. in. Bohr apparently considered Heisenberg's account of the meeting, published after the war was over, to be inaccurate.[4] The secret 1941 message also has a curious moral passivity to it, as if Heisenberg was saying "I hope you stop me before I do something bad," but we should also consider the great risk Heisenberg would have been running if he actually originated the message.

Measurement and Schrödinger's cat

In the previous chapter I briefly mentioned an issue concerning measurement that we are now ready to address carefully. If you hang around a laboratory where quantum-physics experiments are being done and secretly record the physicists' conversations, you'll hear them say many things that assume the probability interpretation of quantum mechanics. Usually they will speak as though the randomness of quantum mechanics enters the picture when something is measured. In the digital camera experiments of the previous chapter, for example, they would casually describe the detection of a photon at one of the pixels as if the moment of detection was when the photon was forced to "make up its mind." Although this mental cartoon usually works fairly well as a description of things they experience in the lab, it cannot ultimately be correct, because it attributes a special role to measurement, which is really just a physical process like all other physical processes.

If we are to find an interpretation that avoids giving any spe-

[3]A Historical Perspective on Copenhagen, David C. Cassidy, Physics Today, July 2000, p. 28, http://www.aip.org/pt/vol-53/iss-7/p28.html

[4]Bohr drafted several replies, but never published them for fear of hurting Heisenberg and his family. Bohr's papers were to be sealed for 50 years after his death, but his family released them early, in February 2002. The texts, including English translations, are available at http://www.nbi.dk/NBA/papers/docs/cover.html.

cial role to measurement processes, then we must think of the entire laboratory, including the measuring devices and the physicists themselves, as one big quantum-mechanical system made out of protons, neutrons, electrons, and photons. In other words, we should take quantum physics seriously as a description not just of microscopic objects like atoms but of human-scale ("macroscopic") things like the apparatus, the furniture, and the people.

The most celebrated example is called the Schrödinger's cat experiment. Luckily for the cat, there probably was no actual experiment — it was simply a "thought experiment" that the physicist the German theorist Schrödinger discussed with his colleagues. Schrödinger wrote:

> One can even construct quite burlesque cases. A cat is shut up in a steel container, together with the following diabolical apparatus (which one must keep out of the direct clutches of the cat): In a Geiger tube [radiation detector] there is a tiny mass of radioactive substance, so little that in the course of an hour perhaps one atom of it disintegrates, but also with equal probability not even one; if it does happen, the counter [detector] responds and ... activates a hammer that shatters a little flask of prussic acid [filling the chamber with poison gas]. If one has left this entire system to itself for an hour, then one will say to himself that the cat is still living, if in that time no atom has disintegrated. The first atomic disintegration would have poisoned it.

Now comes the strange part. Quantum mechanics describes the particles the cat is made of as having wave properties, including the property of superposition. Schrödinger describes the wavefunction of the box's contents at the end of the hour:

The wavefunction of the entire system would express this situation by having the living and the dead cat mixed ... in equal parts [50/50 proportions]. The uncertainty originally restricted to the atomic domain has been transformed into a macroscopic uncertainty...

At first Schrödinger's description seems like nonsense. When you opened the box, would you see two ghostlike cats, as in a doubly exposed photograph, one dead and one alive? Obviously not. You would have a single, fully material cat, which would either be dead or very, very upset. But Schrödinger has an equally strange and logical answer for that objection. In the same way that the quantum randomness of the radioactive atom spread to the cat and made its wavefunction a random mixture of life and death, the randomness spreads wider once you open the box, and your own wavefunction becomes a mixture of a person who has just killed a cat and a person who hasn't.

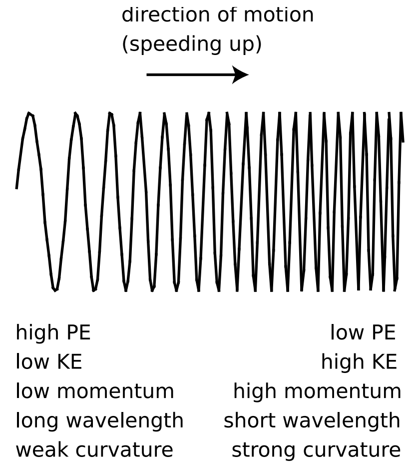
Discussion Questions

A Compare Δp and Δx for the two lowest energy levels of the one-dimensional particle in a box, and discuss how this relates to the uncertainty principle.

B On a graph of Δp versus Δx, sketch the regions that are allowed and forbidden by the Heisenberg uncertainty principle. Interpret the graph: Where does an atom lie on it? An elephant? Can either *p* or *x* be measured with perfect accuracy if we don't care about the other?

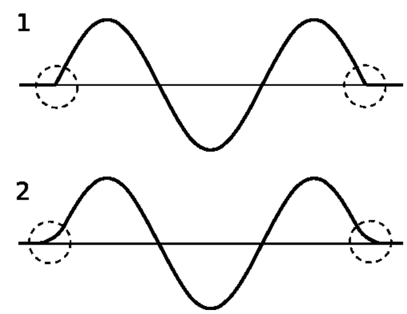
12.3.5 Electrons in electric fields

So far the only electron wave patterns we've considered have been simple sine waves, but whenever an electron finds itself in an electric field, it must have a more complicated wave pattern. Let's consider the example of an electron being accelerated by the electron gun at the back of a TV tube. Newton's laws are not useful, because they implicitly assume that the path taken by the particle is a meaningful concept. Conservation of energy is still valid in quantum physics, however. In terms of energy, the electron is moving from a region of low voltage into a region of higher voltage. Since its charge is negative, it loses electrical energy by moving to a higher voltage, so its kinetic energy increases. As its electrical energy goes down, its kinetic energy goes up by an equal amount, keeping the total energy constant. Increasing kinetic energy implies a growing momentum, and therefore a shortening wavelength, i



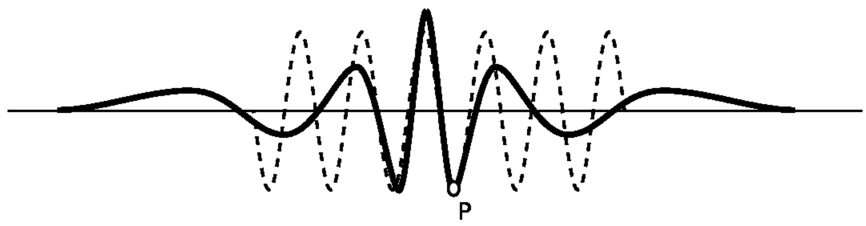
i / An electron in a gentle electric field gradually shortens its wavelength as it gains energy.

The wavefunction as a whole does not have a single well-defined wavelength, but the wave changes so gradually that if you only look at a small part of it you can still pick out a wavelength and relate it to the momentum and energy. (The picture actually exaggerates by many orders of magnitude the rate at which the wavelength changes.)



j / The wavefunction's tails go where classical physics says they shouldn't.

But what if the electric field was stronger? The electric field in a TV is only $\sim 10^5$ N/C, but the electric field within an atom is more like 10^{12} N/C. In figure k, the wavelength changes so rapidly that there is nothing that looks like a sine wave at all. We could get a rough idea of the wavelength in a given region by measuring the distance between two peaks, but that would only be a rough approximation. Suppose we want to know the wavelength at point P. The trick is to construct a sine wave, like the one shown with the dashed line, which matches the curvature of the actual wavefunction as closely as possible near P. The sine wave that matches as well as possible is called the "osculating" curve, from a Latin word meaning "to kiss." The wavelength of the osculating curve is the wavelength that will relate correctly to conservation of energy.



k / A typical wavefunction of an electron in an atom (heavy curve) and the osculating sine wave (dashed curve) that matches its curvature at point P.

Tunneling

We implicitly assumed that the particle-in-a-box wavefunction would cut off abruptly at the sides of the box, j/1, but that would be unphysical. A kink has infinite curvature, and curvature is related to energy, so it can't be infinite. A physically realistic wavefunction must always "tail off" gradually, j/2. In classical physics, a particle can never enter a region in which its interaction energy U would be greater than the amount of energy it has available. But in quantum physics the wavefunction will always have a tail that reaches into the classically forbidden region. If it was not for this effect, called tunneling, the fusion reactions that power the sun would not occur due to the high electrical energy nuclei need in order to get close together! Tunneling is discussed in more detail in the following subsection.

12.3.6 The Schrödinger equation

In subsection 12.3.5 we were able to apply conservation of energy to an electron's wavefunction, but only by using the clumsy graphical technique of osculating sine waves as a measure of the wave's curvature. You have learned a more convenient measure of curvature in calculus: the second derivative. To relate the two approaches, we take the second derivative of a sine wave:

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\sin(2\pi x/\lambda) &= \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{2\pi}{\lambda}\cos\frac{2\pi x}{\lambda}\right)\\ &= -\left(\frac{2\pi}{\lambda}\right)^2\sin\frac{2\pi x}{\lambda}\end{aligned}$$

Taking the second derivative gives us back the same function, but with a minus sign and a constant out in front that is related to the wavelength. We can thus relate the second derivative to the osculating wavelength:

$$[1] \qquad \frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} = -\left(\frac{2\pi}{\lambda}\right)^2\Psi$$

This could be solved for λ in terms of Ψ, but it will turn out below to be more convenient to leave it in this form.

Applying this to conservation of energy, we have

$$\begin{aligned} E &= K + U \\ &= \frac{p^2}{2m} + U \\ &= \frac{(h/\lambda)^2}{2m} + U \end{aligned} \qquad [2]$$

Note that both equation [1] and equation [2] have λ^2 in the denominator. We can simplify our algebra by multiplying both sides of equation [2] by Ψ to make it look more like equation [1]:

$$\begin{aligned} E \cdot \Psi &= \frac{(h/\lambda)^2}{2m}\Psi + U \cdot \Psi \\ &= \frac{1}{2m}\left(\frac{h}{2\pi}\right)^2\left(\frac{2\pi}{\lambda}\right)^2\Psi + U \cdot \Psi \\ &= -\frac{1}{2m}\left(\frac{h}{2\pi}\right)^2\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} + U \cdot \Psi \end{aligned}$$

Further simplification is achieved by using the symbol $\hbar$ (h with a slash through it, read "h-bar") as an abbreviation for $h/2\pi$. We then have the important result known as the **Schrödinger equation**:

$$E \cdot \Psi = -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} + U \cdot \Psi$$

(Actually this is a simplified version of the Schrödinger equation, applying only to standing waves in one dimension.) Physically it is a statement of conservation of energy. The total energy E must be constant, so the equation tells us that a change in interaction energy U must be accompanied by a change in the curvature of the wavefunction. This change in curvature relates to a change in wavelength, which corresponds to a change in momentum and kinetic energy.

Self-Check

Considering the assumptions that were made in deriving the Schrödinger equation, would it be correct to apply it to a photon? To an electron moving at relativistic speeds? ▷ Answer, p. 711

Usually we know right off the bat how U depends on x, so the basic mathematical problem of quantum physics is to find a function $\Psi(x)$ that satisfies the Schrödinger equation for a given interaction-energy function $U(x)$. An equation, such as the Schrödinger equation, that specifies a relationship between a function and its derivatives is known as a differential equation.

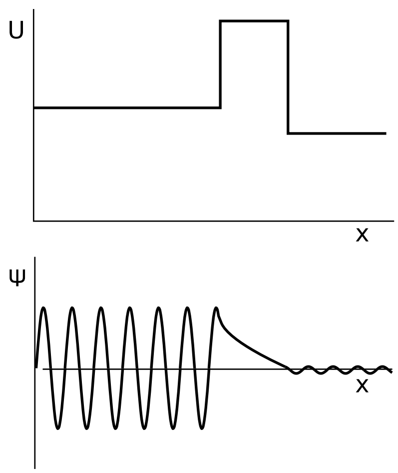
The study of differential equations in general is beyond the mathematical level of this book, but we can gain some important insights by considering the easiest version of the Schrödinger equation, in which the interaction energy U is constant. We can then rearrange the Schrödinger equation as follows:

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} = \frac{2m(U-E)}{\hbar^2}\Psi \qquad ,$$

which boils down to

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} = a\Psi \qquad ,$$

where, according to our assumptions, a is independent of x. We need to find a function whose second derivative is the same as the original function except for a multiplicative constant. The only functions with this property are sine waves and exponentials:



l / Tunneling through a barrier.

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left[\,q\sin(rx+s)\,\right] &= -qr^2\sin(rx+s)\\ \frac{\mathrm{d}^2}{\mathrm{d}x^2}\left[qe^{rx+s}\right] &= qr^2e^{rx+s}\end{aligned}$$

The sine wave gives negative values of a, $a=-r^2$, and the exponential gives positive ones, $a=r^2$. The former applies to the classically allowed region with $U<E$.

This leads us to a quantitative calculation of the tunneling effect discussed briefly in the preceding subsection. The wavefunction evidently tails off exponentially in the classically forbidden region. Suppose, as shown in figure l, a wave-particle traveling to the right encounters a barrier that it is classically forbidden to enter. Although the form of the Schrödinger equation we're using technically does not apply to traveling waves (because it makes no reference to time), it turns out that we can still use it to make a reasonable calculation of the probability that the particle will make it through the barrier. If we let the barrier's width be w, then the ratio of the wavefunction on the left side of the barrier to the wavefunction on the right is

$$\frac{qe^{rx+s}}{qe^{r(x+w)+s}} = e^{-rw} \qquad .$$

Probabilities are proportional to the squares of wavefunctions, so the probability of making it through the barrier is

$$\begin{aligned}P &= e^{-2rw}\\ &= \exp\left(-\frac{2w}{\hbar}\sqrt{2m(U-E)}\right)\end{aligned}$$

Self-Check

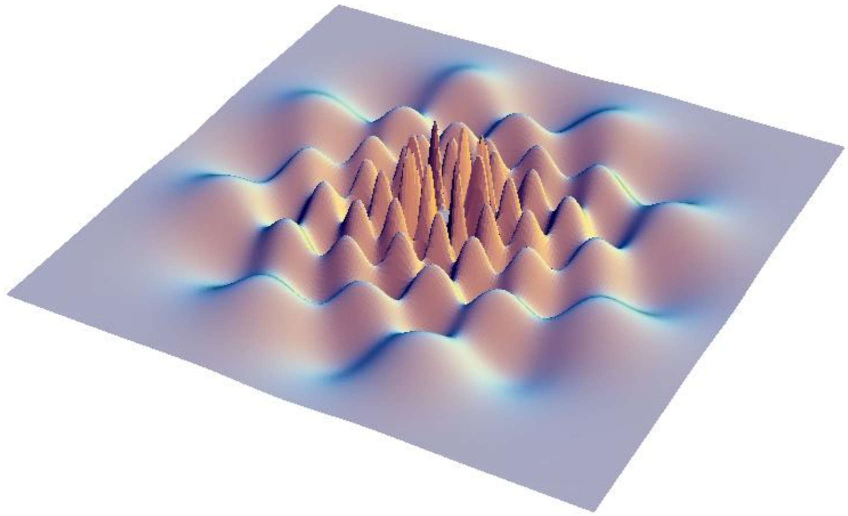
If we were to apply this equation to find the probability that a person can walk through a wall, what would the small value of Planck's constant imply? ▷ Answer, p. 711

Use of complex numbers

In a classically forbidden region, a particle's total energy, $U + K$, is less than its U, so its K must be negative. If we want to keep believing in the equation $K = p^2/2m$, then apparently the momentum of the particle is the square root of a negative number. This is a symptom of the fact that the Schrödinger equation fails to describe all of nature unless the wavefunction and various other quantities are allowed to be complex numbers. In particular it is not possible to describe traveling waves correctly without using complex wavefunctions.

This may seem like nonsense, since real numbers are the only ones that are, well, real! Quantum mechanics can always be related to the real world, however, because its structure is such that the results of measurements always come out to be real numbers. For example, we may describe an electron as having non-real momentum in classically forbidden regions, but its average momentum will always come out to be real (the imaginary parts average out to zero), and it can never transfer a non-real quantity of momentum to another particle.

A complete investigation of these issues is beyond the scope of this book, and this is why we have normally limited ourselves to standing waves, which can be described with real-valued wavefunctions.



12.4 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.

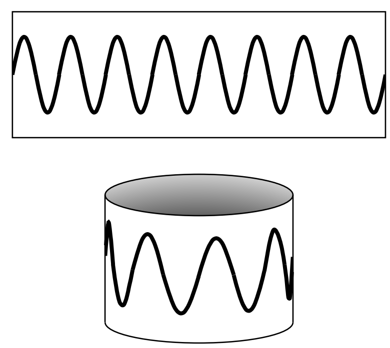
The biggest embarrassment for reductionists is that the job of putting things back together job is usually much harder than the taking them apart. Seventy years after the fundamentals of atomic physics were solved, it is only beginning to be possible to calculate accurately the properties of atoms that 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.

12.4.1 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.)



a / Eight wavelengths fit around this circle (ℓ = 8).

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.

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 for it:

$$\lambda = C/\ell$$

The angular momentum is

$$L = rp \qquad .$$

Here, $r = C/2\pi$, and $p = h/\lambda = h\ell/C$, so

$$\begin{aligned} L &= \frac{C}{2\pi} \cdot \frac{h\ell}{C} \\ &= \frac{h}{2\pi}\ell \end{aligned}$$

In the example of the quantum moat, angular momentum is quantized in units of $h/2\pi$. This makes $h/2\pi$ a pretty important number, so we define the abbreviation $\hbar = h/2\pi$. This symbol is read "h-bar."

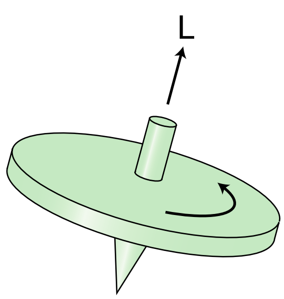
In fact, this is a completely general fact in quantum physics, not just a fact about the quantum moat:

Quantization of angular momentum
The angular momentum of a particle due to its motion through space is quantized in units of $\hbar$.

Self-Check
What is the angular momentum of the wavefunction shown at the beginning of the section? ▷ Answer, p. 711

12.4.2 Angular momentum in three dimensions

Up until now we've only worked with angular momentum in the context of rotation in a plane, for which we could simply use positive and negative signs to indicate clockwise and counterclockwise directions of rotation. A hydrogen atom, however, is unavoidably three-dimensional. Before we worry about three-dimensional angular momentum in quantum mechanics, we need to understand how it works in classical mechanics; the following discussion is a summary of the most important facts from optional section 4.3.



b / The angular momentum vector of a spinning top.

Three-dimensional angular momentum in classical physics

If we are to completely specify the angular momentum of a classical object like a top, b, in three dimensions, it's not enough to say whether the rotation is clockwise or counterclockwise. We must also give the orientation of the plane of rotation or, equivalently, the direction of the top's axis. The convention is to specify the direction of the axis. There are two possible directions along the axis, and as a matter of convention we use the direction such that if we sight along it, the rotation appears clockwise.

Angular momentum can, in fact, be defined as a vector pointing along this direction. This might seem like a strange definition, since nothing actually moves in that direction, but it wouldn't make sense to define the angular momentum vector as being in the direction of motion, because every part of the top has a different direction of motion. Ultimately it's not just a matter of picking a definition that is convenient and unambiguous: the definition we're using is the only one that makes the total angular momentum of a system a conserved quantity if we let "total" mean the vector sum.

As with rotation in one dimension, we cannot define what we mean by angular momentum in a particular situation unless we pick a point as an axis. This is really a different use of the word "axis" than the one in the previous paragraphs. Here we simply mean a point from which we measure the distance r. In the hydrogen atom, the nearly immobile proton provides a natural choice of axis.

Three-dimensional angular momentum in quantum physics

Once we start to think more carefully about the role of angular momentum in quantum physics, it may seem that there is a basic problem: the angular momentum of the electron in a hydrogen atom depends on both its distance from the proton and its momentum, 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, might seem to be forbidden by the Heisenberg uncertainty principle.

Actually the uncertainty principle does place limits on what can be known about a particle's angular momentum vector, but it does not prevent us from knowing its magnitude as an exact integer multiple of $\hbar$. The reason is that in three dimensions, there are really three separate uncertainty principles:

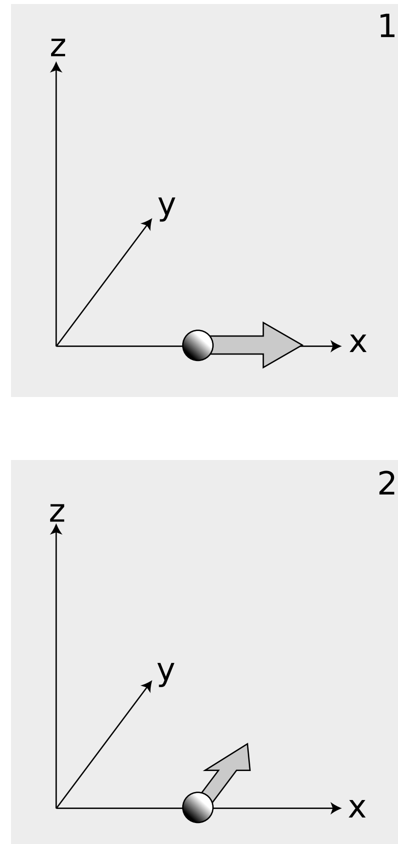
$$\begin{aligned} \Delta p_x \Delta x &\gtrsim h \\ \Delta p_y \Delta y &\gtrsim h \\ \Delta p_z \Delta z &\gtrsim h \end{aligned}$$

Now consider a particle, c/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.

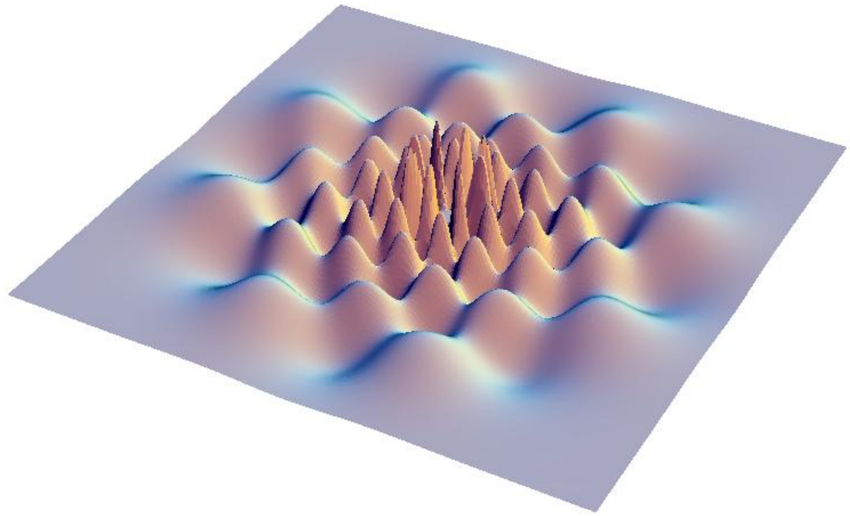
Suppose, on the other hand, a particle finds itself, c/2, at a position x along the x axis, and it is moving parallel to the y axis with momentum p_y. It has angular momentum xp_y about the z axis, and again we can know its angular momentum with unlimited accuracy, because the uncertainty principle on 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 the can be known about an angular momentum vector is its magnitude and one of its three vector components. Both are quantized in units of $\hbar$.



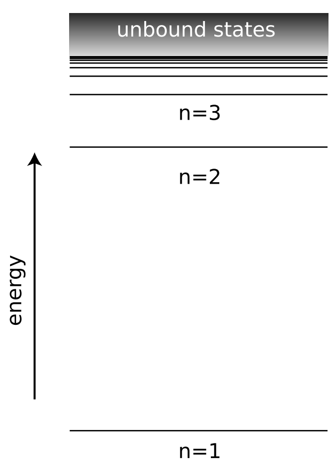
c / Reconciling the uncertainty principle with the definition of angular momentum.



d / A cross-section of a hydrogen wavefunction.

12.4.3 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 d. 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.



e / The energy of a state in the hydrogen atom depends only on its *n* quantum number.

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 12.4.4.

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.[5]

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 from $-\ell$ to ℓ, in steps of 1

Applying these rules, we have the following list of states:

$n = 1, \quad \ell = 0, \quad \ell_z = 0$	one state
$n = 2, \quad \ell = 0, \quad \ell_z = 0$	one state
$n = 2, \quad \ell = 1, \quad \ell_z = -1, 0, \text{ or } 1$	three states
...	

Self-Check
Continue the list for $n = 3$. ▷ Answer, p. 711

Figure f on page 672 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.

Discussion Questions

A The quantum number n is defined as the number of radii at which the wavefunction is zero, including $r = \infty$. Relate this to the features of the figures on the facing page.

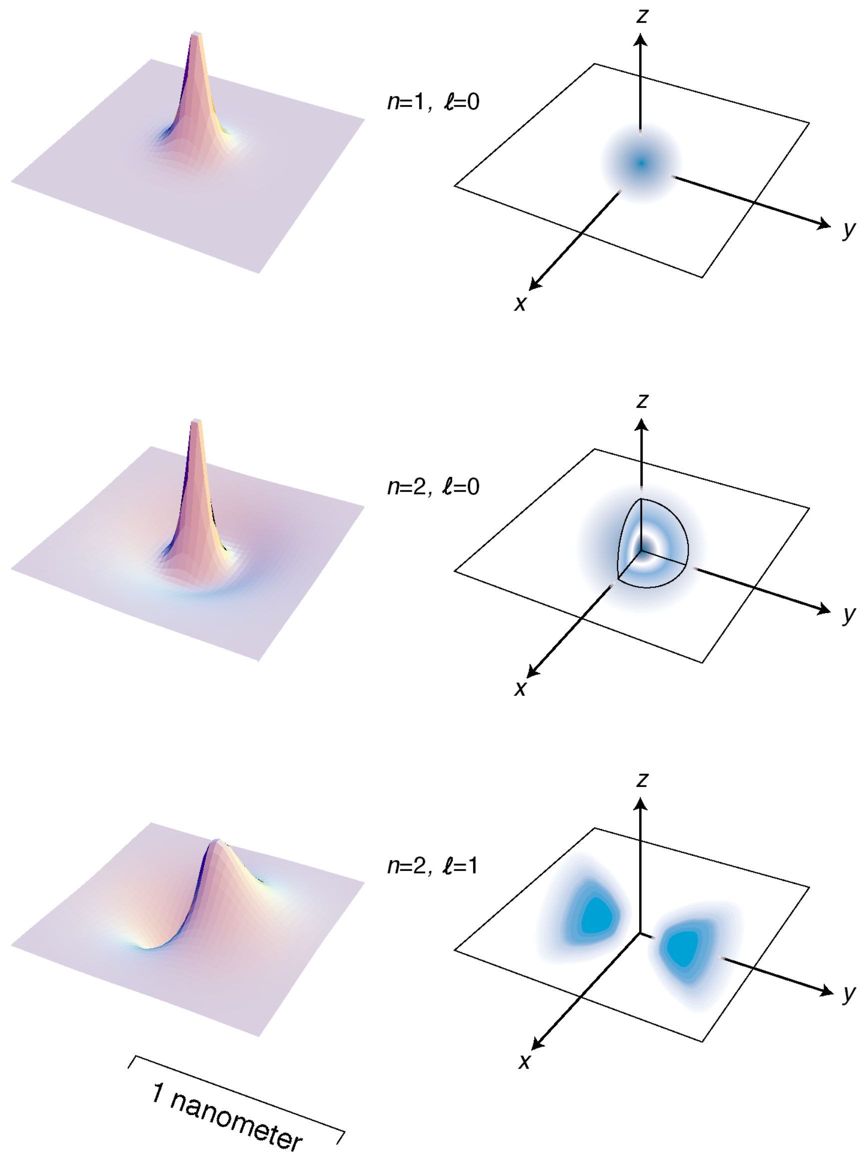
B Based on the definition of n, why can't there be any such thing as an $n = 0$ state?

C Relate the features of the wavefunction plots in figure f to the corresponding features of the probability distribution pictures.

D How can you tell from the wavefunction plots in figure f which ones have which angular momenta?

E Criticize the following incorrect statement: "The $\ell = 8$ wavefunction in figure d has a shorter wavelength in the center because in the center the electron is in a higher energy level."

[5] See page 676 for a note about the two different systems of notations that are used for quantum numbers.



f / The three states of the hydrogen atom having the lowest energies.

F Discuss the implications of the fact that the probability cloud in of the $n = 2$, $\ell = 1$ state is split into two parts.

12.4.4 Energies of states in hydrogen

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 actually 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:

$$E_n = -\frac{2.2 \times 10^{-18}\ \text{J}}{n^2} \qquad ,$$

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:

$$E_n = -\frac{mk^2e^4}{2\hbar^2} \cdot \frac{1}{n^2} \qquad ,$$

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 $\ell = 0$.

A full and correct treatment is impossible at the mathematical level of this book, but we can provide a straightforward explanation for the form of the equation using approximate arguments. 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 12.3.6, 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.

When the electron is at the distance r from the proton, it has zero kinetic energy — in classical terms, this would be the distance at which the electron would have to stop, turn around, and head back toward the proton. Thus when the electron is at distance r, its energy is purely electrical:

[1] $$E = -\frac{ke^2}{r}$$

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

[2] $$\lambda \sim \frac{r}{n} \qquad .$$

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

[3] $$\begin{aligned} &\text{absolute value of total energy} \\ &= \frac{ke^2}{r} \\ &\sim K \\ &= p^2/2m \\ &= (h/\lambda)^2/2m \\ &hn^2/2mr^2 \end{aligned}$$

We now solve the equation $ke^2/r \sim hn^2/2mr^2$ for r and throw away numerical factors we can't hope to have gotten right, yielding

[4] $$r \sim \frac{h^2n^2}{mke^2} \qquad .$$

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

$$E \sim -\frac{mk^2e^4}{h^2n^2} \qquad ,$$

which is correct except for the numerical factors we never aimed to find.

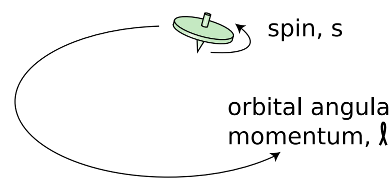
Discussion Questions

A States of hydrogen with *n* greater than about 10 are never observed in the sun. Why might this be?

B Sketch graphs of *r* and *E* versus *n* for the hydrogen, and compare with analogous graphs for the one-dimensional particle in a box.

12.4.5 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 $\hbar/2$.



g / 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:

$n = 1$,	$\ell = 0$,	$\ell_z = 0$,	$s_z = +1/2$ or $-1/2$	two states
$n = 2$,	$\ell = 0$,	$\ell_z = 0$,	$s_z = +1/2$ or $-1/2$	two states
$n = 2$,	$\ell = 1$,	$\ell_z = -1$, 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 (sometimes σ)
s_x	no notation
s_y	no notation
s_z	s

The 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:

$\ell = 0$	$\ell = 1$	$\ell = 2$	$\ell = 3$
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[6] 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,

[6] 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.

but which I simply haven't mentioned yet. The vector $\mathbf{j} = \ell + \mathbf{s}$ 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.

12.4.6 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

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:

I	II	III	IV	V	VI	VII	0
1 H							2 He
3 Li	4 Be	5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	...						

h / The beginning of the periodic table.

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:

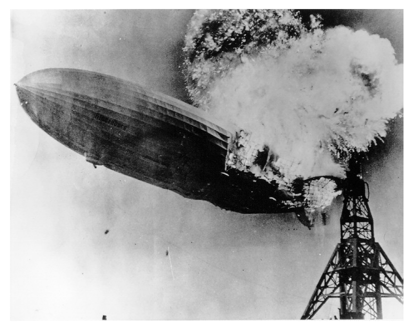
Li 2 electrons in $n = 1$ states, 1 electron in an $n = 2$ state
Be 2 electrons in $n = 1$ states, 2 electrons in $n = 2$ states
...
O 2 electrons in $n = 1$ states, 6 electrons in $n = 2$ states
F 2 electrons in $n = 1$ states, 7 electrons in $n = 2$ states
Ne 2 electrons in $n = 1$ states, 8 electrons 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? 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.



i / Hydrogen is highly reactive.

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.

Problems

The symbols √, , etc. are explained on page 689.

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 Use a calculator to check the approximation that $a^b \approx 1 + b \ln a$, if $b \ll 1$, using some arbitrary numbers. See how good the approximation is for values of b that are not quite as small compared to one.

4 Make up an example of a numerical problem involving a rate of decay where $\Delta t \ll t_{1/2}$, but the exact expression for the rate of decay on page 626 can still be evaluated on a calculator without getting something that rounds off to zero. Check that you get approximately the same result using both methods from subsection 12.1.4 to calculate the number of decays between t and $t + \Delta t$. Keep plenty of significant figures in your results, in order to show the difference between them.

5 Refer to the probability distribution for people's heights in figure e on page 621.
(a) Show that the graph is properly normalized.
(b) Estimate the fraction of the population having heights between 140 and 150 cm. √

6 A blindfolded person fires a gun at a circular target of radius b, and is allowed to continue firing until a shot actually hits it. Any part of the target is equally likely to get hit. We measure the random distance r from the center of the circle to where the bullet went in.
(a) Show that the probability distribution of r must be of the form $D(r) = kr$, where k is some constant. (Of course we have $D(r) = 0$ for $r > b$.)
(b) Determine k by requiring D to be properly normalized. √
(c) Find the average value of r. √
(d) Interpreting your result from part c, how does it compare with $b/2$? Does this make sense? Explain.

7 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. √

(b) Find the average value of t. √

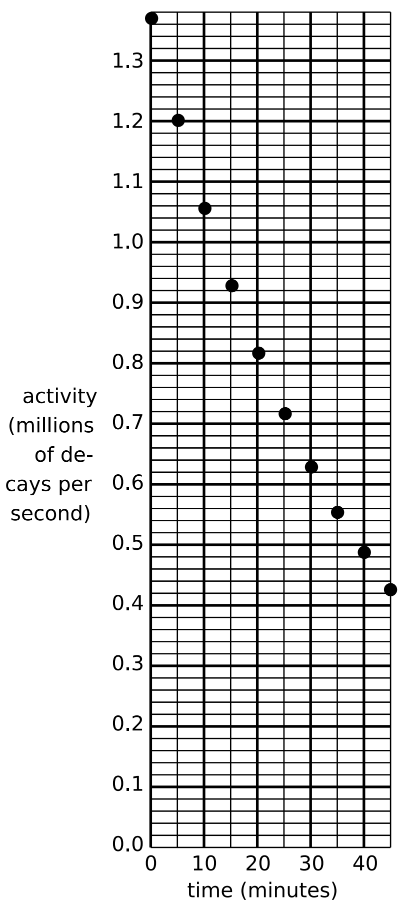
(c) Interpreting your result from part b, how does it compare with $t_{1/2}/2$? Does this make sense? Explain.

8 The speed, v, of an atom in an ideal gas has a probability distribution of the form

$$D(v) = bve^{-cv^2} \qquad ,$$

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. √

(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.) √



Problem 9.

9 (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	6.544	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?

10 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.

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}U atom is in solid rock (as opposed to the earth's molten regions), the He atoms are unable to diffuse out of the rock. This problem involves dating a rock using the known decay properties of uranium 238. Suppose a geologist finds a sample of hardened lava, melts it in a furnace, and finds that it contains 1230 mg of uranium and 2.3 mg of helium. ^{238}U decays by alpha emission, with a half-life of 4.5×10^9 years. The subsequent chain of alpha and electron (beta) decays involves much shorter half-lives, and terminates in the stable nucleus ^{206}Pb. Almost all natural uranium is ^{238}U, and the chemical composition of this rock indicates that there were no decay chains involved other than that of ^{238}U.
(a) How many alphas are emitted per decay chain? [Hint: Use conservation of mass.]
(b) How many electrons are emitted per decay chain? [Hint: Use conservation of charge.]
(c) How long has it been since the lava originally hardened? √

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

15 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.) ✓

16 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 $1\ \text{mW/m}^2$. ✓
(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. ✓

17 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.

18 Generalize the result of problem 17 to the case where the photon bounces off at an angle other than 180 °with respect to its initial direction of motion.

19 In a television, suppose the electrons are accelerated from rest through a voltage difference of 10^4 V. What is their final wavelength? √

20 Use the Heisenberg uncertainty principle to estimate the minimum velocity of a proton or neutron in a ^{208}Pb nucleus, which has a diameter of about 13 fm (1 fm=10^{-15} m). Assume that the speed is nonrelativistic, and then check at the end whether this assumption was warranted. √

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. √

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. √
(b) If a computer memory chip contains 10^8 electric circuits in a 1 cm^2 area, estimate the linear size, in nm, of one such circuit. √
(c) Based on your answers from parts a and b, does an electrical engineer designing such a chip need to worry about wave effects such as diffraction?
(d) Estimate the maximum number of electric circuits that can fit on a 1 cm^2 computer chip before quantum-mechanical effects become important.

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 sin-

gle 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 l on page 664, 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

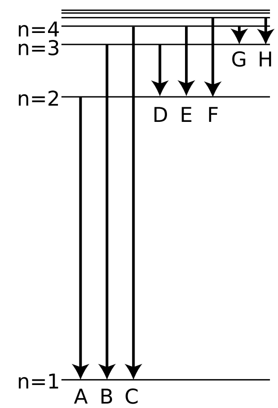
$$\Psi(x) = e^{-bx^2} \qquad ,$$

and relate b to k, m, and $\hbar$. To do this, calculate the second derivative, plug the result into the Schrödinger equation, and then find what value of b would make the equation valid for *all* values of x. This wavefunction turns out to be the ground state. Note that this wavefunction is not properly normalized — don't worry about that.
(b) Sketch a graph showing what this wavefunction looks like.
(c) Let's interpret b. If you changed b, how would the wavefunction look different? Demonstrate by sketching two graphs, one for a smaller value of b, and one for a larger value.
(d) Making k greater means making the atom more tightly bound. Mathematically, what happens to the value of b in your result from part a if you make k greater? Does this make sense physically when you compare with part c? ✓

24 On page 664 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.

25 (a) A distance scale is shown below the wavefunctions and probability densities illustrated in figure f on page 672. Compare this with the order-of-magnitude estimate derived in subsection 12.4.4 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?

26 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.



Problem 26.

27 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.]

28 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. √

29 (a) Verify that Planck's constant has the same units as angular momentum.
(b) Estimate the angular momentum of a spinning basketball, in units of $\hbar$.

30 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?

31 The wavefunction of the electron in the ground state of a hydrogen atom is

$$\Psi = \pi^{-1/2} a^{-3/2} e^{-r/a} \qquad ,$$

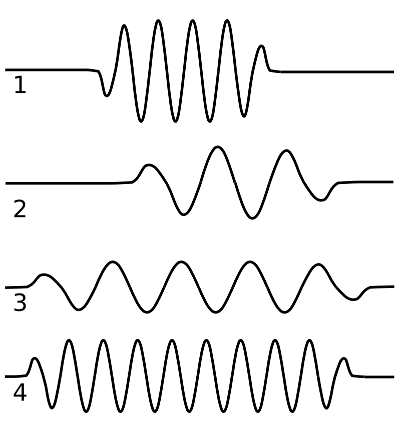
where r is the distance from the proton, and $a = 5.3 \times 10^{-11}$ m is a constant that sets the size of the wave.
(a) Calculate symbolically, without plugging in numbers, the probability that at any moment, the electron is inside the proton. Assume the proton is a sphere with a radius of $b = 0.5$ fm. [Hint: Does it matter if you plug in $r = 0$ or $r = b$ in the equation for the wavefunction?] √
(b) Calculate the probability numerically. √
(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?

32 Use physical reasoning to explain how the equation for the energy levels of hydrogen,

$$E_n = -\frac{mk^2e^4}{2\hbar^2} \cdot \frac{1}{n^2} \qquad ,$$

should be generalized to the case of a heavier atom with atomic number Z that has had all its electrons stripped away except for one.

33 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 12.4.4, 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?

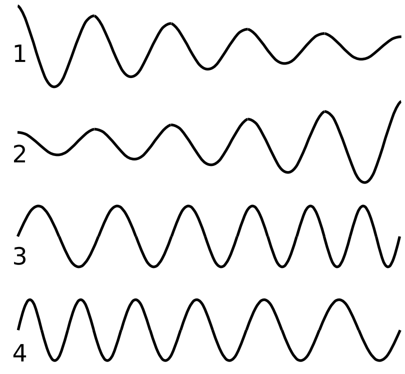


Problem 35.

34 Show that the wavefunction given in problem 31 is properly normalized.

35 (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.

36 Which figure could be an electron speeding up as it moves to the right? Explain.



Problem 36.

Key to symbols:
easy typical challenging difficult very difficult
√ An answer check is available at www.lightandmatter.com.

Exercises

Exercise 12A: Quantum Versus Classical Randomness

1. Imagine the classical version of the particle in a one-dimensional box. Suppose you insert the particle in the box and give it a known, predetermined energy, but a random initial position and a random direction of motion. You then pick a random later moment in time to see where it is. Sketch the resulting probability distribution by shading on top of a line segment. Does the probability distribution depend on energy?

2. Do similar sketches for the first few energy levels of the quantum mechanical particle in a box, and compare with 1.

3. Do the same thing as in 1, but for a classical hydrogen atom in two dimensions, which acts just like a miniature solar system. Assume you're always starting out with the same fixed values of energy and angular momentum, but a position and direction of motion that are otherwise random. Do this for $L = 0$, and compare with a real $L = 0$ probability distribution for the hydrogen atom.

4. Repeat 3 for a nonzero value of L, say L=$\hbar$.

5. Summarize: Are the classical probability distributions accurate? What qualitative features are possessed by the classical diagrams but not by the quantum mechanical ones, or vice-versa?

Appendix 1: Programming with Python

The purpose of this tutorial is to help you get familiar with a computer programming language called Python, which I've chosen because (a) it's free, and (b) it's easy to use interactively. I won't assume you have any previous experience with computer programming; you won't need to learn very much Python, and what little you do need to learn I'll explain explicitly. If you really want to learn Python more thoroughly, there are a couple of excellent books that you can download for free on the Web:

> **How to Think Like a Computer Scientist (Python Version)**, Allen B. Downey, Jeffrey Elkner, Moshe Zadka, http://www.ibiblio.org/obp/
>
> **Dive Into Python**, Mark Pilgrim, http://diveintopython.org/

The first book is meant for people who have never programmed before, while the second is a more complete introduction aimed at veteran programmers who know a different language already.

Using Python as a calculator

Run Python. It will present you with a prompt that looks something like this:

```
Python 2.0
>>>
```

The "`>>>`" is a prompt for you to type something. Try asking it to add 2+2:

```
>>> print 2+2
4
```

In other words, you can use Python just like a calculator. Note that if you're a slow typist, you can usually get away with not typing the word "`print`."

There are only a couple of things to watch out for. First, Python distinguishes between integers and real numbers, so the following gives an unexpected result:

```
>>>  print 2/3
0
```

To get it to treat these values as real numbers, you have to use decimal points:

```
>>> print 2./3.
0.66666666666666666663
```

Multiplication is represented by "*":

```
>>> print 2.*3.
6.0
```

Also, Python doesn't know about its own library of math functions unless you tell it explicitly to load them in:

```
>>> print  sqrt(2.)
Traceback (most recent call last):
File ''<stdin>'', line 1, in ?
NameError: There is no variable named sqrt'
```

Here are the steps you have to go through to calculate the square root of 2 successfully:

```
>>> import math
>>> print math.sqrt(2.)
1.4142135623730951
```

The first line is just something you can make a habit of doing every time you start up Python. In the second line, the name of the square root function had to be prefixed with "`math.`" to tell Python where you wanted to get this "`sqrt`" function from. (All of this may seem like a nuisance if you're just using Python as a calculator, but it's a good way to design a programming language so that names of functions never conflict.)

Try it. Experiment and figure out whether Python's trig functions assume radians or degrees.

Variables

Python lets you define variables and assign values to them using an equals sign:

```
>>> dwarfs=7
>>> print dwarfs
7
>>> print dwarfs+3
10
```

Note that a variable in computer programming isn't quite like a variable in algebra. In algebra, if a=7 then a=7 always, throughout a particular calculation. But in a programming language, the variable name really represents a place in memory where a number can be stored, so you can change its value:

```
>>> dwarfs=37
>>> print dwarfs
37
```

You can even do stuff like this,

```
>>> dwarfs=dwarfs+1
>>> print dwarfs
38
```

In algebra it would be nonsense to have a variable equal to itself plus one, but in a computer program, it's not an assertion that the two things are equal, its a command to calculate the value of the expression on the right side of the equals, and then put that number into the memory location referred to by the variable name on the left.

Try it. What happens if you do dwarfs+1 = dwarfs? Do you understand why?

Functions

Somebody had to teach Python how to do functions like `sqrt`, and it's handy to be able to define your own functions in the same way. Here's how to do it:

```
>>> def double(x):
...   return 2.*x
...
```

The "`...`" prompt means that Python knows your definition isn't complete yet. Note that the indentation is mandatory. The second time you get the "`...`" prompt, you just hit return, indicating you're done entering the function. We can now use the function:

```
>>> print double(5.)
10.0
```

Loops

Suppose we want to add up all the numbers from 0 to 99.

Automating this kind of thing is exactly what computers are best at, and Python provides a mechanism for this called a loop:

```
>>> sum=0
>>> for j in range(100):
...   sum=sum+j
...
>>> print sum
4950
```

The stuff that gets repeated — the inside of the loop — has to be indented, just like in a function definition. Python always counts loops starting from 0, so for `j in range(100)` actually causes `j` to range from 0 to 99, not from 1 to 100.

Appendix 2: Miscellany

Automated search for the brachistochrone

See page 56.

```
d=.01
c1=.61905
c2=-.94427
a = 1.
b = 1.
for i in range(100):
  bestt = 99.
  for j in range(3):
    for k in range(3):
      try_c1 = c1+(j-1)*d
      try_c2 = c2+(k-1)*d
      t = timeb(a,b,try_c1,try_c2,100000)
      if t<bestt :
        bestc1 = try_c1
        bestc2 = try_c2
        bestj = j
        bestk = k
        bestt = t
  c1 = bestc1
  c2 = bestc2
  c3 = (b-c1*a-c2*a**2)/(a**3)
  print c1, c2, c3, bestt
  if (bestj == 1) and (bestk == 1) :
    d = d*.5
```

Derivation of the steady state for damped, driven oscillations

Using the trig identities for the sine of a sum and cosine of a sum, we can change equation [2] on page 128 into the form

$$\begin{aligned}&\left[(-m\omega^2+k)\cos\delta - b\omega\sin\delta - F_m/A\right]\sin\omega t \\ &+\left[(-m\omega^2+k)\sin\delta + b\omega\cos\delta\right]\cos\omega t = 0 \qquad .\end{aligned}$$

Both the quantities in square brackets must equal zero, which gives us two equations we can use to determine the unknowns A and δ. The results are

$$\begin{aligned}\delta &= \tan^{-1}\frac{b\omega}{m\omega^2-k} \\ &= \tan^{-1}\frac{\omega\omega_o}{Q(\omega_o^2-\omega^2)}\end{aligned}$$

and

$$A = \frac{F_m}{\sqrt{(m\omega^2 - k)^2 + b^2\omega^2}}$$
$$= \frac{F_m}{m\sqrt{(\omega^2 - \omega_o^2)^2 + \omega_o^2\omega^2 Q^{-2}}} \quad .$$

Proofs relating to angular momentum

Uniqueness of the cross product

The vector cross product as we have defined it has the following properties:
(1) It does not violate rotational invariance.
(2) It has the property $\mathbf{A} \times (\mathbf{B} + \mathbf{C}) = \mathbf{A} \times \mathbf{B} + \mathbf{A} \times \mathbf{C}$.
(3) It has the property $\mathbf{A} \times (k\mathbf{B}) = k(\mathbf{A} \times \mathbf{B})$, where k is a scalar.

Theorem: The definition we have given is the only possible method of multiplying two vectors to make a third vector which has these properties, with the exception of trivial redefinitions which just involve multiplying all the results by the same constant or swapping the names of the axes. (Specifically, using a left-hand rule rather than a right-hand rule corresponds to multiplying all the results by -1.)

Proof: We prove only the uniqueness of the definition, without explicitly proving that it has properties (1) through (3).

Using properties (2) and (3), we can break down any vector multiplication $(A_x\hat{\mathbf{x}} + A_y\hat{\mathbf{y}} + A_z\hat{\mathbf{z}}) \times (B_x\hat{\mathbf{x}} + B_y\hat{\mathbf{y}} + B_z\hat{\mathbf{z}})$ into terms involving cross products of unit vectors.

A "self-term" like $\hat{\mathbf{x}} \times \hat{\mathbf{x}}$ must either be zero or lie along the x axis, since any other direction would violate property (1). If was not zero, then $(-\hat{\mathbf{x}}) \times (-\hat{\mathbf{x}})$ would have to lie in the opposite direction to avoid breaking rotational invariance, but property (3) says that $(-\hat{\mathbf{x}}) \times (-\hat{\mathbf{x}})$ is the same as $\hat{\mathbf{x}} \times \hat{\mathbf{x}}$, which is a contradiction. Therefore the self-terms must be zero.

An "other-term" like $\hat{\mathbf{x}} \times \hat{\mathbf{y}}$ could conceivably have components in the x-y plane and along the z axis. If there was a nonzero component in the x-y plane, symmetry would require that it lie along the diagonal between the x and y axes, and similarly the in-the-plane component of $(-\hat{\mathbf{x}}) \times \hat{\mathbf{y}}$ would have to be along the other diagonal in the x-y plane. Property (3), however, requires that $(-\hat{\mathbf{x}}) \times \hat{\mathbf{y}}$ equal $-(\hat{\mathbf{x}} \times \hat{\mathbf{y}})$, which would be along the original diagonal. The only way it can lie along both diagonals is if it is zero.

We now know that $\hat{\mathbf{x}} \times \hat{\mathbf{y}}$ must lie along the z axis. Since we are not interested in trivial differences in definitions, we can fix $\hat{\mathbf{x}} \times \hat{\mathbf{y}} = \hat{\mathbf{z}}$, ignoring peurile possibilities such as $\hat{\mathbf{x}} \times \hat{\mathbf{y}} = 7\hat{\mathbf{z}}$ or the left-handed definition $\hat{\mathbf{x}} \times \hat{\mathbf{y}} = -\hat{\mathbf{z}}$. Given $\hat{\mathbf{x}} \times \hat{\mathbf{y}} = \hat{\mathbf{z}}$, the symmetry of space requires that similar relations hold for $\hat{\mathbf{y}} \times \hat{\mathbf{z}}$ and $\hat{\mathbf{z}} \times \hat{\mathbf{x}}$, with at most a difference in sign. A difference in sign could always be eliminated by swapping the names of some of the axes, so ignoring possible trivial differences in definitions we can assume that the cyclically related set of relations $\hat{\mathbf{x}} \times \hat{\mathbf{y}} = \hat{\mathbf{z}}$, $\hat{\mathbf{y}} \times \hat{\mathbf{z}} = \hat{\mathbf{x}}$, and $\hat{\mathbf{z}} \times \hat{\mathbf{x}} = \hat{\mathbf{y}}$ holds. Since the arbitrary cross-product with which we started can be broken down into these simpler ones, the cross product is uniquely defined.

The choice of axis theorem

Theorem: Suppose a system of material particles conserves angular momentum in one frame of reference, with the axis taken to be at the origin. Then conservation of angular momentum is unaffected if the origin is relocated or if we change to a frame of reference that is in constant-velocity motion with respect to the first one.

Proof: In the original frame of reference, angular momentum is conserved, so we have $\mathrm{d}\mathbf{L}/\mathrm{d}t=0$. From example 28 on page 216, this derivative can be rewritten as

$$\frac{\mathrm{d}\mathbf{L}}{\mathrm{d}t} = \sum_i \mathbf{r}_i \times \mathbf{F}_i \qquad ,$$

where $\mathbf{F}_i$ is the total force acting on particle i. In other words, we're adding up all the torques on all the particles.

By changing to the new frame of reference, we have changed the position vector of each particle according to $\mathbf{r}_i \rightarrow \mathbf{r}_i + \mathbf{k} - \mathbf{u}t$, where $\mathbf{k}$ is a constant vector that indicates the relative position of the new origin at $t = 0$, and $\mathbf{u}$ is the velocity of the new frame with respect to the old one. The forces are all the same in the new frame of reference, however. In the new frame, the rate of change of the angular momentum is

$$\begin{aligned}\frac{\mathrm{d}\mathbf{L}}{\mathrm{d}t} &= \sum_i (\mathbf{r}_i + \mathbf{k} - \mathbf{u}t) \times \mathbf{F}_i \\ &= \sum_i \mathbf{r}_i \times \mathbf{F}_i + (\mathbf{k} - \mathbf{u}t) \times \sum_i \mathbf{F}_i \qquad .\end{aligned}$$

The first term is the expression for the rate of change of the angular momentum in the original frame of reference, which is zero by assumption. The second term vanishes by Newton's third law. The rate of change of the angular momentum is therefore zero in the new frame of reference.

The spin theorem

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.

Proof: Let $\mathbf{r}_{cm}$ be the position of the center of mass. The total angular momentum is

$$\mathbf{L} = \sum_i \mathbf{r}_i \times \mathbf{p}_i \qquad ,$$

which can be rewritten as

$$\mathbf{L} = \sum_i (\mathbf{r}_{cm} + \mathbf{r}_i - \mathbf{r}_{cm}) \times \mathbf{p}_i \qquad ,$$

where $\mathbf{r}_i - \mathbf{r}_{cm}$ is particle i's position relative to the center of mass. We then have

$$\begin{aligned}\mathbf{L} &= \mathbf{r}_{cm} \times \sum_i \mathbf{p}_i + \sum_i (\mathbf{r}_i - \mathbf{r}_{cm}) \times \mathbf{p}_i \\ &= \mathbf{r}_{cm} \times \mathbf{p}_{total} + \sum_i (\mathbf{r}_i - \mathbf{r}_{cm}) \times \mathbf{p}_i \\ &= \mathbf{r}_{cm} \times m_{total}\mathbf{v}_{cm} + \sum_i (\mathbf{r}_i - \mathbf{r}_{cm}) \times \mathbf{p}_i \qquad .\end{aligned}$$

The first and second terms in this expression correspond to the quantities (2) and (1), respectively.

Proofs relating to thermodynamics

Temperature and pressure in insulated compression or expansion

We wish to calculate the relationship between temperature and pressure when a monoatomic ideal gas is compressed or expanded without being able to exchange heat with the rest of the universe. This result will be needed in deriving the efficiency of a Carnot engine below.

The loss of thermal energy from the gas equals the work it does on its container (e.g. by pushing a piston out), and using the result of homework problem 2 on page 260, the work done in an infinitesimal expansion equals $P\mathrm{d}V$, so

$$\mathrm{d}E + P\mathrm{d}V = 0 \qquad .$$

(If the gas had not been insulated, then there would have been a third term for the heat gained or lost by heat conduction.) From section 5.2 we have $PV = (2/3)E$ for a monoatomic ideal gas, so $\mathrm{d}E = (3/2)P\mathrm{d}V + (3/2)V\mathrm{d}P$, and the equation above becomes

$$0 = (5/2)P\mathrm{d}V + (3/2)V\mathrm{d}P \qquad .$$

Rearranging, we have

$$5\frac{\mathrm{d}V}{V} = -3\frac{\mathrm{d}P}{P} \qquad .$$

Integrating both sides gives

$$5\ln V = -3\ln P + \text{constant},$$

and taking exponentials on both sides yields

$$V^5 \propto P^{-3} \qquad .$$

We now wish to reexpress this in terms of pressure and temperature. Eliminating $V \propto (T/P)$ gives

$$T \propto P^b,$$

where we write b in place of $2/5$ to emphasize that the numerical value of this exponent is not important to any of our later arguments. Indeed, the exponent would be different for a gas that was not monoatomic.

Heat transfer in constant-temperature compression or expansion

Having now analyzed the behavior of a monoatomic ideal gas as is relevant to the insulated strokes of a Carnot engine, we turn to the work done during the constant-temperature strokes. Integrating the equation $\mathrm{d}W = P\mathrm{d}V$, we have $W = \int P\mathrm{d}V$. Since the thermal energy of an ideal gas depends only on its temperature, there is no change in the thermal energy of the gas during this constant-temperature process. Conservation of energy therefore tells us that work done by the gas must be exactly balanced by the amount of heat transferred in from the reservoir.

$$\begin{aligned} Q &= W \\ &= \int P\mathrm{d}V \end{aligned}$$

For our proof of the efficiency of the Carnot engine, we need only the ratio of Q_H to Q_L, so we neglect constants of proportionality, and simply subsitutde $P \propto T/V$, giving

$$Q \propto \int \frac{T}{V} \mathrm{d}V$$
$$Q \propto T \ln \frac{V_2}{V_1}$$
$$Q \propto T \ln \frac{P_1}{P_2}$$

Efficiency of a Carnot engine

The efficiency of a heat engine is

$$\text{efficiency} = 1 - \frac{Q_L}{Q_H} \qquad .$$

Making use of the result from the previous proof for a Carnot engine with a monoatomic ideal gas as its working gas, we have

$$\text{efficiency} = 1 - \frac{T_L \ \ln(P_4/P_3)}{T_H \ \ln(P_1/P_2)},$$

where the subscripts 1, 2, 3, and 4 refer to figures d–g on page 241. We have shown above that the temperature is proportional to P^b on the insulated strokes 2-3 and 4-1, the pressures must be related by $P_2/P_3 = P_1/P_4$, which can be rearranged as $P_4/P_3 = P_1/P_2$, and we therefore have

$$\text{efficiency} = 1 - \frac{T_L}{T_H} \qquad .$$

Proofs relating to relativity

Combination of velocities

We proceed by transforming from the x, t frame to the x', t' frame moving relative to it at a velocity v_1, and then from that to a third frame, x'', t'', moving with respect to the second at v_2. The result must be equivalent to a single transformation from x, t to x'', t'' using the combined velocity. Transforming from x, t to x', t' gives

$$x' = \gamma_1 x - v_1 \gamma_1 t$$
$$t' = -v_1 \gamma_1 x + \gamma_1 t \qquad ,$$

and plugging this into the second transformation results in

$$x'' = \gamma_2(\gamma_1 x - v_1 \gamma_1 t) - v_2 \gamma_2(-v_1 \gamma_1 x + \gamma_1 t)$$
$$t'' = \ldots + \ldots \qquad ,$$

where "..." indicates terms that we don't need in order to complete the derivation. Collecting terms gives

$$x'' = (\ldots)x - (v_1 + v_2)\gamma_1 \gamma_2 t \qquad ,$$

where the coefficient of t, $-(v_1 + v_2)\gamma_1\gamma_2$, must be the same as it would have been in a direct transformation from x, t to x'', t'':

$$-v_{combined}\gamma_{combined} = -(v_1 + v_2)\gamma_1\gamma_2$$

Straightforward algebra then produces the equation on page 732.

Relativistic momentum

We want to show that if $p = m\gamma v$, then any collision that conserves momentum in the center of mass frame will also conserve momentum in any other frame. The whole thing is restricted to two-body collisions in one dimension in which no kinetic energy is changed to any other form, so it is not a general proof that $p = m\gamma v$ forms a consistent part of the theory of relativity. This is just the minimum test we want the equation to pass.

Let the new frame be moving at a velocity u with respect to the center of mass and let Γ (capital gamma) be $1/\sqrt{1-u^2}$. Then the total momentum in the new frame (at any moment before or after the collision) is

$$p' = m_1\gamma_1' v_1' + m_2\gamma_2' v_2' \qquad .$$

The velocities v_1' and v_2' result from combining v_1 and v_2 with u, so making use of the result from the previous proof,

$$\begin{aligned} p' &= m_1(v_1+u)\Gamma\gamma_1 + m_2(v_2+u)\Gamma\gamma_2 \\ &= (m_1\gamma_1 v_1 + m_2\gamma_2 v_2)\Gamma + (m_1\gamma_1 + m_2\gamma_2)\Gamma u \\ &= p\Gamma + (K_1 + m_1 + K_2 + m_2)\Gamma u \qquad . \end{aligned}$$

If momentum is conserved in the center of mass frame, then there is no change in p, the momentum in the center of mass frame, after the collision. The first term is therefore the same before and after, and the second term is also the same before and after because mass is invariant, and we have assumed no kinetic energy was converted to other forms of energy. (We shouldn't expect the proof to work if kinetic energy is changed to other forms, because we have not taken into account the effects of any other forms of mass-energy.)

Relativistic work-kinetic energy theorem

This is a straightforward application of calculus, albeit with a couple of tricks to make it easier to do without recourse to a table of integrals. The kinetic energy of an object of mass m moving with velocity v equals the work done in accelerating it to that speed from rest:

$$\begin{aligned} K &= \int_{v=0}^{v} F\,\mathrm{d}x \\ &= \int_{v=0}^{v} \frac{\mathrm{d}p}{\mathrm{d}t}\,\mathrm{d}x \\ &= \int_{v=0}^{v} \frac{\mathrm{d}(m\gamma v)}{\mathrm{d}t}\,\mathrm{d}x \\ &= m\int_{v=0}^{v} v\,\mathrm{d}(\gamma v) \\ &= m\int_{v=0}^{v} v^2\,\mathrm{d}\gamma + m\int_{v=0}^{v} v\gamma\,\mathrm{d}v \\ &= m\int_{v=0}^{v} \left(1-\gamma^{-2}\right)\,\mathrm{d}\gamma + m\int_{v=0}^{v} \frac{v\,\mathrm{d}v}{\sqrt{1-v^2}} \end{aligned}$$

$$
\begin{aligned}
&= m\left(\gamma+\frac{1}{\gamma}\right)\Bigg]_{v=0}^{v} - m\sqrt{1-v^2}\Big]_{v=0}^{v}\\
&= m\left(\gamma+\frac{1}{\gamma}-\sqrt{1-v^2}\right)\Bigg]_{v=0}^{v}\\
&= m\gamma]_{v=0}^{v}\\
&= m(\gamma-1)
\end{aligned}
$$

Change in inertia with heating

We prove here that the inertia of a heated object (its apparent mass) increases by an amount equal to the heat.[7] Suppose an object moving with velocity v_{cm} consists of molecules with masses $m_1, m_2, \ldots$, which are moving relative to the origin at velocities $v_{o1}, v_{o2}, \ldots$ and relative to the object's center of mass at velocities $v_1, v_2, \ldots$ The total momentum is

$$
\begin{aligned}
p_{total} &= \sum m_j v_{oj} \gamma_{oj}\\
&= \sum m_j (v_{cm}+v_j)\gamma_{cm}\gamma_j
\end{aligned}
$$

where we have used the result from the first proof above. Rearranging,

$$p_{total} = \gamma_{cm}\left(\sum m_j\gamma_j v_{cm} + \sum m_j\gamma_j v_j\right)$$

The second term, which is the total momentum in the c.m. frame, vanishes.

$$p_{total} = \left(\sum m_j\gamma_j\right)\gamma_{cm} v_{cm}$$

The quantity in parentheses is the total mass plus the total thermal energy.

Different Forms of Maxwell's Equations

First we reproduce Maxwell's equations as stated on page 587, in integral form, using the SI (meter-kilogram-second) system of units, with the coupling constants written in terms of k and c:

$$
\begin{aligned}
\Phi_E &= 4\pi k q_{in}\\
\Phi_B &= 0\\
\Gamma_E &= -\frac{\partial \Phi_B}{\partial t}\\
c^2\Gamma_B &= \frac{\partial \Phi_E}{\partial t} + 4\pi k I_{through}
\end{aligned}
$$

Homework problem 36 on page 607 deals with rewriting these in terms of $\epsilon_o = 1/4\pi k$ and $\mu_o = 4\pi k/c^2$ rather than k and c.

For the reader who has been studying the optional sections giving Maxwell's equations in differential form, here is a summary:

[7] This proof really only applies to an ideal gas, which expresses all of its heat energy as kinetic energy. In general heat energy is expressed partly as kinetic energy and partly as electrical potential energy.

$$\begin{aligned}
\operatorname{div}\mathbf{E} &= 4\pi k\rho \\
\operatorname{div}\mathbf{B} &= 0 \\
\operatorname{curl}\mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\
c^2\operatorname{curl}\mathbf{B} &= \frac{\partial \mathbf{E}}{\partial t} + 4\pi k\,\mathbf{j}
\end{aligned}$$

Although all engineering and most scientific work these days is done in the SI (mks) system, one may still encounter the older cgs (centimeter-gram-second) system, especially in astronomy and particle physics. The mechanical units in this system include the dyne ($\text{g}\cdot\text{cm}/\text{s}^2$) for force, and the erg ($\text{g}\cdot\text{cm}^2/\text{s}^2$) for energy. The system is extended to electrical units by taking $k = 1$ as a matter of definition, so the Coulomb force law is $F = q_1q_2/r^2$. This equation indirectly defines a unit of charge called the elestrostatic unit, with 1 C $= 2.998 \times 10^9$ esu, the factor of 2.998 arising from the speed of light. The unit of voltage is the statvolt, 1 statvolt = 299.8 V. In this system, the electric and magnetic fields have the same units, dynes/esu, but to avoid confusion, magnetic fields are normally written using the equivalent unit of gauss, 1 gauss=1 dyne/esu=10^{-4} T. The force on a charged particle is $\mathbf{F} = q\mathbf{E} + q\frac{\mathbf{v}}{c} \times \mathbf{B}$, which differs from the mks version by the $1/c$ factor in the magnetic term. Maxwell's equations are:

$$\begin{aligned}
\Phi_E &= 4\pi q_{in} \\
\Phi_B &= 0 \\
\Gamma_E &= -\frac{1}{c}\frac{\partial \Phi_B}{\partial t} \\
\Gamma_B &= \frac{1}{c}\frac{\partial \Phi_E}{\partial t} + \frac{4\pi}{c} I_{through}
\end{aligned}$$

Appendix 3: Photo Credits

Except as specifically noted below or in a parenthetical credit in the caption of a figure, all the illustrations in this book are by under my own copyright, and are copyleft licensed under the same license as the rest of the book.

In some cases it's clear from the date that the figure is public domain, but I don't know the name of the artist or photographer; I would be grateful to anyone who could help me to give proper credit. I have assumed that images that come from U.S. government web pages are copyright-free, since products of federal agencies fall into the public domain. When "PSSC Physics" is given as a credit, it indicates that the figure is from the second edition of the textbook entitled Physics, by the Physical Science Study Committee; these are used according to a blanket permission given in the later PSSC College Physics edition, which states on the copyright page that "The materials taken from the original and second editions and the Advanced Topics of PSSC PHYSICS included in this text will be available to all publishers for use in English after December 31, 1970, and in translations after December 31, 1975."

In a few cases, I have made use of images under the fair use doctrine. However, I am not a lawyer, and the laws on fair use are vague, so you should not assume that it's legal for you to use these images. In particular, fair use law may give you less leeway than it gives me, because I'm using the images for educational purposes, and giving the book away for free. Likewise, if the photo credit says "courtesy of ...," that means the copyright owner gave me permission to use it, but that doesn't mean you have permission to use it.

14 *Portrait of Monsieur Lavoisier and His Wife:* Jacques-Louis David, 1788. **15** *Astronaut:* NASA. **22** *Portrait of Galileo Galilei:* Justus Sustermans, 1636. **24** *Foucault and pendulum:* contemporary, ca. 1851. **31** *Wind tunnel:* NASA. **33** *Portrait of James Joule:* contemporary. **34,36,37** *Infrared photographs:* Courtesy of M. Vollmer and K.P. Möllmann, Univ. Appl. Sciences, Brandenburg, Germany, www.fh-brandenburg.de/~piweb/projekte/thermo_galerie_eng.html. **42** *Skateboarder:* Courtesy of J.D. Rogge, www.sonic.net/~shawn. **80** *Colliding balls:* PSSC Physics. **171** *Spider oscillations:* Emile, Le Floch, and Vollrath, *Nature* 440:621 (2006). **91** *Ion drive:* NASA. **93** *Halley's comet:* W. Liller. **93** *Nucleus of Halley's comet:* European Space Agency. **96** *Colliding galazies:* NASA. **101** *Wrench:* PSSC Physics. **101** *Highjumper:* Courtesy of Dunia Young. **131** *Nimitz freeway:* Courtesy of U.C. Berkeley Earth Sciences and Map Library. **135** *Descartes:* French postal stamp. **150** *Solar sail (artist's rendering):* Wikipedia user Paranoid, GFDL licensed. **159** *Breaking trail:* Art by Walter E. Bohl. Image courtesy of the University of Michigan Museum of Art/School of Information and Library Studies. **179** *High jump:* Thomas Eakins, public domain. **219** *Explorer I:* NASA/JPL, public domain. **240** *Carnot:* contemporary. **247** *Space junk:* STK-generated images courtesy of CSSI (www.centerforspace.com). **248** *Boltzmann's tomb:* Wikipedia user Daderot, GFDL licensed. **254** *Otto cycle:* based on an animation by Wikipedia user UtzOnBike, GFDL licensed. **283** *Mount Wilson:* Andrew Dunn, cc-by-sa licensed. **284** *X-15 shock wave:* NASA, public domain. **??** *Flute:* Wikipedia user Grendelkhan, GFDL licensed. **305** *Einstein:* "Professor Einstein's Visit to the United States," The Scientific Monthly 12:5 (1921), p. 483, public domain. **315** *RHIC collision:* Courtesy of RHIC. **339** *Eclipse:* 1919. **349** *Lightning:* C. Clark, NOAA photo library. **363** *Millikan:* contemporary. **367** *Thomson:* Harper's Monthly, 1904. **394** *Neutrino detector:* Copyrighted, not covered by the book's copyleft. **376** *nuclear fuel pellets:* US DOE, public domain. **389** *nuclear power plant:* Wikipedia user Stefan Kuhn, GFDL licensed. **394** *Raft in neutrino detector:* Kamioka Observatory, ICRR (Institute for Cosmic Ray Research), The University of Tokyo. **396** *GAMMASPHERE:* Courtesy of C.J. Lister and R.V.F. Janssens. **396** *H bomb test:* public domain product of US DOE, Ivy Mike test. **396** *Fatu Hiva Rainforest:* Wikipedia user Makemake, GFDL licensed. **396** *fusion reactor:* "These images may be used free of charge for educational purposes but please use the acknowledgement 'photograph courtesy of EFDA-JET"'. **396** *sun:* SOHO (ESA & NASA). **399** *Chernobyl village:* Wikipedia user Slawojarek, GFDL licensed. **399** *Chernobyl map:* CIA Handbook of International Economic Statistics, 1996, public domain. **402** *UNILAC:* Copyrighted, not covered by the book's copyleft. **408** *Knifefish:* Courtesy of Greg DeGreef. **420** *Superconducting accelerator segment:* Courtesy of Argonne National Laboratory, managed and operated by the University of Chicago for the U.S. Department of Energy under contract no. W-31-109-ENG-38. **456** *LIGO:* California Institute of Technology. **464** *Topographical map:* United States Geological Survey, 19th century, uncopyrighted.. **481** *Capacitors:* Wikipedia user de:Benutzer:Honina, GFDL licensed. **481** *Inductors:* Wikipedia user de:Benutzer:Honina, GFDL licensed. **490** *Ballasts:* Magnetic ballast: Wikimedia Commons use Atlant, CC-BY; solid-state ballast: Wikipedia user Anton, GFDL. **576** *Faraday banknote:* fair use. **577** *Ascending and Descending:* (c) 1960, M.C. Escher. **613** *Mount St. Helens:* Public-domain image by Austin Post, USGS. **630** *Ozone maps:* NASA/GSFC TOMS Team. **631** *Photon interference photos:* Courtesy of Prof. Lyman Page. **644** *Wicked witch:* W.W. Denslow, 1900. Quote from The Wizard of Oz, by L. Frank Baum, 1900. **679** *Hindenburg:* Public domain product of the U.S. Navy.

Appendix 4: Hints and Solutions

Hints

Hints for Chapter 2

Page 82, problem 16: You can use either the chain-rule technique from page 44 or the technique prescribed in problem 15 on p. 82. The positions and velocities of the two masses are related to each other, and you'll need to use this relationship to eliminate one mass's position and velocity and get everything in terms of the other mass's position and velocity. The relationship between the two positions will involve some extraneous variables like the length of the string, which won't have any effect on your final result.

Page 82, problem 17: This is similar to problem 16, but you're trying to find the combination of masses that will result in *zero* acceleration. In this problem, the distance dropped by one weight is different from, but still related to, the distance by which the other weight rises. Try relating the heights of the two weights to each other, so you can get the total gravitational energy in terms of only one of these heights.

Page 82, problem 18: This is similar to problem 17, in that you're looking for a setup that will give zero acceleration, and the distance the middle weight rises or falls is not the same as the distance the other two weights fall or rise. The simplest approach is to get the three heights in terms of θ, so that you can write the gravitational energy in terms of θ.

Page 82, problem 19: This is very similar to problems 16 and 17.

Page 82, problem 20: The first two parts can be done more easily by setting $a = 1$, since the value of a only changes the distance scale.

Page 83, problem 22: The condition for a circular orbit contains three unknowns, v, g, and r, so you can't just solve it for r. You'll need more equations to make three equations in three unknowns. You'll need a relationship between g and r, and also a relationship between v and r that uses the given fact that it's supposed to take 24 hours for an orbit.

Page 83, problem 25: What does the total energy have to be if the projectile's velocity is exactly escape velocity? Write down conservation of energy, change v to $\mathrm{d}r/\mathrm{d}t$, separate the variables, and integrate.

Page 83, problem 26: The analytic approach is a little cumbersome, although it can be done by using approximations like $1/\sqrt{1+\epsilon} \approx 1 - (1/2)\epsilon$. A more straightforward, brute-force method is simply to write a computer program that calculates U/m for a given point in spherical coordinates. By trial and error, you can fairly rapidly find the r that gives a desired value of U/m.

Page 85, problem 33: Use calculus to find the minimum of U.

Page 85, problem 35: The spring constant of this spring, k, is *not* the quantity you need in the equation for the period. What you need in that equation is the second derivative of the spring's energy with respect to the position of the thing that's oscillating. You need to start by finding the energy stored in the spring as a function of the vertical position, y, of the mass.

This is similar to example 22 on page 78.

Page 85, problem 37: The variables x_1 and x_2 will adjust themselves to reach an equilibrium. Write down the total energy in terms of x_1 and x_2, then eliminate one variable, and find the equilibrium value of the other. Finally, eliminate both x_1 and x_2 from the total energy, getting it just in terms of b.

Hints for Chapter 3

Page 166, problem 15: Write down two equations, one for Newton's second law applied to each object. Solve these for the two unknowns T and a.

Page 171, problem 46: You can use the geometric interpretation of the dot product.

Page 166, problem 20: The whole expression for the amplitude has maxima where the stuff inside the square root is at a minimum, and vice versa, so you can save yourself a lot of work by just working on the stuff inside the square root. For normal, large values of Q, the there are two extrema, one at $\omega = 0$ and one at resonance; one of these is a maximum and one is a minimum. You want to find out at what value of Q the zero-frequency extremum switches over from being a maximum to being a minimum.

Hints for Chapter 4

Page 221, problem 11: There are three forces on the wheel at first, but only two when it lifts off. Normal forces are always perpendicular to the surface of contact. Note that the corner of the step cannot be perfectly sharp, so the surface of contact for this force really coincides with the surface of the wheel.

Page 221, problem 12: How is this different from the case where you whirl a rock in a circle on a string and gradually pull in the string?

Page 223, problem 24: The acceleration and the tension in the string are unknown. Use $\tau = \mathrm{d}L/\mathrm{d}t$ and $F = ma$ to determine these two unknowns.

Page 224, problem 35: You'll need the result of problem 19 in order to relate the energy and momentum of a rigidly rotating body. Since this relationship involves a variable raised to a power, you can't just graph the data and get the moment of inertia directly. You can either graph the logarithms of both quantities, or manipulate one of the variables to make the graph linear.

If you want to use the latter approach, here is an example from another context. Suppose you were given a table of the masses, m, of cubical pieces of wood, whose sides had various lengths, b. You want to find a best-fit value for the density of the wood. The relationship is $m = \rho b^3$. The graph of m versus b would be a curve, and you would not have any easy way to get the density from such a graph. But by graphing m versus b^3, you can produce a graph that is linear, and whose slope equals the density.

Hints for Chapter 6

Page 302, problem 12: The answers to the two parts are not the same.

Page 302, problem 9: (a) The most straightforward approach is to apply the equation $\partial^2 y/\partial t^2 = (T/\mu)\partial^2 y/\partial x^2$. Although this equation was developed in the main text in the context of a straight string with a curvy wave on it, it works just as well for a circular loop; the left-hand side is simply the inward acceleration of any point on the rope. Note, however, that we've been assuming the string was (at least approximately) parallel to the x axis, which will only be true if you choose a specific value of x. You need to get an equation for y in terms of x in order to evaluate the right-hand side.

Hints for Chapter 7

Page 341, problem 1: Your calculator will round γ off to one. Use the approximation $(1+\epsilon)^p \approx 1 + p\epsilon$.

Page 342, problem 8: You can just find γ, and avoid finding v.

Page 343, problem 13: Write down an equation for the motion of one wavefront in the source's frame, and then a second equation for the motion of the next wavefront in the source's frame. Then transform to the observer's frame and find the separation in time between the arrival of the first and second wavefronts at the same point in the observer's frame.

Hints for Chapter 8

Page 404, problem 6: The force on the lithium ion is the vector sum of all the forces of all the quadrillions of sodium and chlorine atoms, which would obviously be too laborious to calculate. Nearly all of these forces, however, are canceled by a force from an ion on the opposite side of the lithium.

Hints for Chapter 10

Page 526, problem 11: Use the approximation $(1 + \epsilon)^p \approx 1 + p\epsilon$, which is valid for small ϵ.

Page 528, problem 22: The math is messy if you put the origin of your polar coordinates at the center of the disk. It comes out much simpler if you put the origin at the edge, right on top of the point at which we're trying to compute the voltage.

Page 529, problem 24: Since we have $t \ll r$, the volume of the membrane is essentially the same as if it was unrolled and flattened out, and the field's magnitude is nearly constant.

Page 529, problem 26: First find the energy stored in a spherical shell extending from r to $r + \mathrm{d}r$, then integrate to find the total energy.

Hints for Chapter 11

Page 604, problem 23: A stable system has low energy; energy would have to be added to change its configuration.

Page 607, problem 38: We're ignoring the fact that the light consists of little wavepackets, and imagining it as a simple sine wave. But wait, there's more good news! The energy density depends on the squares of the fields, which means the squares of some sine waves. Well, when you square a sine wave that varies from -1 to $+1$, you get a sine wave that goes from 0 to $+1$, and the average value of that sine wave is $1/2$. That means you don't have to do an integral like $U = \int(\mathrm{d}U/\mathrm{d}V)\mathrm{d}V$. All you have to do is throw in the appropriate factor of $1/2$, and you can pretend that the fields have their constant values $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{B}}$ everywhere.

Page 608, problem 39: Use Faraday's law, and choose an Ampèrian surface that is a disk of radius R sandwiched between the plates.

Page 608, problem 40: (a) Magnetic fields are created by currents, so once you've decided how currents behave under time-reversal, you can figure out how magnetic fields behave.

Answers to Self-Checks

Answers to Self-Checks for Chapter 1

Page 16: The stream has to spread out. When the velocity becomes zero, it seems like the cross-sectional area has to become infinite. In reality, this is the point where the water turns around and comes back down. The infinity isn't real; it occurs mathematically because we used a simplified model of the the stream of water, assuming, for instance, that the water's velocity is always straight up.

Page 18: A positive Δx means the object is moving in the same direction as the positive x axis. A negative Δx means it's going the opposite direction.

Page 24: It lands at the bottom of the mast, just like it would if the ship was at rest. Galilean relativity says that the experiment can't come out any differently just because the ship is in motion. From the point of view of someone on the ship, the ship is at rest, and the water is what's moving.

Page 28: Galilean relativity says that experiments can't come out differently just because they're performed while in motion. The tilting of the surface tells us the train is accelerating, but it doesn't tell us anything about the train's velocity at that instant. The person in the train might say the bottle's velocity was zero (but changing), whereas a person working in a reference frame attached to the dirt outside says it's moving; they don't agree on velocities. They *do* agree on accelerations. The person in the train has to agree that the train is accelerating, since otherwise there's no reason for the funny tilting effect.

Page 29: Yes. In U.S. currency, for instance, the quantum of money is one cent.

Answers to Self-Checks for Chapter 2

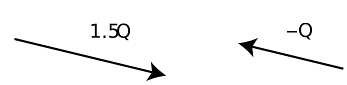
Page 51: There are two reasonable possibilities we could imagine — neither of which ends up making much sense — if we insist on the straight-line trajectory. (1) If the car has constant speed along the line, then in the * frame we see it going straight down at constant speed. It makes sense that it goes straight down in the * frame of reference, since in that frame it was never moving horizontally, and there's no reason for it to start. However, it doesn't make sense that it goes down with constant speed, since falling objects are supposed to speed up the whole time they fall. This violates both Galilean relativity and conservation of energy. (2) If it's speeding up and moving along a diagonal line in the original frame, then it might be conserving energy in one frame or the other. But if it's speeding up along a line, then as seen in the original frame, both its vertical motion and its horizontal motion must be speeding up. If its horizontal velocity is increasing in the original frame, then it can't be zero and remain zero in the * frame. This violates Galilean relativity, since in the * frame the car apparently starts moving sideways for no reason.

Answers to Self-Checks for Chapter 3

Page 112: Frictionless (or nearly frictionless) ice can certainly make a normal force, since otherwise a hockey puck would sink into the ice. Friction is not possible without a normal force, however: we can see this from the equation, or from common sense, e.g. while sliding down a rope you don't get any friction unless you grip the rope.

Page 142: $\mathbf{F} = m\mathbf{a}$

Page 143:



Answers to Self-Checks for Chapter 4

Page 190: Torques 1, 2, and 4 all have the same sign, because they are trying to twist the wrench clockwise. The sign of 3 is opposite to the signs of 1, 2, and 4. The magnitude of 3 is the greatest, since it has a large r and the force is nearly all perpendicular to the wrench. Torques

1 and 2 are the same because they have the same values of r and $F_\perp$. Torque 4 is the smallest, due to its small r.

Page 198: One person's θ-t graph would simply be shifted up or down relative to the others. The derivative equals the slope of the tangent line, and this slope isn't changed when you shift the graph, so both people would agree on the angular velocity.

Page 200: Reversing the direction of ω also reverses the direction of motion, and this is reflected by the relationship between the plus and minus signs. In the equation for the radial acceleration, ω is squared, so even if ω is negative, the result is positive. This makes sense because the acceleration is always inward in circular motion, never outward.

Page 211: All the rotations around the x axis give ω vectors along the positive x axis (thumb pointing along positive x), and all the rotations about the y axis have ω vectors with positive y components.

Page 214: For example, if we take $(\mathbf{A} \times \mathbf{B})_x = A_y B_z - B_y A_z$ and reverse the A's and B's, we get $(\mathbf{B} \times \mathbf{A})_x = B_y A_z - A_y B_z$, which is just the negative of the original expression.

Answers to Self-Checks for Chapter 5

Page 231: Solids can exert shear forces. A solid could be in an equilibrium in which the shear forces were canceling the forces due to sideways pressure gradients. For example, if I push on a brick wall, it will give by perhaps a millionth of an inch, but it will reach an equilibrium, in which the shear forces cancel out the effect of the pressure gradient.

Page 232: (1) Not valid. The equation only applies to fluids. (2) Valid. The density of the air is nearly constant between the top and bottom of the building. (3) Not valid. There is a large difference is the density of the air between the top and the bottom of the mountain. (4) Not valid, because g isn't constant throughout the interior of the earth. (5) Not valid, because the air is flowing around the wing. The air is accelerating, so it is not in equilibrium.

Page 250: Heating the gas at constant pressure requires adding heat to it, which increases its entropy. To increase the gas's pressure while keeping its temperature constant, we would have to compress it, which would give it a smaller volume to inhabit, and therefore fewer possible positions for each atom. The whole thing has to be proportional to n because entropy is additive.

Answers to Self-Checks for Chapter 6

Page 266: The leading edge is moving up, the trailing edge is moving down, and the top of the hump is motionless for one instant.

Page 286: The energy of a wave is usually proportional to the square of the amplitude. Squaring a negative number gives a positive result, so the energy is the same

Page 287: A substance is invisible to sonar if the speed of sound waves in it is the same as in water. Reflections occur only at boundaries between media in which the wave speed is different.

Page 288: No. A material object that loses kinetic energy slows down, but a wave is not a material object. The velocity of a wave ordinarily only depends on the medium, not on the amplitude. The speed of soft sound, for example, is the same as the speed of loud sound.

Page 295: No. To get the best possible interference, the thickness of the coating must be such that the second reflected wave train lags behind the first by an integer number of wavelengths. Optimal performance can therefore only be produced for one specific color of light. The typical greenish color of the coatings shows that it does the worst job for green light.

Page 296: The period is the time required to travel a distance $2L$ at speed v, $T = 2L/v$. The frequency is $f = 1/T = v/2L$.

Page 300: The wave pattern will look like this:  Three quarters of a wavelength fit in the tube, so the wavelength is three times shorter than that of the lowest-frequency mode, in which one quarter of a wave fits. Since the wavelength is smaller by a factor of three, the frequency is three times higher. Instead of $f_o,\ 2f_o,\ 3f_o,\ 4f_o,\ \ldots$, the pattern of wave frequencies of this air column goes $f_o,\ 3f_o,\ 5f_o,\ 7f_o,\ \ldots$

Answers to Self-Checks for Chapter 7

Page 311: At $v = 0$, we get $\gamma = 1$, so $t_1 = t_2$. There is no time distortion unless the two frames of reference are in relative motion.

Page 318: A relates distance to distance, so it is unitless, and similarly for D. Multiplying B by a time has to give a distance, so B has units of m/s. Multiplying C by distance has to give a time, so C has units of s/m.

Page 322: With $v = 0$, we have $\gamma = 1$, and the Lorentz transformation becomes simply $x' = x$, $y' = y$, $z' = z$, and $t' = t$. That makes sense, because we assumed the two coordinate systems coincided at $t = t' = 0$, and if they're not in relative motion, they'll continue to coincide exactly.

Answers to Self-Checks for Chapter 8

Page 353: Either type can be involved in either an attraction or a repulsion. A positive charge could be involved in either an attraction (with a negative charge) or a repulsion (with another positive), and a negative could participate in either an attraction (with a positive) or a repulsion (with a negative).

Page 354: It wouldn't make any difference. The roles of the positive and negative charges in the paper would be reversed, but there would still be a net attraction.

Page 365: Yes. In U.S. currency, the quantum of money is the penny.

Page 388: Thomson was accelerating electrons, which are negatively charged. This apparatus is supposed to accelerated atoms with one electron stripped off, which have positive net charge. In both cases, a particle that is between the plates should be attracted by the forward plate and repelled by the plate behind it.

Page 397: The hydrogen-1 nucleus is simple a proton. The binding energy is the energy required to tear a nucleus apart, but for a nucleus this simple there is nothing to tear apart.

Answers to Self-Checks for Chapter 9

Page 422: The large amount of power means a high rate of conversion of the battery's chemical energy into heat. The battery will quickly use up all its energy, i.e. "burn out."

Answers to Self-Checks for Chapter 10

Page 457: The reasoning is exactly analogous to that used in example 1 on page 455 to derive an equation for the gravitational field of the earth. The field is $F/q_t = (kQq_t/r^2)/q_t = kQ/r^2$.

Page 463:

$$\begin{aligned}E_x &= -\frac{\mathrm{d}V}{\mathrm{d}x}\\ &= -\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{kQ}{r}\right)\\ &= \frac{kQ}{r^2}\end{aligned}$$

Page 464: (a) The voltage (height) increases as you move to the east or north. If we let the positive x direction be east, and choose positive y to be north, then $\mathrm{d}V/\mathrm{d}x$ and $\mathrm{d}V/\mathrm{d}y$ are both positive. This means that E_x and E_y are both negative, which makes sense, since the water is flowing in the negative x and y directions (south and west).
(b) The electric fields are all pointing away from the higher ground. If this was an electrical map, there would have to be a large concentration of charge all along the top of the ridge, and especially at the mountain peak near the south end.

Page 476: (a) The energy density depends on $\mathbf{E}\cdot\mathbf{E}$, which equals $E_x^2+E_y^2+E_z^2$.
(b) Since E_x is squared, reversing its sign has no effect on the energy density. This makes sense, because otherwise we'd be saying that the positive and negative x axes in space were somehow physically different in their behavior, which would violate the symmetry of space.

Page 476:

$$\begin{aligned}\mathrm{N^{-1}m^{-2}C^2V^2m^{-2}m^2m} &= \mathrm{N^{-1}m^{-1}C^2V^2}\\ &= \mathrm{N^{-1}m^{-1}J^2}\\ &= \mathrm{J^{-1}J^2}\\ &= \mathrm{J}\end{aligned}$$

Page 484: Yes. The mass has the same kinetic energy regardless of which direction it's moving. Friction coverts mechanical energy into heat at the same rate whether the mass is sliding to the right or to the left. The spring has an equilibrium length, and energy can be stored in it either by compressing it ($x<0$) or stretching it ($x>0$).

Page 485: Velocity, v, is the derivative of position, x, with respect to time. This is exactly analogous to $I=\mathrm{d}q/\mathrm{d}t$.

Page 495: The impedance depends on the frequency at which the capacitor is being driven. It isn't just a single value for a particular capacitor.

Page 503: Say we're looking for $u=\sqrt{z}$, i.e., we want a number u that, multiplied by itself, equals z. Multiplication multiplies the magnitudes, so the magnitude of u can be found by taking the square root of the magnitude of z. Since multiplication also adds the arguments of the numbers, squaring a number doubles its argument. Therefore we can simply divide the argument of z by two to find the argument of u. This results in one of the square roots of z. There is another one, which is $-u$, since $(-u)^2$ is the same as u^2. This may seem a little odd: if u was chosen so that doubling its argument gave the argument of z, then how can the same be true for $-u$? Well for example, suppose the argument of z is $4\,°$. Then $\arg u=2\,°$, and $\arg(-u)=182\,°$. Doubling 182 gives 364, which is actually a synonym for 4 degrees.

Page 505: Only $\cos(6t - 4)$ can be represented by a complex number. Although the graph of $\cos^2 t$ does have a sinusoidal shape, it varies between 0 and 1, rather than -1 and 1, and there is no way to represent that using complex numbers. The function $\tan t$ doesn't even have a sinusoidal shape.

Page 512: The quantity $4\pi k q_{in}$ is now negative, so we'd better get a negative flux on the other side of Gauss' theorem. We do, because each field vector $\mathbf{E}_j$ is inward, while the corresponding area vector, $\mathbf{A}_j$, is outward. Vectors in opposite directions make negative dot products.

Answers to Self-Checks for Chapter 11

Page 549: For instance, imagine a small sphere around the negative charge, which we would sketch on the two-dimensional paper as a circle. The field points inward at every point on the sphere, so all the contributions to the flux are negative. There is no cancellation, and the total flux is negative, which is consistent with Gauss' law, since the sphere encloses a negative charge. Copying the same surface onto the field of the bar magnet, however, we find that there is inward flux on the top and outward flux on the bottom, where the surface is inside the magnet. According to Gauss' law for magnetism, these cancel exactly, which is plausible based on the figure.

Page 547: From the top panel of the figure, where the magnetic field is turned off, we can see that the beam leaves the cathode traveling upward, so in the bottom figure the electrons must be circlng in the counterclockwise direction. To produce circular motion, the force must be towards the center of the circle. We can arbitrarily pick a point on the circle at which to analyze the vectors — let's pick the right-hand side. At this point, the velocity vector of the electrons is upward. Since the electrons are negatively charged, the force $q\mathbf{v} \times \mathbf{B}$ is given by $-\mathbf{v} \times \mathbf{B}$, not $+\mathbf{v} \times \mathbf{B}$. Circular orbits are produced when the motion is in the plane perpendicular to the field, so the field must be either into or out of the page. If the field was into the page, the right-hand rule would give $\mathbf{v} \times \mathbf{B}$ to the left, which is towards the center, but the force would be in the direction of $-\mathbf{v} \times \mathbf{B}$, which would be outwards. The field must be out of the page.

Page 553: Plugging $z = 0$ into the equation gives $B_z = 4kI/c^2h$. This is simply twice the field of a single wire at a distance h. At this location, the fields contributed by the two wires are parallel, so vector addition simply gives a vector twice as strong.

Page 565: The circulation around the Ampèrian surface we used was counterclockwise, since the field on the bottom was to the right. Applying the right-hand rule, the current $I_{through}$ must have been out of the page at the top of the solenoid, and into the page at the bottom.

Page 565: The quantity ℓ came in because we set $\eta = NI/\ell$. Based on that, it's clear that ℓ represents the length of the solenoid, not the length of the wire.

Page 566: Doubling the radius of the solenoid would mean that every distance in the problem would be doubled, which would tend to make the fields weaker, since fields fall off with distance. However, doubling the radius would also mean that we had twice as much wire, and therefore twice as many moving charges to create magnetic fields. Since the magnetic field of a wire falls off like $1/r$, it's not surprising that the first effect amounts to exactly a factor of 1/2, which is exactly enough to cancel out the factor of 2 from the second effect.

Page 578: Unless the engine is already turning over, the permanent magnet isn't spinning, so there is no change in the magnetic field. Only a changing magnetic field creates an induced electric field.

Page 584: Let's get all the electrical units in terms of Teslas. Electric field units can be expressed as $\mathrm{T}\cdot\mathrm{m/s}$. The circulation of the electric field has units of electric field multiplied by distance, or $\mathrm{T}\cdot\mathrm{m}^2/\mathrm{s}$. On the right side, the derivative $\partial\mathbf{B}/\partial t$ has units of T/s, and multiplying this my area gives units of $\mathrm{T}\cdot\mathrm{m}^2/\mathrm{s}$, just like on the left side.

Answers to Self-Checks for Chapter 12

Page 618: (1) Most people would think they were positively correlated, but it's possible that they're independent. (2) These must be independent, since there is no possible physical mechanism that could make one have any effect on the other. (3) These cannot be independent, since dying today guarantees that you won't die tomorrow.

Page 621: The area under the curve from 130 to 135 cm is about 3/4 of a rectangle. The area from 135 to 140 cm is about 1.5 rectangles. The number of people in the second range is about twice as much. We could have converted these to actual probabilities (1 rectangle = 5 cm $\times$ 0.005 cm^{-1} = 0.025), but that would have been pointless because we were just going to compare the two areas.

Page 636: The axes of the graph are frequency and photon energy, so its slope is Planck's constant. It doesn't matter if you graph $e\Delta V$ rather than $W + e\Delta V$, because that only changes the y-intercept, not the slope.

Page 648: Wavelength is inversely proportional to momentum, so to produce a large wavelength we would need to use electrons with very small momenta and energies. (In practical terms, this isn't very easy to do, since ripping an electron out of an object is a violent process, and it's not so easy to calm the electrons down afterward.)

Page 658: Under the ordinary circumstances of life, the accuracy with which we can measure position and momentum of an object doesn't result in a value of $\Delta p\Delta x$ that is anywhere near the tiny order of magnitude of Planck's constant. We run up against the ordinary limitations on the accuracy of our measuring techniques long before the uncertainty principle becomes an issue.

Page 663: No. The equation $K = p^2/2m$ is nonrelativistic, so it can't be applied to an electron moving at relativistic speeds. Photons always move at relativistic speeds, so it can't be applied to them either.

Page 665: Dividing by Planck's constant, a small number, gives a large negative result inside the exponential, so the probability will be very small.

Page 668: If you trace a circle going around the center, you run into a series of eight complete wavelengths. Its angular momentum is $8\hbar$.

Page 671: $n = 3$, $\ell = 0$, $\ell_z = 0$: one state; $n = 3$, $\ell = 1$, $\ell_z = -1$, 0, or 1: three states; $n = 3$, $\ell = 2$, $\ell_z = -2$, -1, 0, 1, or 2: five states

Answers

Answers for Chapter 2

Page 85, problem 36: (b) $\omega = \sqrt{3k/m}$.

Page 85, problem 37: $K = k_1k_2/(k_1 + k_2) = 1/(1/k_1 + 1/k_2)$

Answers for Chapter 3

Page 164, problem 4: After the collision it is moving at 1/3 of its initial speed, in the same direction it was initially going (it "follows through").

Page 166, problem 20: $Q = 1/\sqrt{2}$

Page 167, problem 22: (a) 7×10^{-8} radians, or about 4×10^{-6} degrees.

Page 168, problem 27: (a) $R = (2v^2/g)\sin\theta\cos\theta$ (c) 45 °

Page 168, problem 28: (a) The optimal angle is about 40°, and the resulting range is about 122 meters, which is about the length of a home run. (b) It goes about 8 meters farther. For comparison with reality, the stadium's web site claims a home run goes about 11 meters farther there than in a sea-level stadium.

Answers for Chapter 5

Page 261, problem 6: (c) $n \approx 16$

Page 262, problem 9: (a) $\sim 2-10\%$ (b) 5% (c) The high end for the body's actual efficiency is higher than the limit imposed by the laws of thermodynamics. However, the high end of the 1-5 watt range quoted in the problem probably includes large people who aren't just lying around. Still, it's impressive that the human body comes so close to the thermodynamic limit.

Page 262, problem 10: (a) Looking up the relevant density for air, and converting everything to mks, we get a frequency of 730 Hz. This is on the right order of magnitude, which is promising, considering the crudeness of the approximation. (b) This brings the result down to 400 Hz, which is amazingly close to the observed frequency of 300 Hz.

Answers for Chapter 6

Page 302, problem 6: (b) $g/2$

Page 303, problem 13: The actual length of a clarinet is about 67 cm from the tip of the mouthpiece to the end of the bell. Because the behavior of the clarinet and its coupling to air outside it is a little more complex than that of a simple tube enclosing a cylindrical air column, your answer will be close to this value, but not exactly equal to it.

Page 302, problem 9: (a) $T = \mu\omega^2 r^2$

Page 303, problem 16: (a) $f = 4\alpha/(1+\alpha)^2$ (b) $v_2 = \sqrt{v_1 v_3}$

Answers for Chapter 7

Page 342, problem 11: (c) $p = E/c$

Answers for Chapter 10

Page 528, problem 19: (a) $E = 2k\lambda/R$.

Answers for Chapter 11

Page 600, problem 9: (a) $I = \lambda v$.

Page 601, problem 10: (b) $2kI_1I_2L/c^2R$.

Solutions

Solutions for Chapter 2

Page 80, problem 1: (a) The energy stored in the gasoline is being changed into heat via frictional heating, and also probably into sound and into energy of water waves. Note that the kinetic energy of the propeller and the boat are not changing, so they are not involved in the energy transformation. (b) The crusing speed would be greater by a factor of the cube root of 2, or about a 26% increase.

Page 80, problem 2: We don't have actual masses and velocities to plug in to the equation, but that's OK. We just have to reason in terms of ratios and proportionalities. Kinetic energy is proportional to mass and to the square of velocity, so B's kinetic energy equals

$(13.4\text{ J})(3.77)/(2.34)^2 = 9.23$ J.

Page 80, problem 3: Room temperature is about 20 °C. The fraction of the power that actually goes into heating the water is

$$\frac{(250\text{ g})/(0.24\text{ J/g}\,^\circ\text{C}) \times (100\,^\circ\text{C-}20\,^\circ\text{C})/126\text{ s}}{1.25\times10^8\text{ J/s}} = 0.53$$

So roughly half of the energy is wasted. The wasted energy might be in several forms: heating of the cup, heating of the oven itself, or leakage of microwaves from the oven.

Page 81, problem 5:

$$\begin{aligned}
E_{total,i} &= E_{total,f} \\
U_i + heat_i &= U_f + heat_f + K_f \\
\frac{1}{2}mv^2 &= U_i - U_f + heat_i - heat_f \\
&= -\Delta U - \Delta heat \\
v &= \sqrt{2\left(\frac{-\Delta U - \Delta heat}{m}\right)} \\
&= 6.4\text{ m/s}
\end{aligned}$$

Solutions for Chapter 3

Page 166, problem 17: (a) Spring constants in parallel add, so the spring constant has to be proportional to the cross-sectional area. Two springs in series give half the spring constant, three springs in series give 1/3, and so on, so the spring constant has to be inversely proportional to the length. Summarizing, we have $k \propto A/L$.
(b) With the Young's modulus, we have $k = (A/L)E$.The spring constant has units of N/m, so the units of E would have to be N/m^2.

Page 170, problem 39: We want to find out about the velocity vector $\mathbf{v}_{BG}$ of the bullet relative to the ground, so we need to add Annie's velocity relative to the ground $\mathbf{v}_{AG}$ to the bullet's velocity vector $\mathbf{v}_{BA}$ relative to her. Letting the positive x axis be east and y north, we have

$$\begin{aligned}
v_{BA,x} &= (140\text{ mi/hr})\cos 45\,^\circ \\
&= 100\text{ mi/hr} \\
v_{BA,y} &= (140\text{ mi/hr})\sin 45\,^\circ \\
&= 100\text{ mi/hr}
\end{aligned}$$

and

$$\begin{aligned}
v_{AG,x} &= 0 \\
v_{AG,y} &= 30\text{ mi/hr} \qquad .
\end{aligned}$$

The bullet's velocity relative to the ground therefore has components

$$v_{BG,x} = 100\text{ mi/hr}$$

and

$$v_{BG,y} = 130 \text{ mi/hr} \qquad .$$

Its speed on impact with the animal is the magnitude of this vector

$$\begin{aligned}
|\mathbf{v}_{BG}| &= \sqrt{(100 \text{ mi/hr})^2 + (130 \text{ mi/hr})^2} \\
&= 160 \text{ mi/hr}
\end{aligned}$$

(rounded off to two significant figures).

Page 170, problem 40: Since its velocity vector is constant, it has zero acceleration, and the sum of the force vectors acting on it must be zero. There are three forces acting on the plane: thrust, lift, and gravity. We are given the first two, and if we can find the third we can infer the plane's mass. The sum of the y components of the forces is zero, so

$$\begin{aligned}
0 &= F_{thrust,y} + F_{lift,y} + F_{g,y} \\
&= |\mathbf{F}_{thrust}| \sin\theta + |\mathbf{F}_{lift}| \cos\theta - mg \qquad .
\end{aligned}$$

The mass is

$$\begin{aligned}
m &= (|\mathbf{F}_{thrust}| \sin\theta + |\mathbf{F}_{lift}| \cos\theta)/g \\
&= 6.9 \times 10^4 \text{ kg} \qquad .
\end{aligned}$$

Page 170, problem 41: (a) Since the wagon has no acceleration, the total forces in both the x and y directions must be zero. There are three forces acting on the wagon: T, $\mathbf{F}_g$, and the normal force from the ground, $\mathbf{F}_n$. If we pick a coordinate system with x being horizontal and y vertical, then the angles of these forces measured counterclockwise from the x axis are $90\,^\circ - \phi$, $270\,^\circ$, and $90\,^\circ + \theta$, respectively. We have

$$\begin{aligned}
F_{x,total} &= T\cos(90\,^\circ - \phi) + F_g \cos(270\,^\circ) + F_n cos(90\,^\circ + \theta) \\
F_{y,total} &= T\sin(90\,^\circ - \phi) + F_g \sin(270\,^\circ) + F_n sin(90\,^\circ + \theta) \qquad ,
\end{aligned}$$

which simplifies to

$$\begin{aligned}
0 &= T\sin\phi - F_n \sin\theta \\
0 &= T\cos\phi - F_g + F_n \cos\theta.
\end{aligned}$$

The normal force is a quantity that we are not given and do not with to find, so we should choose it to eliminate. Solving the first equation for $F_n = (\sin\phi/\sin\theta)T$, we eliminate F_n from the second equation,

$$0 = T\cos\phi - F_g + T\sin\phi\cos\theta/\sin\theta$$

and solve for T, finding

$$T = \frac{F_g}{\cos\phi + \sin\phi\cos\theta/\sin\theta}$$

Multiplying both the top and the bottom of the fraction by $\sin\theta$, and using the trig identity for $\sin(\theta+\phi)$ gives the desired result,

$$T = \frac{\sin\theta}{\sin(\theta+\phi)} F_g s$$

(b) The case of $\phi = 0$, i.e. pulling straight up on the wagon, results in $T = F_g$: we simply support the wagon and it glides up the slope like a chair-lift on a ski slope. In the case of $\phi = 180° - \theta$, T becomes infinite. Physically this is because we are pulling directly into the ground, so no amount of force will suffice.

Page 170, problem 42: (a) If there was no friction, the angle of repose would be zero, so the coefficient of static friction, μ_s, will definitely matter. We also make up symbols θ, m and g for the angle of the slope, the mass of the object, and the acceleration of gravity. The forces form a triangle just like the one in example 56 on page 148, but instead of a force applied by an external object, we have static friction, which is less than $\mu_s F_n$. As in that example, $F_s = mg\sin\theta$, and $F_s < \mu_s F_n$, so

$$mg\sin\theta < \mu_s F_n \qquad .$$

From the same triangle, we have $F_n = mgcos\theta$, so

$$mg\sin\theta < \mu_s mg\cos\theta \qquad .$$

Rearranging,

$$\theta < \tan^{-1}\mu_s \qquad .$$

(b) Both m and g canceled out, so the angle of repose would be the same on an asteroid.

Solutions for Chapter 4

Page 220, problem 1: The pliers are not moving, so their angular momentum remains constant at zero, and the total torque on them must be zero. Not only that, but each half of the pliers must have zero total torque on it. This tells us that the magnitude of the torque at one end must be the same as that at the other end. The distance from the axis to the nut is about 2.5 cm, and the distance from the axis to the centers of the palm and fingers are about 8 cm. The angles are close enough to 90°that we can pretend they're 90 degrees, considering the rough nature of the other assumptions and measurements. The result is (300 N)(2.5 cm)=(F)(8 cm), or F=90 N.

Page 225, problem 37: The foot of the rod is moving in a circle relative to the center of the rod, with speed $v = \pi b/T$, and acceleration $v^2/(b/2) = (\pi^2/8)g$. This acceleration is initially upward, and is greater in magnitude than g, so the foot of the rod will lift off without dragging. We could also worry about whether the foot of the rod would make contact with the floor again before the rod finishes up flat on its back. This is a question that can be settled by graphing, or simply by inspection of figure i on page 209. The key here is that the two parts of the acceleration are both independent of m and b, so the result is univeral, and it does suffice to check a graph in a single example. In practical terms, this tells us something about how difficult the trick is to do. Because $\pi^2/8 = 1.23$ isn't much greater than unity, a hit that is just a little too weak (by a factor of $1.23^{1/2} = 1.11$) will cause a fairly obvious qualitative change in the results. This is easily observed if you try it a few times with a pencil.

Solutions for Chapter 7

Page 343, problem 15: (a) The factor of 2 comes from the reversal of the direction of the light ray's momentum. If we pick a coordinate system in which the force on the surface is in the positive direction, then $\Delta p = (-p) - p = -2p$. The question doesn't refer to any particular coordinate system, and is only talking about the magnitude of the force, so let's just say $\Delta p = 2p$. The force is $F = \Delta p/\Delta t = 2p/\Delta t = 2E/c\Delta t = 2P/c$.
(b) $mg = 2P/c$, so $m = 2P/gc$=70 nanograms.

Solutions for Chapter 9

Page 440, problem 1: $\Delta t = \Delta q/I = e/I = 0.16\ \mu\text{s}$

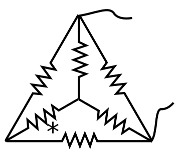
Page 442, problem 15: (a) Conservation of energy gives

$$\begin{aligned} U_A &= U_B + K_B \\ K_B &= U_A - U_B \\ \frac{1}{2}mv^2 &= e\Delta V \\ v &= \sqrt{\frac{2e\Delta V}{m}} \end{aligned}$$

(b) Plugging in numbers, we get 5.9×10^7 m/s. This is about 20% of the speed of light, so the nonrelativistic assumption was good to at least a rough approximation.

Page 443, problem 17: In series, they give 11 kΩ. In parallel, they give $(1/1\ \text{k}\Omega + 1/10\ \text{k}\Omega)^{-1} = 0.9\ \text{k}\Omega$.

Page 445, problem 27: The actual shape is irrelevant; all we care about it whats connected to what. Therefore, we can draw the circuit flattened into a plane. Every vertex of the tetrahedron is adjacent to every other vertex, so any two vertices to which we connect will give the same resistance. Picking two arbitrarily, we have this:



This is unfortunately a circuit that cannot be converted into parallel and series parts, and thats what makes this a hard problem! However, we can recognize that by symmetry, there is zero current in the resistor marked with an asterisk. Eliminating this one, we recognize the whole arrangement as a triple parallel circuit consisting of resistances R, $2R$, and $2R$. The resulting resistance is $R/2$.

Page 445, problem 31: It's much more practical to measure voltage differences. To measure a current, you have to break the circuit somewhere and insert the meter there, but it's not possible to disconnect the circuits sealed inside the board.

Solutions for Chapter 10

Page 526, problem 12: By symmetry, the field is always directly toward or away from the center. We can therefore calculate it along the x axis, where $r = x$, and the result will be valid for any location at that distance from the center.

$$\begin{aligned} E &= -\frac{\mathrm{d}V}{\mathrm{d}x} \\ &= -\frac{\mathrm{d}}{\mathrm{d}x}\left(x^{-1}e^{-x}\right) \\ &= -x^{-2}e^{-x} - x^{-1}e^{-x} \end{aligned}$$

At small x, near the proton, the first term dominates, and the exponential is essentially 1, so we have $E \propto x^{-2}$, as we expect from the Coulomb force law. At large x, the second term dominates, and the field approaches zero faster than an exponential.

Solutions for Chapter 11

Page 608, problem 40: (a) For a material object, $\mathbf{p} = m\mathbf{v}$. The velocity vector reverses itself, but mass is still positive, so the momentum vector is reversed.
(b) In the forward-time universe, conservation of momentum is $\mathbf{p}_{1,i} + \mathbf{p}_{2,i} = \mathbf{p}_{1,f} + \mathbf{p}_{2,f}$. In the backward-time universe, all the momenta are reversed, but that just negates both sides of the equation, so momentum is still conserved.

Appendix 5: Useful Data

.0.7 Notation and terminology, compared with other books

Almost all the notation and terminology in *Simple Nature* is standard, but there are some cases where there is no universal standard, and a very few cases where I've intentionally deviated from a universal standard. The notation used by physicists is also different from that used by electrical and mechanical engineers; I use physics terminology and notation (notably $\sqrt{-1} = i$, not j, and "torque" rather than "moment"), but employ the SI system of units used in engineering, rather than the cgs units favored by some physicists.

Nonstandard terminology:

Potential energy is referred to in this book as *interaction energy*, or according to its type: *gravitational energy*, *electrical energy*, etc.

The potential, in an electrical context, is referred to as *voltage*, e.g. I say that $V = kq/r$ is the voltage surrounding a point charge.

Heat and thermal energy are both referred to as *heat*. This is in keeping with casual usage among scientists, but formal written usage dictates the use of "thermal energy" to mean the kinetic energy an object has because of its molecules' random motion, while "heat" is the transfer of thermal energy.

Notation for which there is no universal standard:

Kinetic energy is written K. Standard notation is K, T, or KE.

Interaction energy is written U. Standard notation is U, V, or PE.

The unit vectors are $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$. Standard notation is either $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$ or $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$.

Distance from an axis in cylindrical coordinates is R. A more common notation in math books is ρ, but this would conflict with the standard physics notation for the charge density.

Vibrations do not have very well standardized terminology or notation. I use "frequency" to refer to both f and ω, depending on the context to make it clear which is meant. The frequency of free, damped oscillations is ω_f, which is only approximately the same as $\omega_o = \sqrt{k/m}$. The full width at half-maximum of the resonance peak (on a plot of energy versus frequency) is $\Delta\omega$.

The coupling constants for electricity and magnetism are written as k and k/c^2. This is standard notation, but it would be more common in SI calculations to see everything expressed in terms of $\epsilon_o = 1/4\pi k$ and $\mu_o = 4\pi k/c^2$. Numerically, we have $k = 8.99 \times 10^9\ \mathrm{N{\cdot}m^2/C^2}$ and $k/c^2 = 10^{-7}\ \mathrm{N/A^2}$, the latter being an exact relation.

.0.8 Notation and units

quantity	unit	symbol
distance	meter, m	$x, \Delta x$
time	second, s	$t, \Delta t$
mass	kilogram, kg	m
density	$\mathrm{kg/m^3}$	ρ
force	newton, 1 N=1 $\mathrm{kg{\cdot}m/s^2}$	F
pressure	1 Pa=1 $\mathrm{N/m^2}$	P
velocity	m/s	v
acceleration	$\mathrm{m/s^2}$	a
gravitational field	J/kg·m or $\mathrm{m/s^2}$	g
energy	joule, J	E
power	watt, 1 W = 1 J/s	P
momentum	kg·m/s	p
angular momentum	$\mathrm{kg{\cdot}m^2/s}$ or J·s	L
period	s	T
wavelength	m	λ
frequency	$\mathrm{s^{-1}}$ or Hz	f
charge	coulomb, C	q
voltage	volt, 1 V = 1 J/C	V
current	ampere, 1 A = 1 C/s	I
resistance	ohm, 1 Ω = 1 V/A	R
capacitance	farad, 1 F = 1 C/V	C
inductance	henry, 1 H = 1 V·s/A	L
electric field	V/m or N/C	E
magnetic field	tesla, 1 T = 1 N·s/C·m	B
focal length	m	f
magnification	unitless	M
index of refraction	unitless	n
electron wavefunction	$\mathrm{m^{-3/2}}$	Ψ

.0.9 Metric prefixes

M-	mega-	10^6
k-	kilo-	10^3
m-	milli-	10^{-3}
μ- (Greek mu)	micro-	10^{-6}
n-	nano-	10^{-9}
p-	pico-	10^{-12}
f-	femto-	10^{-15}

Note that the exponents go in steps of three. The exception is centi-, 10^{-2}, which is used only in the centimeter, and this doesn't require memorization, because a cent is 10^{-2} dollars.

.0.10 Nonmetric units

Nonmetric units in terms of metric ones:

1 inch	= 25.4 mm (by definition)
1 pound (lb)	= 4.5 newtons of force
1 scientific calorie	= 4.18 J
1 nutritional calorie	$= 4.18 \times 10^3$ J
1 gallon	$= 3.78 \times 10^3\ \text{cm}^3$
1 horsepower	= 746 W

The pound is a unit of force, so it converts to newtons, not kilograms. A one-kilogram mass at the earth's surface experiences a gravitational force of $(1\ \text{kg})(9.8\ \text{m/s}^2) = 9.8\ \text{N} = 2.2$ lb. The nutritional information on food packaging typically gives energies in units of calories, but those so-called calories are really kilocalories.

Relationships among U.S. units:

1 foot (ft)	= 12 inches
1 yard (yd)	= 3 feet
1 mile (mi)	= 5280 feet
1 ounce (oz)	= 1/16 pound

.0.11 The Greek alphabet

α	A	alpha	ι	I	iota	ρ	P	rho
β	B	beta	κ	K	kappa	σ	Σ	sigma
γ	Γ	gamma	λ	Λ	lambda	τ	T	tau
δ	Δ	delta	μ	M	mu	υ	Y	upsilon
ϵ	E	epsilon	ν	N	nu	ϕ	Φ	phi
ζ	Z	zeta	ξ	Ξ	xi	χ	X	chi
η	H	eta	o	O	omicron	ψ	Ψ	psi
θ	Θ	theta	π	Π	pi	ω	Ω	omega

.0.12 Fundamental constants

gravitational constant	$G = 6.67 \times 10^{-11}\ \text{J}\cdot\text{m/kg}^2$
Coulomb constant	$k = 8.99 \times 10^9\ \text{J}\cdot\text{m/C}^2$ or $\text{N}\cdot\text{m}^2/\text{C}^2$
quantum of charge	$e = 1.60 \times 10^{-19}$ C
speed of light	$c = 3.00 \times 10^8$ m/s
Planck's constant	$h = 6.63 \times 10^{-34}\ \text{J}\cdot\text{s}$

.0.13 Subatomic particles

particle	mass (kg)	charge	radius (fm)
electron	9.109×10^{-31}	$-e$	$\lesssim 0.01$
proton	1.673×10^{-27}	$+e$	~ 1.1
neutron	1.675×10^{-27}	0	~ 1.1
neutrino	$\sim 10^{-39}$ kg ?	0	?

The radii of protons and neutrons can only be given approximately, since they have fuzzy surfaces. For comparison, a typical atom is about a million fm in radius.

.0.14 Earth, moon, and sun

body	mass (kg)	radius (km)	radius of orbit (km)
earth	5.97×10^{24}	6.4×10^{3}	1.49×10^{8}
moon	7.35×10^{22}	1.7×10^{3}	3.84×10^{5}
sun	1.99×10^{30}	7.0×10^{5}	—

.0.15 The periodic table

1 H																		2 He
3 Li	4 Be												5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg												13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc		22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y		40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	57 La	*	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	89 Ac	**	104 Rf	105 Ha	106	107	108	109	110	111	112	113	114	115	116	117	118

*	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
**	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

.0.16 Atomic masses

These atomic masses are given in atomic mass units (u), where by definition the mass of an atom of the isotope carbon-12 equals 12 u. One atomic mass unit is the same as about 1.66×10^{-27} kg. Data are only given for naturally occurring elements.

Ag	107.9	Eu	152.0	Mo	95.9	Sc	45.0
Al	27.0	F	19.0	N	14.0	Se	79.0
Ar	39.9	Fe	55.8	Na	23.0	Si	28.1
As	74.9	Ga	69.7	Nb	92.9	Sn	118.7
Au	197.0	Gd	157.2	Nd	144.2	Sr	87.6
B	10.8	Ge	72.6	Ne	20.2	Ta	180.9
Ba	137.3	H	1.0	Ni	58.7	Tb	158.9
Be	9.0	He	4.0	O	16.0	Te	127.6
Bi	209.0	Hf	178.5	Os	190.2	Ti	47.9
Br	79.9	Hg	200.6	P	31.0	Tl	204.4
C	12.0	Ho	164.9	Pb	207.2	Tm	168.9
Ca	40.1	In	114.8	Pd	106.4	U	238
Ce	140.1	Ir	192.2	Pt	195.1	V	50.9
Cl	35.5	K	39.1	Pr	140.9	W	183.8
Co	58.9	Kr	83.8	Rb	85.5	Xe	131.3
Cr	52.0	La	138.9	Re	186.2	Y	88.9
Cs	132.9	Li	6.9	Rh	102.9	Yb	173.0
Cu	63.5	Lu	175.0	Ru	101.1	Zn	65.4
Dy	162.5	Mg	24.3	S	32.1	Zr	91.2
Er	167.3	Mn	54.9	Sb	121.8		

Appendix 6: Summary

Notation and units are summarized on page 719.

Chapter 1, Conservation of Mass, page 13

Conservation laws are the foundation of physics. A conservation law states that a certain quantity can be neither created nor destroyed; the total amount of it remains the same.

Mass is a conserved quantity in classical physics, i.e. physics before Einstein. This is plausible, since we know that matter is composed of subatomic particles; if the particles are neither created or destroyed, then it makes sense that the total mass will remain the same. There are two ways of defining mass.

Gravitational mass is defined by measuring the effect of gravity on a particular object, and comparing with some standard object, taking care to test both objects at a location where the strength of gravity is the same.

Inertial mass is defined by measuring how much a particular object resists a change in its state of motion. For instance, an object placed on the end of a spring will oscillate if the spring is initially compressed, and a more massive object will take longer to complete one oscillation.

Inertial and gravitational mass are equivalent: experiments show to a very high degree of precision that any two objects with the same inertial mass have the same gravitational mass as well.

The definition of inertial mass depends on a correct but counterintuitive assumption: that an object resists a change in its state of motion. Most people intuitively believe that motion has a natural tendency to slow down. This cannot be correct as a general statement, because "to slow down" is not a well-defined concept unless we specify what we are measuring motion relative to. This insight is credited to Galileo, and the general principle of *Galilean relativity* states that the laws of physics are the same in all inertial frames of reference. In other words, there is no way to distinguish a moving frame of reference from one that is at rest. To establish which frames of reference are inertial, we first must find one inertial frame in which objects appear to obey Galilean relativity. The surface of the earth is an inertial frame to a reasonably good approximation, and the frame of reference of the stars is an even better one. Once we have found one inertial frame of reference, any other frame is inertial which is moving in a straight line at constant velocity relative to the first one. For instance, if the surface of the earth is an approximately inertial frame, then a train traveling in a straight line at constant speed is also approximately an inertial frame.

The unit of mass is the kilogram, which, along with the meter and the second, forms the basis for the SI system of units (also known as the mks system). A fundamental skill in science is to know the definitions of the most common metric prefixes, which are summarized on page 719, and to be able to convert among them.

One consequence of Einstein's theory of special relativity is that *mass can be converted to energy and energy to mass.* This prediction has been verified amply by experiment. Thus the conserved quantity is not really mass but rather the total "mass-energy," $m + E/c^2$, where c is the speed of light. Since the speed of light is a large number, the E/c^2 term is ordinarily small in everyday life, which is why we can usually neglect it.

Chapter 2, Conservation of Energy, page 33

We observe that certain processes are physically impossible. For example, there is no process that can heat up an object without using up fuel or having some other side effect such as cooling a different object. We find that we can neatly separate the possible processes from the impossible by defining a single numerical quantity, called *energy*, which is conserved. Energy comes in many forms, such as heat, motion, sound, light, the energy required to melt a solid, and gravitational energy (e.g. the energy that depends on the distance between a rock and the earth). Because it has so many forms, we can arbitrarily choose one form, heat, in order to define a standard unit for our numerical scale of energy. Energy is measured in units of joules (J), and one joule can be defined as the amount of energy required in order to raise the termperature of a certain amount of water by a certain number of degrees. (The numbers are not worth memorizing.) *Power* is defined as the rate of change of energy $P = \mathrm{d}E/\mathrm{d}t$, and the unit of power is the watt, $1\ \mathrm{W} = 1\ \mathrm{J/s}$.

Once we have defined one type of energy numerically, we can perform experiments that establish the mathematical rules governing other types of energy. For example, in his paddlewheel experiment, James Joule allowed weights to drop through a certain height and spin paddlewheels inside sealed canisters of water, thereby heating the water through friction. Since in this book we define the joule unit in terms of the temperature of water, we can think of the paddlewheel experiment as establishing a rule for the *gravitational energy* of a mass which is at a certain height,

$$\mathrm{d}U_g = mg\,\mathrm{d}y \qquad ,$$

where $\mathrm{d}U_g$ is the infinitesimal change in the gravitational energy of a mass m when its height is changed by an infinitesimal amount $\mathrm{d}y$ in the vertical direction. The quantity g is called the *gravitational field*, and at the earth's surface it has a numerical value of about $10\ \mathrm{J/kg{\cdot}m}$. That is, about 10 joules of energy are required in order to raise a one-kilogram mass by one meter. (The gravitational field g also has the interpretation that when we drop an object, its acceleration, $\mathrm{d}^2y/\mathrm{d}t^2$, is equal to g.)

Using similar techniques, we find that the energy of a moving object, called its *kinetic energy*, is given by

$$K = \frac{1}{2}mv^2 \qquad ,$$

where m is its mass and v its velocity. The proportionality factor equals 1/2 exactly by the design of the SI system of units, and since the SI is based on the meter, the kilogram, and the second, the joule is considered to be a derived unit, $1\ \mathrm{J} = 1\ \mathrm{kg{\cdot}m^2/s^2}$.

When the interaction energy U has a local maximum or minimum with respect to the position of an object ($\mathrm{d}U/\mathrm{d}x = 0$), then the object is in *equilibrium* at that position. For example, if a weight is hanging from a rope, and is initially at rest at the bottom, then it must remain at rest, because this is a position of minimum gravitational energy U_g; to move, it would have to increase both its kinetic and its gravitational energy, which would violate conservation of energy, since the total energy would increase.

Since kinetic energy is independent of the direction of motion, conservation of energy is often insufficient to predict the direction of an object's motion. However, many of the physically impossible motions can be ruled out by the trick of imposing conservation of energy in some other frame of reference. By this device, we can solve the important problem of *projectile motion:* even if the projectile has horizontal motion, we can imagine ourselves in a frame of reference in which we are moving along with the projectile horizontally. In this frame of reference, the projectile has no horizontal motion, and its vertical motion has constant acceleration g. Switching back to a frame of reference in which its horizontal velocity is not zero, we find that a projectile's horizontal and vertical motions are independent, and that the horizontal motion is at constant velocity.

Even in one-dimensional motion, it is seldom possible to solve real-world problems and predict the motion of an object in closed form. However, there are straightforward numerical techniques for solving such problems.

From observations of the motion of the planets, we infer that the *gravitational interaction between any two objects* is given by $U_g = -Gm_1m_2/r$, where r is the distance between them. When the sizes of the objects are not small compared to their separation, the definition of r becomes vague; for this reason, we should interpret this fundamentally as the law governing the gravitational interactions between individual atoms. However, in the special case of a spherically symmetric mass distribution, there is a shortcut: the *shell theorem* states that the gravitational interaction between a spherically symmetric shell of mass and a particle on the outside of the shell is the same as if the shell's mass had all been concentrated at its center. An astronomical body like the earth can be broken down into concentric shells of mass, and so its gravitational interactions with external objects can also be calculated simply by using the center-to-center distance.

Energy appears to come in a bewildering variety of forms, but matter is made of atoms, and thus if we restrict ourselves to the study of mechanical systems (containing material objects, not light), all the forms of energy we observe must be explainable in terms of the behabior and interactions of atoms. Indeed, at the atomic level the picture is much simpler. Fundamentally, all the familiar forms of mechanical energy arise from either the kinetic energy of atoms or the energy they have because they interact with each other via gravitational or electrical forces. For example, when we stretch a spring, we distort the latticework of atoms in the metal, and this change in the interatomic distances involves an increase in the atoms' electrical energies.

An equilibrium is a local minimum of $U(x)$, and up close, any minimum looks like a parabola. Therefore, small oscillations around an equilibrium exhibit universal behavior, which depends only on the object's mass, m, and on the tightness of curvature of the minimum, parametrized by the quantity $k = \mathrm{d}^2U/\mathrm{d}x^2$. The oscillations are sinusoidal as a function of time, and the period is $T = 2\pi\sqrt{m/k}$, independent of amplitude. When oscillations are small enough for these statements to be good approximations, we refer to them the oscillations as *simple harmonic*.

Chapter 3, Conservation of Momentum, page 89

Since the kinetic energy of a material object depends on v^2, it isn't obvious that conservation of energy is consistent with Galilean relativity. Even if a certain mechanical system conserves energy in one frame of reference, the velocities involved will be different as measured in another frame, and therefore so will the kinetic energies. It turns out that consistency is achieved only if there is a new conservation law, conservation of *momentum*,

$$\mathbf{p} = m\mathbf{v} \qquad .$$

In one dimension, the direction of motion is described using positive and negative signs of the velocity $\mathbf{v}$, and since mass is always positive, the momentum carries the same sign. Thus conservation of momentum, unlike conservation of energy, makes direct predictions about the direction of motion. Although this line of argument was based on the assumption of a mechanical system, momentum need not be mechanical. Light has momentum.

A moving object's momentum equals the sum of the momenta of all its atoms. To avoid having to carry out this sum, we can use the concept of the *center of mass*. The center of mass can be defined as a kind of weighted average of the positions of all the atoms in the object,

$$\mathbf{x}_{cm} = \frac{\sum m_j \mathbf{x}_j}{\sum m_j} \qquad ,$$

and although the definition does involve a sum, we can often locate the center of mass by symmetry or by physically determining an object's balance point. The total momentum of the object is then given by

$$\mathbf{p}_{total} = m_{total} \mathbf{v}_{cm} \qquad .$$

The rate of transfer of momentum is called *force*, $\mathbf{F} = \mathrm{d}\mathbf{p}/\mathrm{d}t$, and is measured in units of newtons, 1 N = 1 kg·m/s^2. As a direct consequence of conservation of momentum, we have the following statements, known as *Newton's laws of motion:*

If the total force on an object is zero, it remains in the same state of motion.

$\mathbf{F} = \mathrm{d}\mathbf{p}/\mathrm{d}t$

Forces always come in pairs: if object A exerts a force on object B, then object B exerts a force on object A which is the same strength, but in the opposite direction.

Although the fundamental forces at the atomic level are gravity, electromagnetism, and nuclear forces, we use a different and more practical classification scheme in everyday situations. In this scheme, the forces between solid objects are described as follows:

A normal force, F_n*,*	is perpendicular to the surface of contact, and prevents objects from passing through each other by becoming as strong as necessary (up to the point where the objects break). "Normal" means perpendicular.
Static friction, F_s*,*	is parallel to the surface of contact, and prevents the surfaces from starting to slip by becoming as strong as necessary, up to a maximum value of $F_{s,max}$. "Static" means not moving, i.e. not slipping.
Kinetic friction, F_k*,*	is parallel to the surface of contact, and tends to slow down any slippage once it starts. "Kinetic" means moving, i.e. slipping.

Work is defined as the transfer of energy by a force. ("By a force" is meant to exclude energy transfer by heat conduction.) The *work theorem* states that when a force occurs at a single point of contact, the amount of energy transferred by that force is given by $\mathrm{d}W = \mathbf{F} \cdot \mathrm{d}\mathbf{x}$, where $\mathrm{d}\mathbf{x}$ is the distance traveled by the point of contact. The *kinetic energy theorem* is $\mathrm{d}K_{cm} = \mathbf{F}_{total} \cdot \mathrm{d}\mathbf{x}_{cm}$, where $\mathrm{d}K_{cm}$ is the change in the energy, $(1/2)mv_{cm}^2$, an object possesses due to the motion of its center of mass, $\mathbf{F}_{total}$ is the total force acting on the object, and $\mathrm{d}\mathbf{x}_{cm}$ is the distance traveled by the center of mass.

The relationship between force and interaction energy is $U = -\mathrm{d}F/\mathrm{d}x$. Any interaction can be described either by giving the force as a function of distance or the interaction energy as a function of distance; the other quantity can then be found by integration or differentiation.

An oscillator subject to friction will, if left to itself, suffer a gradual decrease in the amplitude of its motion as mechanical energy is transformed into heat. The *quality factor*, Q, is defined as the number of oscillations required for the mechanical energy to fall off by a factor of $e^{2\pi} \approx 535$. To maintain an oscillation indefinitely, an external force must do work to replace this energy. We assume for mathematical simplicity that the external force varies sinusoidally with time, $F = F_m \sin \omega t$. If this force is applied for a long time, the motion approaches a steady state, in which the oscillator's motion is sinusoidal, matching the driving force in frequency but not in phase. The amplitude of this steady-state motion, A, exhibits the phenomenon of *resonance:* the amplitude is maximized at a driving frequency which, for large Q, is essentially the same as the natural frequency of the free vibrations, ω_f (and for large Q this is also nearly the same as $\omega_o = \sqrt{k/m}$). When the energy of the steady-state oscillations is graphed as a function of frequency, both the height and the width of the resonance peak depend on Q. The peak is taller for greater Q, and its full width at half-maximum is $\Delta\omega \approx \omega_o/Q$. For small values of Q, all these approximations become worse, and at $Q < 1/2$ qualitatively different behavior sets in.

For three-dimensional motion, a moving object's motion can be described by three different velocities, $v_x = \mathrm{d}x/\mathrm{d}t$, and similarly for v_y and v_z. Thus conservation of momentum becomes three different conservation laws: conservation of $p_x = mv_x$, and so on. The principle of *rotational invariance* says that the laws of physics are the same regardless of how we change the orientation of our laboratory: there is no preferred direction in space. As a consequence of this, no matter how we choose our x, y, and z coordinate axes, we will still have conservation of p_x, p_y, and p_z. To simplify notation, we define a momentum *vector*, $\mathbf{p}$, which is a single symbol that stands for all the momentum information contained in the components p_x, p_y, and p_z. The concept of a vector is more general than its application to the momentum: any quantity that has a direction in space is considered a vector, as opposed to a *scalar* like time or temperature. The following table summarizes some vector operations.

operation	*definition*
$\lvert\mathbf{vector}\rvert$	$\sqrt{vector_x^2 + vector_y^2 + vector_z^2}$
$\mathbf{vector} + \mathbf{vector}$	Add component by component.
$\mathbf{vector} - \mathbf{vector}$	Subtract component by component.
$\mathbf{vector} \cdot scalar$	Multiply each component by the scalar.
$\mathbf{vector} \,/\, scalar$	Divide each component by the scalar.

Differentiation and integration of vectors is defined component by component.

There is only one meaningful (rotationally invariant) way of defining a multiplication of vectors whose result is a scalar, and it is known as the vector *dot product:*

$$\begin{aligned}\mathbf{b} \cdot \mathbf{c} &= b_x c_x + b_y c_y + b_z c_z \\ &= |\mathbf{b}|\,|\mathbf{c}| \cos\theta_{bc} \qquad .\end{aligned}$$

The dot product has most of the usual properties associated with multiplication, except that there is no "dot division."

Chapter 4, Conservation of Angular Momentum, page 179

Angular momentum is a conserved quantity. For motion confined to a plane, the angular momentum of a material particle is

$$L = mv_\perp r \qquad ,$$

where r is the particle's distance from the point chosen as the axis, and $v_\perp$ is the component of its velocity vector that is perpendicular to the line connecting the particle to the axis. The choice of axis is arbitrary. In a plane, only two directions of rotation are possible, clockwise and counterclockwise. One of these is considered negative and the other positive. Geometrically, angular momentum is related to rate at which area is swept out by the line segment connecting the particle to the axis.

Torque is the rate of change of angular momentum, $\tau = \mathrm{d}L/\mathrm{d}t$. The torque created by a given force can be calculated using any of the relations

$$\begin{aligned}\tau &= rF\sin\theta_{rF} \\ &= rF_\perp \\ &= r_\perp F \qquad ,\end{aligned}$$

where the subscript $\perp$ indicates a component perpendicular to the line connecting the axis to the point of application of the force.

In the special case of a *rigid body* rotating in a single plane, we define

$$\omega = \frac{\mathrm{d}\theta}{\mathrm{d}t} \qquad \text{[angular velocity]}$$

and

$$\alpha = \frac{\mathrm{d}\omega}{\mathrm{d}t} \qquad , \qquad \text{[angular acceleration]}$$

in terms of which we have

$$L = I\omega$$

and

$$\tau = I\alpha \qquad ,$$

where the *moment of inertia*, I, is defined as

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

summing over all the atoms in the object (or using calculus to perform a continuous sum, i.e. an integral). The relationship between the angular quantities and the linear ones is

$v_t = \omega r$ [tangential velocity of a point

$v_r = 0$ [radial velocity of a point

$a_t = \alpha r$. [radial acceleration of a point

at a distance r from the axis]$a_r = \omega^2 r$ [radial acceleration of a point

at a distance r from the axis]

In three dimensions, torque and angular momentum are vectors, and are expressed in terms of the vector *cross product*, which is the only rotationally invariant way of defining a multiplication of two vectors that produces a third vector:

$$\begin{aligned}\mathbf{L} &= \mathbf{r} \times \mathbf{p} \\ \boldsymbol{\tau} &= \mathbf{r} \times \mathbf{F}\end{aligned}$$

In general, the cross product of vectors $\mathbf{b}$ and $\mathbf{c}$ has magnitude

$$|\mathbf{b} \times \mathbf{c}| = |\mathbf{b}|\,|\mathbf{c}|\,\sin\theta_{bc} \qquad ,$$

which can be interpreted geometrically as the area of the parallelogram formed by the two vectors when they are placed tail-to-tail. The direction of the cross product lies along the line which is perpendicular to both vectors; of the two such directions, we choose the one that is right-handed, in the sense that if we point the fingers of the flattened right hand along $\mathbf{b}$, then bend the knuckles to point the fingers along $\mathbf{c}$, the thumb gives the direction of $\mathbf{b} \times \mathbf{c}$. In terms of components, the cross product is

$$\begin{aligned}(\mathbf{b} \times \mathbf{c})_x &= b_y c_z - c_y b_z \\ (\mathbf{b} \times \mathbf{c})_y &= b_z c_x - c_z b_x \\ (\mathbf{b} \times \mathbf{c})_z &= b_x c_y - c_x b_y\end{aligned}$$

The cross product has the disconcerting properties

$$\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a} \qquad \text{[noncommutative]}$$

and

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) \neq (\mathbf{a} \times \mathbf{b}) \times \mathbf{c} \qquad \text{[nonassociative]} \qquad ,$$

and there is no “cross-division.”

For rigid-body rotation in three dimensions, we define an angular velocity vector $\boldsymbol{\omega}$, which lies along the axis of rotation and bears a right-hand relationship to it. Except in special cases, there is no scalar moment of inertia for which $\mathbf{L} = I\boldsymbol{\omega}$; the moment of inertia must be expressed as a matrix.

Chapter 5, Thermodynamics, page 227

A fluid is any gas or liquid, but not a solid; fluids do not exhibit shear forces. A fluid in equilibrium exerts a force on any surface which is proportional to the surface’s area and perpendicular to the surface. We can therefore define a quantity called the *pressure*, P, which is ratio of force to area,

$$P = \frac{F_\perp}{A} \qquad ,$$

where the subscript $\perp$ indicates the component of the fluid’s force which is perpendicular to the surface.

Usually it is only the difference in pressure between the two sides of a surface that is physically significant. Pressure doesn’t just “press down” on things; air pressure upward under your chin is the same as air pressure downward on your shoulders. In a fluid acted on by gravity, pressure varies with depth according to the equation

$$\mathrm{d}P = -\rho\mathbf{g} \cdot \mathrm{d}\mathbf{y}.$$

This equation is only valid if the fluid is in equilibrium, and if g and r are constant with respect to height.

Temperature can be defined according to the volume of an ideal gas under conditions of standard pressure. The Kelvin scale of temperature used throughout this book equals zero at absolute zero, the temperature at which all random molecular motion ceases, and equals 273 K at the freezing point of water. We can get away with using the Celsius scale as long as we are only interested in temperature differences; a difference of 1 degree C is the same as a difference of 1 degree K.

It is an observed fact that ideal gases obey the *ideal gas law*,

$$PV = nkT \qquad ,$$

and this equation can be explained by the kinetic theory of heat, which states that the gas experts pressure on its container because its molecules are constantly in motion. In the kinetic theory of heat, the temperature of a gas is proportional to the average energy per molecule.

Not all the heat energy in an object can be extracted to do mechanical work. We therefore describe heat as a lower grade of energy than other forms of energy. Entropy is a measure of how much of a system's energy is inaccessible to being extracted, even by the most efficient heat engine; a high entropy corresponds to a low grade of energy. The change in a system's entropy when heat Q is deposited into it is

$$\Delta S = \frac{Q}{T} \qquad .$$

The efficiency of any heat engine is defined as

$$\text{efficiency} = \frac{\text{energy we get in useful form}}{\text{energy we pay for}} \qquad ,$$

and the efficiency of a Carnot engine, the most efficient of all, is

$$\text{efficiency} = 1 - \frac{T_L}{T_H} \qquad .$$

These results are all closely related. For instance, example 11 on page 242 uses $\Delta S = Q/T$ and efficiency $= 1 - T_L/T_H$ to show that a Carnot engine doesn't change the entropy of the universe.

Fundamentally, entropy is defined as the being proportional to the natural logarithm of the number of states available to a system, and the above equation then serves as a definition of temperature. The entropy of a closed system always increases; this is the second law of thermodynamics.

Chapter 6, Waves, page 263

Wave motion differs in three important ways from the motion of material objects:

Waves obey the principle of superposition. When two waves collide, they simply add together.

The medium is not transported along with the wave. The motion of any given point in the medium is a vibration about its equilibrium location, not a steady forward motion.

The velocity of a wave depends on the medium, not on the amount of energy in the wave. (For some types of waves, notably water waves, the velocity may also depend on the shape of the wave.)

Sound waves consist of increases and decreases (typically very small ones) in the density of the air. Light is a wave, but it is a vibration of electric and magnetic fields, not of any physical medium. Light can travel through a vacuum.

A periodic wave is one that creates a periodic motion in a receiver as it passes it. Such a wave has a well-defined period and frequency, and it will also have a wavelength, which is the distance in space between repetitions of the wave pattern. The velocity, frequency, and wavelength of a periodic wave are related by the equation

$$v = f\lambda \qquad .$$

A wave emitted by a moving source will undergo a *Doppler shift* in wavelength and frequency. The shifted wavelength is given by the equation

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

where v is the velocity of the waves and u is the velocity of the source, taken to be positive or negative so as to produce a Doppler-lengthened wavelength if the source is receding and a Doppler-shortened one if it approaches. A similar shift occurs if the observer is moving, and in general the Doppler shift depends approximately only on the relative motion of the source and observer if their velocities are both small compared to the waves' velocity. (This is not just approximately but exactly true for light waves, and this fact forms the basis of Einstein's theory of relativity.)

Whenever a wave encounters the boundary between two media in which its speeds are different, part of the wave is reflected and part is transmitted. The reflection is always reversed front-to-back, but may also be inverted in amplitude. Whether the reflection is inverted depends on how the wave speeds in the two media compare, e.g. a wave on a string is uninverted when it is reflected back into a segment of string where its speed is lower. The greater the difference in wave speed between the two media, the greater the fraction of the wave energy that is reflected. Surprisingly, a wave in a dense material like wood will be strongly reflected back into the wood at a wood-air boundary.

A one-dimensional wave confined by highly reflective boundaries on two sides will display motion which is periodic. For example, if both reflections are inverting, the wave will have a period equal to twice the time required to traverse the region, or to that time divided by an integer. An important special case is a sinusoidal wave; in this case, the wave forms a stationary pattern composed of a superposition of sine waves moving in opposite direction.

Chapter 7, Relativity, page 305

Einstein's *principle of relativity* states that both light and matter obey the same laws of physics in any inertial frame of reference, regardless of the frame's orientation, position, or constant-velocity motion. The laws of physics require a specific value for the speed of light in a vacuum, $c = 3.0 \times 10^8$ m/s, so this principle implies that *the speed of light must be the same in all frames of reference*, even when it seems intuitively that this is impossible because the frames are in relative motion. This strange constancy of the speed of light was experimentally supported by the 1887 Michelson-Morley experiment.

Based only on this principle, Einstein showed that *time and space are measured differently by different observers* if the observers are in motion relative to one another. A clock appears to run fastest when observed in a frame of reference moving along with it. In other frames of reference, the clock runs more slowly by a factor

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

where v is the velocity of the clock as measured in this frame of reference. Likewise, a meter stick appears to be longest to an observer moving along with it, and is contracted by the same factor γ according to observers in other frames.

To simplify notation, we adopt a system of *natural units*, in which the speed of light equals one by definition, and both times and distances are measured in units of seconds. One second of distance is how far light travels in one second. To change natural-unit equations back to metric units, we must multiply terms by factors of c as necessary in order to make the units of all the terms on both sides of the equation come out right. In *natural units*, the gamma factor, for instance, is

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

All of the equations in the remainder of this summary are expressed in natural units.

In general, the relationship between space and time coordinates as measured by different observers is given by the *Lorentz transformations:*

$$\begin{aligned} x' &= \gamma x - \gamma v t \\ t' &= -\gamma v x + \gamma t, \end{aligned}$$

where v is the velocity of the x', t' frame with respect to the x, t frame.

Some of the main implications of these equations are:

Nothing can move faster than the speed of light.

There is no well-defined concept of simultaneity for events occurring at different points in space.

Just as a rotation mixes up the x, y, and z space coordinates, a Lorentz transformation mixes time with the coordinates. Space and time are not separate; we refer to them together as *spacetime*

In the same way that a rotation preserves lengths and angles, a Lorentz transformation preserves certain relationships between events. For one thing, it preserves the *light cone*, which is defined as the set of all events which, starting from one event, can be reached only by traveling at exactly the speed of light. Only events inside each other's light cones can be causally related. More generally, the Lorentz transformations preserve the spacetime *interval* is defined as

$$\Delta x^2 + \Delta y^2 + \Delta z^2 - \Delta t^2 \qquad .$$

Other quantities besides space and time, including momentum, force, and energy, are distorted when transformed from one frame to another, just as time and space are. But some quantities, notably spacetime intervals, masses, electric charges, and the speed of light, are invariant: they are the same in all frames.

If object A moves at velocity u relative to object B, and B moves at velocity v relative to object C, the combination of the velocities, i.e. A's velocity relative to C, is not given by $u+v$ but rather by

$$v_{combined} = \frac{u+v}{1+uv} \qquad .$$

Relativistic momentum is

$$p = m\gamma v \qquad ,$$

and kinetic energy is

$$K = m(\gamma - 1) \qquad .$$

A consequence of the theory of relativity is that mass and energy do not obey separate conservation laws. Instead, the conserved quantity is the *mass-energy.* Mass and energy may be converted into each other according to the famous equation $E = mc^2$, or, in natural units

$$E = m \qquad .$$

Chapter 8, Atoms and Electromagnetism, page 349

All the forces we encounter in everyday life boil down to two basic types: gravitational forces and electrical forces. A force such as friction or a "sticky force" arises from electrical forces between individual atoms.

Just as we use the word mass to describe how strongly an object participates in gravitational forces, we use the word *charge* for the intensity of its electrical forces. There are two types of charge. Two charges of the same type repel each other, but objects whose charges are different attract each other. Charge is measured in units of coulombs (C).

Mobile charged particle model: A great many phenomena are easily understood if we imagine matter as containing two types of charged particles, which are at least partially able to move around.

Positive and negative charge: Ordinary objects that have not been specially prepared have both types of charge spread evenly throughout them in equal amounts. The object will then tend not to exert electrical forces on any other object, since any attraction due to one type of charge will be balanced by an equal repulsion from the other. (We say "tend not to" because bringing the object near an object with unbalanced amounts of charge could cause its charges to separate from each other, and the force would no longer cancel due to the unequal distances.) It therefore makes sense to describe the two types of charge using positive and negative signs, so that an unprepared object will have zero *total* charge.

The *Coulomb force law* states that the magnitude of the electrical force between two charged particles is given by

$$|F| = \frac{k|q_1||q_2|}{r^2} \qquad .$$

Conservation of charge: An even more fundamental reason for using positive and negative signs for charge is that with this definition the total charge of a closed system is a conserved quantity.

Quantization of charge: Millikan's oil drop experiment showed that the total charge of an object could only be an integer multiple of a basic unit of charge, e. This supported the idea the the “flow” of electrical charge was the motion of tiny particles rather than the motion of some sort of mysterious electrical fluid.

Einstein’s analysis of Brownian motion was the first definitive proof of the existence of atoms. Thomson’s experiments with vacuum tubes demonstrated the existence of a new type of microscopic particle with a very small ratio of mass to charge. Thomson correctly interpreted these as building blocks of matter even smaller than atoms: the first discovery of subatomic particles. These particles are called electrons.

The above experimental evidence led to the first useful model of the interior structure of atoms, called the raisin cookie model. In the raisin cookie model, an atom consists of a relatively large, massive, positively charged sphere with a certain number of negatively charged electrons embedded in it.

Rutherford and Marsden observed that some alpha particles from a beam striking a thin gold foil came back at angles up to 180 degrees. This could not be explained in the then-favored raisin-cookie model of the atom, and led to the adoption of the planetary model of the atom, in which the electrons orbit a tiny, positively-charged nucleus. Further experiments showed that the nucleus itself was a cluster of positively-charged protons and uncharged neutrons.

Radioactive nuclei are those that can release energy. The most common types of radioactivity are alpha decay (the emission of a helium nucleus), beta decay (the transformation of a neutron into a proton or vice-versa), and gamma decay (the emission of a type of very-high-frequency light). Stars are powered by nuclear fusion reactions, in which two light nuclei collide and form a bigger nucleus, with the release of energy.

Human exposure to ionizing radiation is measured in units of millirem. The typical person is exposed to about 100 mrem worth of natural background radiation per year.

Chapter 9, DC Circuits, page 407

All electrical phenomena are alike in that that arise from the presence or motion of charge. Most practical electrical devices are based on the motion of charge around a complete circuit, so that the charge can be recycled and does not hit any dead ends. The most useful measure of the flow of charge is *current*,

$$I = \frac{\mathrm{d}q}{\mathrm{d}t} \qquad .$$

An electrical device whose job is to transform energy from one form into another, e.g. a lightbulb, uses power at a rate which depends both on how rapidly charge is flowing through it and on how much work is done on each unit of charge. The latter quantity is known as the voltage difference between the point where the current enters the device and the point where the current leaves it. Since there is a type of electrical energy associated with electrical forces, the amount of work they do is equal to the difference in potential energy between the two points, and we therefore define voltage differences directly in terms of electrical energy,

$$\Delta V = \frac{\Delta U_{elec}}{q} \qquad .$$

The rate of power dissipation is

$$P = I\Delta V \qquad .$$

Many important electrical phenomena can only be explained if we understand the mechanisms of current flow at the atomic level. In metals, currents are carried by electrons, in liquids by ions. Gases are normally poor conductors unless their atoms are subjected to such intense electrical forces that the atoms become ionized.

Many substances, including all solids, respond to electrical forces in such a way that the flow of current between two points is proportional to the voltage difference between those points (assuming the voltage difference is small). Such a substance is called ohmic, and an object made out of an ohmic substance can be rated in terms of its resistance,

$$R = \frac{\Delta V}{I}$$

An important corollary is that a perfect conductor, with $R = 0$, must have constant voltage everywhere within it.

A schematic is a drawing of a circuit that standardizes and stylizes its features to make it easier to understand. Any circuit can be broken down into smaller parts. For instance, one big circuit may be understood as two small circuits in series, another as three circuits in parallel. When circuit elements are combined in parallel and in series, we have two basic rules to guide us in understanding how the parts function as a whole:

> *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.
>
> *The loop rule:* 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.

The simplest application of these rules is to pairs of resistors combined in series or parallel. In such cases, the pair of resistors acts just like a single unit with a certain resistance value, called their equivalent resistance. Resistances in series add to produce a larger equivalent resistance,

$$R = R_1 + R_2 \qquad ,$$

because the current has to fight its way through both resistances. Parallel resistors combine to produce an equivalent resistance that is smaller than either individual resistance,

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} \qquad ,$$

because the current has two different paths open to it.

An important example of resistances in parallel and series is the use of voltmeters and ammeters in resistive circuits. A voltmeter acts as a large resistance in parallel with the resistor across which the voltage drop is being measured. The fact that its resistance is not infinite means that it alters the circuit it is being used to investigate, producing a lower equivalent resistance. An ammeter acts as a small resistance in series with the circuit through which the current is to be determined. Its resistance is not quite zero, which leads to an increase in the resistance of the circuit being tested.

Chapter 10, Fields, page 451

Newton conceived of a universe where forces reached across space instantaneously, but we now know that there is a delay in time before a change in the configuration of mass and charge in one corner of the universe will make itself felt as a change in the forces experienced far away.

We imagine the outward spread of such a change as a ripple in an invisible universe-filling *field of force.*

As an alternative to our earlier energy-based definition, we can define the *gravitational field* at a given point as the force per unit mass exerted on objects inserted at that point, and likewise the *electric field* is defined as the force per unit charge. These fields are vectors, and the fields generated by multiple sources add according to the rules of vector addition.

The *relationship between the electric field and the voltage* is

$$\frac{\partial V}{\partial x} = -E_x$$
$$\frac{\partial V}{\partial y} = -E_y$$
$$\frac{\partial V}{\partial z} = -E_z \qquad ,$$

which can be notated more compactly as a gradient,

$$\mathbf{E} = -\nabla V \qquad .$$

Fields of force contain energy, and the density of energy is proportional to the square of the magnitude of the field,

$$\mathrm{d}U_g = -\frac{1}{8\pi G} g^2 \mathrm{d}v$$
$$\mathrm{d}U_e = \frac{1}{8\pi k} E^2 \mathrm{d}v$$
$$\mathrm{d}U_m \propto B^2 \mathrm{d}v$$

The equation for the energy stored in the magnetic field is given explicitly in the next chapter; for now, we only need the fact that it behaves in the same general way as the first two equations: the energy density is proportional to the square of the field. In the case of static electric fields, we can calculate potential energy either using the previous definition in terms of mechanical work or by calculating the energy stored in the fields. If the fields are not static, the old method gives incorrect results and the new one must be used.

Capacitance, C, and inductance, L, are defined as

$$U_C = \frac{1}{2C} q^2$$

and

$$U_L = \frac{L}{2} I^2 \qquad ,$$

measured in units of farads and henries, respectively. The voltage across a capacitor or inductor is given by

$$V_C = \frac{q}{C}$$

or

$$|V_L| = \left| L \frac{\mathrm{d}I}{\mathrm{d}t} \right| \qquad .$$

In the equation for the inductor, the direction of the voltage drop (plus or minus sign) is such that the inductor resists the change in current. Although the equation for the voltage across an inductor follows directly from fundamental arguments concerning the energy stored in the magnetic field, the result is a surprise: the voltage drop implies the existence of electric fields which are not created by charges. This is an *induced electric field*, discussed in more detail in the next chapter.

A series LRC circuit exhibits *oscillation*, and, if driven by an external voltage, resonates. The Q of the circuit relates to the resistance value. For large Q, the resonant frequency is

$$\omega \approx \frac{1}{\sqrt{LC}} \qquad .$$

A series RC or RL circuit exhibits exponential *decay*,

$$q = q_o \exp\left(-\frac{t}{RC}\right)$$

or

$$I = I_o \exp\left(-\frac{R}{L}t\right) \qquad ,$$

and the quantity RC or L/R is known as the time constant.

When driven by a sinusoidal AC voltage with amplitude $\tilde{V}$, a capacitor, resistor, or inductor responds with a current having amplitude

$$\tilde{I} = \frac{\tilde{V}}{Z} \qquad ,$$

where the *impedance*, Z, is a frequency-dependent quantity having units of ohms:

$$\begin{aligned} Z_C &= \frac{1}{\omega C} \\ Z_R &= R \\ Z_L &= \omega L \qquad . \end{aligned}$$

In a capacitor, the current has a phase that is $90°$ ahead of the voltage, while in an inductor the current is $90°$ behind. When a voltage source is driving a load through a transmission line, the maximum power is delivered to the load when the impedances of the line and the load are matched.

The impedances defined in this way do not combine in series and parallel according to the same rules as resistances. However, the usual series and parallel behavior can be recovered by redefining them as complex numbers:

$$\begin{aligned} Z_C &= -\frac{i}{\omega C} \\ Z_R &= R \\ Z_L &= i\omega L \end{aligned}$$

The complex phases are to be interpreted as phase relationships between the oscillating voltages and currents.

Gauss' law states that for any region of space, the flux through the surface,

$$\Phi = \sum \mathbf{E}_j \cdot \mathbf{A}_j \qquad ,$$

is related by

$$\Phi = 4\pi k q_{in}$$

to the charge enclosed within the surface.

Chapter 11, Electromagnetism, page 539

Relativity implies that there must be an interaction between moving charges and other moving charges. This *magnetic* interaction is in addition to the usual electrical one. The magnetic field can be defined in terms of the magnetic force exerted on a test charge,

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

or, alternatively, in terms of the torque on a magnetic test dipole,

$$|B| = \frac{\tau}{|\mathbf{m}_t| \sin\theta} \qquad ,$$

where θ is the angle between the dipole vector and the field. The magnetic dipole moment $\mathbf{m}$ of a loop of current has magnitude $m = IA$, and is in the (right-handed) direction perpendicular to the loop.

The magnetic field has no sources or sinks. Gauss' law for magnetism is

$$\Phi_B = 0 \qquad .$$

The external magnetic field of a long, straight wire is

$$B = \frac{2kI}{c^2 R} \qquad ,$$

forming a right-handed circular pattern around the wire.

The energy of the magnetic field is

$$\mathrm{d}U_m = \frac{c^2}{8\pi k} B^2 \mathrm{d}v \qquad .$$

The magnetic field resulting from a set of currents can be computed by finding a set of dipoles that combine to give those currents. The field of a dipole is

$$B_z = \frac{km}{c^2}\left(3\cos^2\theta - 1\right) r^{-3}$$
$$B_R = \frac{km}{c^2}\left(3\sin\theta\,\cos\theta\right) r^{-3} \qquad ,$$

which reduces to $B_z = km/c^2r^3$ in the plane perpendicular to the dipole moment. By constructing a current loop out of dipoles, one can prove the *Biot-Savart law*,

$$\mathrm{d}\mathbf{B} = \frac{kI\,\mathrm{d}\boldsymbol{\ell} \times \mathbf{r}}{c^2 r^3} \qquad ,$$

which gives the field when we integrate over a closed current loop. All of this is valid only for static magnetic fields.

Ampère's law is another way of relating static magnetic fields to the static currents that created them, and it is more easily extended to nonstatic fields than is the Biot-Savart law. Ampère's law states that the *circulation* of the magnetic field,

$$\Gamma_B = \sum \mathbf{s}_j \cdot \mathbf{B}_j \qquad ,$$

around the edge of a surface is related to the current $I_{through}$ passing through the surface,

$$\Gamma = \frac{4\pi k}{c^2} I_{through} \qquad .$$

In the general nonstatic case, the fundamental laws of physics governing electric and magnetic fields are *Maxwell's equations*, which state that for any closed surface, the fluxes through the surface are

$$\Phi_E = 4\pi k q_{in} \qquad \text{and}$$
$$\Phi_B = 0 \qquad .$$

For any surface that is not closed, the circulations around the edges of the surface are given by

$$\Gamma_E = -\frac{\partial \Phi_B}{\partial t} \qquad \text{and}$$
$$c^2\Gamma_B = \frac{\partial \Phi_E}{\partial t} + 4\pi k I_{through} \qquad .$$

The most important result of Maxwell's equations is the existence of *electromagnetic waves* which propagate at the velocity of light — that's what light is. The waves are transverse, and the electric and magnetic fields are perpendicular to each other. There are no purely electric or purely magnetic waves; their amplitudes are always related to one another by $B = E/c$. They propagate in the right-handed direction given by the cross product $\mathbf{E}\times\mathbf{B}$, and carry momentum $p = U/c$.

Chapter 12, Quantum Physics, page 613

Quantum physics differs from classical physics in many ways, the most dramatic of which is that certain processes at the atomic level, such as radioactive decay, are random rather than deterministic. There is a method to the madness, however: quantum physics still rules out any process that violates conservation laws, and it also offers methods for calculating probabilities numerically. The most important of these generic methods is the law of independent probabilities, which states that if two random events are not related in any way, then the probability that they will both occur equals the product of the two probabilities,

$$\text{probability of A and B} = P_A P_B \qquad \text{[if A and B are independent]} \qquad .$$

When discussing a random variable x that can take on a continuous range of values, we cannot assign any finite probability to any particular value. Instead, we define the probability distribution $D(x)$, defined so that its integral over some range of x gives the probability of that range.

In radioactive decay, the time that a radioactive atom has a 50% chance of surviving is called the half-life, $t_{1/2}$. The probability of surviving for two half-lives is $(1/2)(1/2) = 1/4$, and so on. In general, the probability of surviving a time t is given by

$$P_{surv}(t) = 0.5^{t/t_{1/2}} \qquad .$$

Related quantities such as the rate of decay and probability distribution for the time of decay are given by the same type of exponential function, but multiplied by certain constant factors.

Around the turn of the twentieth century, experiments began to show problems with the classical wave theory of light. In any experiment sensitive enough to detect very small amounts of light energy, it becomes clear that light energy cannot be divided into chunks smaller than a certain amount. Measurements involving the photoelectric effect demonstrate that this smallest unit of light energy equals hf, where f is the frequency of the light and h is a number known as Planck's constant. We say that light energy is quantized in units of hf, and we interpret this quantization as evidence that light has particle properties as well as wave properties. Particles of light are called photons.

The only method of reconciling the wave and particle natures of light that has stood the test of experiment is the probability interpretation: the probability that the particle is at a given location is proportional to the square of the amplitude of the wave at that location.

One important consequence of wave-particle duality is that we must abandon the concept of the path the particle takes through space. To hold on to this concept, we would have to contradict the well established wave nature of light, since a wave can spread out in every direction simultaneously.

Light is both a particle and a wave. Matter is both a particle and a wave. The equations that connect the particle and wave properties are the same in all cases:

$$E = hf$$
$$p = h/\lambda$$

Unlike the electric and magnetic fields that make up a photon-wave, the electron wavefunction is not directly measurable. Only the square of the wavefunction, which relates to probability, has direct physical significance.

A particle that is bound within a certain region of space is a standing wave in terms of quantum physics. The two equations above can then be applied to the standing wave to yield some important general observations about bound particles:

1. The particle's energy is quantized (can only have certain values).

2. The particle has a minimum energy.

3. The smaller the space in which the particle is confined, the higher its kinetic energy must be.

These immediately resolve the difficulties that classical physics had encountered in explaining observations such as the discrete spectra of atoms, the fact that atoms don't collapse by radiating away their energy, and the formation of chemical bonds.

A standing wave confined to a small space must have a short wavelength, which corresponds to a large momentum in quantum physics. Since a standing wave consists of a superposition of two traveling waves moving in opposite directions, this large momentum should actually be interpreted as an equal mixture of two possible momenta: a large momentum to the left, or a large momentum to the right. Thus it is not possible for a quantum wave-particle to be confined to a small space without making its momentum very uncertain. In general, the Heisenberg uncertainty principle states that it is not possible to know the position and momentum of a particle simultaneously with perfect accuracy. The uncertainties in these two quantities must satisfy the approximate inequality

$$\Delta p \Delta x \gtrsim h \qquad .$$

When an electron is subjected to electric forces, its wavelength cannot be constant. The "wavelength" to be used in the equation $p = h/\lambda$ should be thought of as the wavelength of the sine wave that most closely approximates the curvature of the wavefunction at a specific point.

Infinite curvature is not physically possible, so realistic wavefunctions cannot have kinks in them, and cannot just cut off abruptly at the edge of a region where the particle's energy would be insufficient to penetrate according to classical physics. Instead, the wavefunction "tails off" in the classically forbidden region, and as a consequence it is possible for particles to "tunnel" through regions where according to classical physics they should not be able to penetrate. If this quantum tunneling effect did not exist, there would be no fusion reactions to power our sun, because the energies of the nuclei would be insufficient to overcome the electrical repulsion between them.

Hydrogen, with one proton and one electron, is the simplest atom, and more complex atoms can often be analyzed to a reasonably good approximation by assuming their electrons occupy states that have the same structure as the hydrogen atom's. The electron in a hydrogen atom exchanges very little energy or angular momentum with the proton, so its energy and angular momentum are nearly constant, and can be used to classify its states. The energy of a hydrogen state depends only on its n quantum number.

In quantum physics, the angular momentum of a particle moving in a plane is quantized in units of $\hbar$. Atoms are three-dimensional, however, so the question naturally arises of how to deal with angular momentum in three dimensions. In three dimensions, angular momentum is a vector in the direction perpendicular to the plane of motion, such that the motion appears clockwise if viewed along the direction of the vector. Since angular momentum depends on both position and momentum, the Heisenberg uncertainty principle limits the accuracy with which one can know it. The most the can be known about an angular momentum vector is its magnitude and one of its three vector components, both of which are quantized in units of $\hbar$.

In addition to the angular momentum that an electron carries by virtue of its motion through space, it possesses an intrinsic angular momentum with a magnitude of $\hbar/2$. Protons and neutrons also have spins of $\hbar/2$, while the photon has a spin equal to $\hbar$.

Particles with half-integer spin obey the Pauli exclusion principle: only one such particle can exist is a given state, i.e., with a given combination of quantum numbers.

We can enumerate the lowest-energy states of hydrogen as follows:

$n=1$,	$\ell=0$,	$\ell_z=0$,	$s_z=+1/2$ or $-1/2$	two states
$n=2$,	$\ell=0$,	$\ell_z=0$,	$s_z=+1/2$ or $-1/2$	two states
$n=2$,	$\ell=1$,	$\ell_z=-1$, 0, or 1,	$s_z=+1/2$ or $-1/2$	six states
...				...

The periodic table can be understood in terms of the filling of these states. The nonreactive noble gases are those atoms in which the electrons are exactly sufficient to fill all the states up to a given n value. The most reactive elements are those with one more electron than a noble gas element, which can release a great deal of energy by giving away their high-energy electron, and those with one electron fewer than a noble gas, which release energy by accepting an electron.

Index