Two events are given as points on a graph of position versus time. Joan of Arc helps to restore Charles VII to the throne. At a later time and a different position, Joan of Arc is sentenced to death.

A change of units distorts an $x$-$t$ graph. This graph depicts exactly the same events as figure a. The only change is that the $x$ and $t$ coordinates are measured using different units, so the grid is compressed in $t$ and expanded in $x$.

A convention we'll use to represent a distortion of time and space.
same universal laws of motion to the solar system and to the earth; this contradicted Aristotle, who believed that, for example, a rock would naturally want to move in a certain special direction (down) in order to reach a certain special location (the earth's surface). Property 3 is the reason that Einstein called his theory "relativity," but Galileo and Newton believed exactly the same thing to be true, as dramatized by Galileo's run-in with the Church over the question of whether the earth could really be in motion around the sun. Property 4 would probably surprise most people only because it asserts in such a weak and specialized way something that they feel deeply must be true. The only really strange item on the list is 5, but the Hafele-Keating experiment forces it upon us.

If it were not for property 5, we could imagine that figure d would give the correct transformation between frames of reference in motion relative to one another. Let's say that observer 1, whose grid coincides with the gray rectangle, is a hitch-hiker standing by the side of a road. Event A is a raindrop hitting his head, and event B is another raindrop hitting his head. He says that A and B occur at the same location in space. Observer 2 is a motorist who drives by without stopping; to him, the passenger compartment of his car is at rest, while the asphalt slides by underneath. He says that A and B occur at different points in space, because during the time between the first raindrop and the second, the hitch-hiker has moved backward. On the other hand, observer 2 says that events A and C occur in the same place, while the hitch-hiker disagrees. The slope of the grid-lines is simply the velocity of the relative motion of each observer relative to the other.

Figure d has familiar, comforting, and eminently sensible behavior, but it also happens to be wrong, because it violates property 5. The distortion of the coordinate grid has only moved the vertical lines up and down, so both observers agree that events like B and C are simultaneous. If this was really the way things worked, then all observers could synchronize all their clocks with one another for once and for all, and the clocks would never get out of sync. This contradicts the results of the Hafele-Keating experiment, in which all three clocks were initially synchronized in Washington, but later went out of sync because of their different states of motion.

It might seem as though we still had a huge amount of wiggle room available for the correct form of the distortion. It turns out, however, that properties 1-5 are sufficient to prove that there is only one answer, which is the one found by Einstein in 1905. To see why this is, let's work by a process of elimination.

Figure e shows a transformation that might seem at first glance to be as good a candidate as any other, but it violates property 3, that motion is relative, for the following reason. In observer 2's frame of reference, some of the grid lines cross one another. This
means that observers 1 and 2 disagree on whether or not certain events are the same. For instance, suppose that event A marks the arrival of an arrow at the bull’s-eye of a target, and event B is the location and time when the bull’s-eye is punctured. Events A and B occur at the same location and at the same time. If one observer says that A and B coincide, but another says that they don’t, we have a direct contradiction. Since the two frames of reference in figure e give contradictory results, one of them is right and one is wrong. This violates property 3, because all inertial frames of reference are supposed to be equally valid. To avoid problems like this, we clearly need to make sure that none of the grid lines ever cross one another.

The next type of transformation we want to kill off is shown in figure f, in which the grid lines curve, but never cross one another. The trouble with this one is that it violates property 1, the uniformity of time and space. The transformation is unusually “twisty” at A, whereas at B it’s much more smooth. This can’t be correct, because the transformation is only supposed to depend on the relative state of motion of the two frames of reference, and that given information doesn’t single out a special role for any particular point in spacetime. If, for example, we had one frame of reference rotating relative to the other, then there would be something special about the axis of rotation. But we’re only talking about inertial frames of reference here, as specified in property 3, so we can’t have rotation; each frame of reference has to be moving in a straight line at constant speed. For frames related in this way, there is nothing that could single out an event like A for special treatment compared to B, so transformation f violates property 1.

The examples in figures e and f show that the transformation we’re looking for must be linear, meaning that it must transform lines into lines, and furthermore that it has to take parallel lines to parallel lines. Einstein wrote in his 1905 paper that “...on account of the property of homogeneity [property 1] which we ascribe to time and space, the [transformation] must be linear.”

Applying this to our diagrams, the original gray rectangle, which is a special type of parallelogram containing right angles, must be transformed into another parallelogram. There are three types of transformations, figure g, that have this property. Case I is the Galilean transformation of figure d on page 402, which we’ve already ruled out.

Case II can also be discarded. Here every point on the grid rotates counterclockwise. What physical parameter would determine the amount of rotation? The only thing that could be relevant would be $v$, the relative velocity of the motion of the two frames of reference with respect to one another. But if the angle of rotation was pro-

---

In the units that are most convenient for relativity, the transformation has symmetry about a 45-degree diagonal line.

Interpretation of the Lorentz transformation. The slope indicated in the figure gives the relative velocity of the two frames of reference. Events A and B that were simultaneous in frame 1 are not simultaneous in frame 2, where event A occurs to the right of the \( t = 0 \) line represented by the left edge of the grid, but event B occurs to its left.

Portion of the grid, proportional to \( v \), then for large enough velocities the grid would have left and right reversed, and this would violate property 4, causality: one observer would say that event A caused a later event B, but another observer would say that B came first and caused A.

The only remaining possibility is case III, which I’ve redrawn in figure h with a couple of changes. This is the one that Einstein predicted in 1905. The transformation is known as the Lorentz transformation, after Hendrik Lorentz (1853-1928), who partially anticipated Einstein’s work, without arriving at the correct interpretation. The distortion is a kind of smooching and stretching, as suggested by the hands. Also, we’ve already seen in figures a-c on page 401 that we’re free to stretch or compress everything as much as we like in the horizontal and vertical directions, because this simply corresponds to choosing different units of measurement for time and distance. In figure h I’ve chosen units that give the whole drawing a convenient symmetry about a 45-degree diagonal line. Ordinarily it wouldn’t make sense to talk about a 45-degree angle on a graph whose axes had different units. But in relativity, the symmetric appearance of the transformation tells us that space and time ought to be treated on the same footing, and measured in the same units.

As in our discussion of the Galilean transformation, slopes are interpreted as velocities, and the slope of the near-horizontal lines in figure i is interpreted as the relative velocity of the two observers. The difference between the Galilean version and the relativistic one is that now there is smooching happening from the other side as well. Lines that were vertical in the original grid, representing simultaneous events, now slant over to the right. This tells us that, as required by property 5, different observers do not agree on whether events that occur in different places are simultaneous. The Hafele-Keating experiment tells us that this non-simultaneity effect is fairly small, even when the velocity is as big as that of a passenger jet, and this is what we would have anticipated by the correspondence principle. The way that this is expressed in the graph is that if we
pick the time unit to be the second, then the distance unit turns out to be hundreds of thousands of miles. In these units, the velocity of a passenger jet is an extremely small number, so the slope \( v \) in figure i is extremely small, and the amount of distortion is tiny — it would be much too small to see on this scale.

The only thing left to determine about the Lorentz transformation is the size of the transformed parallelogram relative to the size of the original one. Although the drawing of the hands in figure h may suggest that the grid deforms like a framework made of rigid coat-hanger wire, that is not the case. If you look carefully at the figure, you’ll see that the edges of the smooshed parallelogram are actually a little longer than the edges of the original rectangle. In fact what stays the same is not lengths but areas, as proved in the caption to figure j.

**Proof that Lorentz transformations don’t change area:** We first subject a square to a transformation with velocity \( v \), and this increases its area by a factor \( R(v) \), which we want to prove equals 1. We chop the resulting parallelogram up into little squares and finally apply a \(-v\) transformation; this changes each little square’s area by a factor \( R(-v) \), so the whole figure’s area is also scaled by \( R(-v) \). The final result is to restore the square to its original shape and area, so \( R(v)R(-v) = 1 \). But \( R(v) = R(-v) \) by property 2 of spacetime on page 401, which states that all directions in space have the same properties, so \( R(v) = 1 \).

### 7.2.2 The \( \gamma \) factor

With a little algebra and geometry (homework problem 7, page 458), one can use the equal-area property to show that the factor \( \gamma \) (Greek letter gamma) defined in figure k is given by the equation

\[
\gamma = \frac{1}{\sqrt{1 - v^2}}.
\]

If you’ve had good training in physics, the first thing you probably think when you look at this equation is that it must be nonsense, because its units don’t make sense. How can we take something with units of velocity squared, and subtract it from a unitless 1? But remember that this is expressed in our special relativistic units, in which the same units are used for distance and time. We refer
The ruler is moving in frame 1, represented by a square, but at rest in frame 2, shown as a parallelogram. Each picture of the ruler is a snapshot taken at a certain moment as judged according to frame 2’s notion of simultaneity. An observer in frame 1 judges the ruler’s length instead according to frame 1’s definition of simultaneity, i.e., using points that are lined up vertically on the graph. The ruler appears shorter in the frame in which it is moving. As proved in figure m, the length contracts from \( L \) to \( L/\gamma \).

As proved in figures l and m, lengths are also distorted:

**Length contraction**

A meter-stick appears longest to an observer who is at rest relative to it. An observer moving relative to the meter-stick at \( v \) observes the stick to be shortened by a factor of \( \gamma \).

**Time dilation**

A clock runs fastest in the frame of reference of an observer who is at rest relative to the clock. An observer in motion relative to the clock at speed \( v \) perceives the clock as running more slowly by a factor of \( \gamma \).

Here’s why we care about \( \gamma \). Figure k defines it as the ratio of two times: the time between two events as expressed in one coordinate system, and the time between the same two events as measured in the other one. The interpretation is:

**Time dilation**

A clock runs fastest in the frame of reference of an observer who is at rest relative to the clock. An observer in motion relative to the clock at speed \( v \) perceives the clock as running more slowly by a factor of \( \gamma \).

To these as natural units. In this system, velocities are always unitless. This sort of thing happens frequently in physics. For instance, before James Joule discovered conservation of energy, nobody knew that heat and mechanical energy were different forms of the same thing, so instead of measuring them both in units of joules as we would do now, they measured heat in one unit (such as calories) and mechanical energy in another (such as foot-pounds). In ordinary metric units, we just need an extra conversion factor \( c \), and the equation becomes

\[
\gamma = \frac{1}{\sqrt{1 - \left(\frac{v}{c}\right)^2}}.
\]

What is \( \gamma \) when \( v = 0 \)? What does this mean? 

**Answer, p. 1062**

self-check A

What is \( \gamma \) when \( v = 0 \)? What does this mean?
Figure n shows the behavior of $\gamma$ as a function of $v$.

**Changing an equation from natural units to SI example 1**

Often it is easier to do all of our algebra in natural units, which are simpler because $c = 1$, and all factors of $c$ can therefore be omitted. For example, suppose we want to solve for $v$ in terms of $\gamma$. In natural units, we have $\gamma = \frac{1}{\sqrt{1 - v^2}}$, so $\gamma^{-2} = 1 - v^2$, and $v = \sqrt{1 - \gamma^{-2}}$.

This form of the result might be fine for many purposes, but if we wanted to find a value of $v$ in SI units, we would need to reinsert factors of $c$ in the final result. There is no need to do this throughout the whole derivation. By looking at the final result, we see that there is only one possible way to do this so that the results make sense in SI, which is to write $v = c\sqrt{1 - \gamma^{-2}}$.

**Motion of a ray of light example 2**

▷ The motion of a certain ray of light is given by the equation $x = -t$. Is this expressed in natural units, or in SI units? Convert to the other system.

▷ The equation is in natural units. It wouldn’t make sense in SI units, because we would have meters on the left and seconds on the right. To convert to SI units, we insert a factor of $c$ in the only possible place that will cause the equation to make sense: $x = -ct$.

**An interstellar road trip example 3**

Alice stays on earth while her twin Betty heads off in a spaceship for Tau Ceti, a nearby star. Tau Ceti is 12 light-years away, so even though Betty travels at 87% of the speed of light, it will take her a long time to get there: 14 years, according to Alice.

Betty experiences time dilation. At this speed, her $\gamma$ is 2.0, so that the voyage will only seem to her to last 7 years. But there is perfect symmetry between Alice’s and Betty’s frames of reference, so Betty agrees with Alice on their relative speed; Betty sees herself as being at rest, while the sun and Tau Ceti both move backward at 87% of the speed of light. How, then, can she observe Tau Ceti to get to her in only 7 years, when it should take 14 years to travel 12 light-years at this speed?
We need to take into account length contraction. Betty sees the distance between the sun and Tau Ceti to be shrunk by a factor of 2. The same thing occurs for Alice, who observes Betty and her spaceship to be foreshortened.

\textit{The correspondence principle} \textbf{example 4}

The correspondence principle requires that $\gamma$ be close to 1 for the velocities much less than $c$ encountered in everyday life. In natural units, $\gamma = (1 - v^2)^{-1/2}$. For small values of $\epsilon$, the approximation $(1 + \epsilon)^p \approx 1 + p\epsilon$ holds (see p. 1022). Applying this approximation, we find $\gamma \approx 1 + v^2/2$.

As expected, this gives approximately 1 when $v$ is small compared to 1 (i.e., compared to $c$, which equals 1 in natural units).

In problem 32 on p. 464 we rewrite this in SI units.

Figure n on p. 407 shows that the approximation is \textit{not} valid for large values of $v/c$. In fact, $\gamma$ blows up to infinity as $v$ gets closer and closer to $c$.

\textit{Large time dilation} \textbf{example 5}

The time dilation effect in the Hafele-Keating experiment was very small. If we want to see a large time dilation effect, we can’t do it with something the size of the atomic clocks they used; the kinetic energy would be greater than the total megatonnage of all the world’s nuclear arsenals. We can, however, accelerate subatomic particles to speeds at which $\gamma$ is large. For experimental particle physicists, relativity is something you do all day before heading home and stopping off at the store for milk. An early, low-precision experiment of this kind was performed by Rossi and Hall in 1941, using naturally occurring cosmic rays. Figure q shows a 1974 experiment of a similar type which verified the time dilation predicted by relativity to a precision of about one part per thousand.

Particles called muons (named after the Greek letter $\mu$, “myoo”) were produced by an accelerator at CERN, near Geneva. A muon is essentially a heavier version of the electron. Muons undergo radioactive decay, lasting an average of only 2.197 $\mu$s before they evaporate into an electron and two neutrinos. The 1974 experiment was actually built in order to measure the magnetic properties of muons, but it produced a high-precision test of time dilation as a byproduct. Because muons have the same electric charge as electrons, they can be trapped using magnetic fields. Muons were injected into the ring shown in figure q, circling around it until they underwent radioactive decay. At the speed at which these muons were traveling, they had $\gamma = 29.33$, so on the average they

---

$^2$Bailey at al., Nucl. Phys. B150(1979) 1
lasted 29.33 times longer than the normal lifetime. In other words, they were like tiny alarm clocks that self-destructed at a randomly selected time. Figure p shows the number of radioactive decays counted, as a function of the time elapsed after a given stream of muons was injected into the storage ring. The two dashed lines show the rates of decay predicted with and without relativity. The relativistic line is the one that agrees with experiment.

Colliding nuclei show relativistic length contraction.

An example of length contraction

Figure r shows an artist's rendering of the length contraction for the collision of two gold nuclei at relativistic speeds in the RHIC accelerator in Long Island, New York. The gold nuclei would appear nearly spherical (or just slightly lengthened like an American football) in frames moving along with them, but in the laboratory's frame, they both appear drastically foreshortened as they approach the point of collision. The later pictures show the nuclei merging to form a hot soup, observed at RHIC in 2010, in which the quarks are no longer confined inside the protons and neutrons.
Example 7: In the garage's frame of reference, the bus is moving, and can fit in the garage due to its length contraction. In the bus's frame of reference, the garage is moving, and can't hold the bus due to its length contraction.

The garage paradox

One of the most famous of all the so-called relativity paradoxes has to do with our incorrect feeling that simultaneity is well defined. The idea is that one could take a schoolbus and drive it at relativistic speeds into a garage of ordinary size, in which it normally would not fit. Because of the length contraction, the bus would supposedly fit in the garage. The driver, however, will perceive the garage as being contracted and thus even less able to contain the bus.

The paradox is resolved when we recognize that the concept of fitting the bus in the garage “all at once” contains a hidden assumption, the assumption that it makes sense to ask whether the front and back of the bus can simultaneously be in the garage. Observers in different frames of reference moving at high relative speeds do not necessarily agree on whether things happen simultaneously. As shown in figure s, the person in the garage's frame can shut the door at an instant B he perceives to be simultaneous with the front bumper's arrival A at the back wall of the garage, but the driver would not agree about the simultaneity of these two events, and would perceive the door as having shut long after she plowed through the back wall.
Let’s think a little more about the role of the 45-degree diagonal in the Lorentz transformation. Slopes on these graphs are interpreted as velocities. This line has a slope of 1 in relativistic units, but that slope corresponds to $c$ in ordinary metric units. We already know that the relativistic distance unit must be extremely large compared to the relativistic time unit, so $c$ must be extremely large. Now note what happens when we perform a Lorentz transformation: this particular line gets stretched, but the new version of the line lies right on top of the old one, and its slope stays the same. In other words, if one observer says that something has a velocity equal to $c$, every other observer will agree on that velocity as well. (The same thing happens with $-c$.)

**Velocities don’t simply add and subtract.**

This is counterintuitive, since we expect velocities to add and subtract in relative motion. If a dog is running away from me at 5 m/s relative to the sidewalk, and I run after it at 3 m/s, the dog’s velocity in my frame of reference is 2 m/s. According to everything we have learned about motion, the dog must have different speeds in the two frames: 5 m/s in the sidewalk’s frame and 2 m/s in mine. But velocities are measured by dividing a distance by a time, and both distance and time are distorted by relativistic effects, so we actually shouldn’t expect the ordinary arithmetic addition of velocities to hold in relativity; it’s an approximation that’s valid at velocities that are small compared to $c$.

**A universal speed limit**

For example, suppose Janet takes a trip in a spaceship, and accelerates until she is moving at 0.6$c$ relative to the earth. She then launches a space probe in the forward direction at a speed relative to her ship of 0.6$c$. We might think that the probe was then moving at a velocity of 1.2$c$, but in fact the answer is still less than $c$ (problem 1, page 457). This is an example of a more general fact about relativity, which is that $c$ represents a universal speed limit. This is required by causality, as shown in figure t.

**Light travels at $c$.**

Now consider a beam of light. We’re used to talking casually about the “speed of light,” but what does that really mean? Motion is relative, so normally if we want to talk about a velocity, we have to specify what it’s measured relative to. A sound wave has a certain speed relative to the air, and a water wave has its own speed relative to the water. If we want to measure the speed of an ocean wave, for example, we should make sure to measure it in a frame of reference at rest relative to the water. But light isn’t a vibration of a physical medium; it can propagate through the near-perfect vacuum of outer space.
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 h₁. 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.

space, as when rays of sunlight travel to earth. This seems like a paradox: light is supposed to have a specific speed, but there is no way to decide what frame of reference to measure it in. The way out of the paradox is that light must travel at a velocity equal to c. Since all observers agree on a velocity of c, regardless of their frame of reference, everything is consistent.
The Michelson-Morley experiment

The constancy of the speed of light had in fact already been observed when Einstein was an 8-year-old boy, but because nobody could figure out how to interpret it, the result was largely ignored. In 1887 Michelson and Morley set up a clever apparatus to measure any difference in the speed of light beams traveling east-west and north-south. The motion of the earth around the sun at 110,000 km/hour (about 0.01% of the speed of light) is to our west during the day. Michelson and Morley believed that light was a vibration of a mysterious medium called the ether, so they expected that the speed of light would be a fixed value relative to the ether. As the earth moved through the ether, they thought they would observe an effect on the velocity of light along an east-west line. For instance, if they released a beam of light in a westward direction during the day, they expected that it would move away from them at less than the normal speed because the earth was chasing it through the ether. They were surprised when they found that the expected 0.01% change in the speed of light did not occur.

Example 8

The ring laser gyroscope

If you’ve flown in a jet plane, you can thank relativity for helping you to avoid crashing into a mountain or an ocean. Figure v shows a standard piece of navigational equipment called a ring laser gyroscope. A beam of light is split into two parts, sent around the perimeter of the device, and reunited. Since the speed of light is constant, we expect the two parts to come back together at the same time. If they don’t, it’s evidence that the device has been rotating. The plane’s computer senses this and notes how much rotation has accumulated.

Example 9

No frequency-dependence

Relativity has only one universal speed, so it requires that all light waves travel at the same speed, regardless of their frequency and wavelength. Presently the best experimental tests of the invariance of the speed of light with respect to wavelength come from astronomical observations of gamma-ray bursts, which are sudden outpourings of high-frequency light, believed to originate from a supernova explosion in another galaxy. One such observation, in 2009, found that the times of arrival of all the different frequencies in the burst differed by no more than 2 seconds out of a total time in flight on the order of ten billion years!

3http://arxiv.org/abs/0908.1832

Section 7.2 Distortion of space and time 413
Discussion Questions

A A person in a spaceship moving at 99.99999999% of the speed of light relative to Earth shines a flashlight forward through dusty air, so the beam is visible. What does she see? What would it look like to an observer on Earth?

B A question that students often struggle with is whether time and space can really be distorted, or whether it just seems that way. Compare with optical illusions or magic tricks. How could you verify, for instance, that the lines in the figure are actually parallel? Are relativistic effects the same, or not?

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

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

E Figure r from page 409, depicting the collision of two nuclei at the RHIC accelerator, is reproduced below. What would the shapes of the two nuclei look like to a microscopic observer riding on the left-hand nucleus? To an observer riding on the right-hand one? Can they agree on what is happening? If not, why not — after all, shouldn’t they see the same thing if they both compare the two nuclei side-by-side at the same instant in time?

w/ Discussion question E: colliding nuclei show relativistic length contraction.

F If you stick a piece of foam rubber out the window of your car while driving down the freeway, the wind may compress it a little. Does it make sense to interpret the relativistic length contraction as a type of strain that pushes an object’s atoms together like this? How does this relate to discussion question E?
G  The graph shows three galaxies. The axes are drawn according to an observer at rest relative to the galaxy 2, so that that galaxy is always at the same x coordinate. Intelligent species in the three different galaxies develop radio technology independently, and at some point each begins to actively send out signals in an attempt to communicate with other civilizations. Events a, b, and c mark the points at which these signals begin spreading out across the universe at the speed of light. Find the events at which the inhabitants of galaxy 2 detect the signals from galaxies 1 and 3. According to 2, who developed radio first, 1 or 3? On top of the graph, draw a new pair of position and time axes, for the frame in which galaxy 3 is at rest. According to 3, in what order did events a, b, and c happen?

Discussion question G.

H  The machine-gunner in the figure sends out a spray of bullets. Suppose that the bullets are being shot into outer space, and that the distances traveled are trillions of miles (so that the human figure in the diagram is not to scale). After a long time, the bullets reach the points shown with dots which are all equally far from the gun. Their arrivals at those points are events A through E, which happen at different times. Sketch these events on a position-time graph. The chain of impacts extends across space at a speed greater than \( c \). Does this violate special relativity?
7.2.4 No action at a distance

The Newtonian picture

The Newtonian picture of the universe has particles interacting with each other by exerting forces from a distance, and these forces are imagined to occur without any time delay. For example, suppose that super-powerful aliens, angered when they hear disco music in our AM radio transmissions, come to our solar system on a mission to cleanse the universe of our aesthetic contamination. They apply a force to our sun, causing it to go flying out of the solar system at a gazillion miles an hour. According to Newton’s laws, the gravitational force of the sun on the earth will immediately start dropping off. This will be detectable on earth, and since sunlight takes eight minutes to get from the sun to the earth, the change in gravitational force will, according to Newton, be the first way in which earthlings learn the bad news — the sun will not visibly start receding until a little later. Although this scenario is fanciful, it shows a real feature of Newton’s laws: that information can be transmitted from one place in the universe to another with zero time delay, so that transmission and reception occur at exactly the same instant. Newton was sharp enough to realize that this required a nontrivial assumption, which was that there was some completely objective and well-defined way of saying whether two things happened at exactly the same instant. He stated this assumption explicitly: “Absolute, true, and mathematical time, of itself, and from its own nature flows at a constant rate without regard to anything external. . .”

Time delays in forces exerted at a distance

Relativity forbids Newton’s instantaneous action at a distance. For suppose that instantaneous action at a distance existed. It
would then be possible to send signals from one place in the universe to another without any time lag. This would allow perfect synchronization of all clocks. But the Hafele-Keating experiment demonstrates that clocks A and B that have been initially synchronized will drift out of sync if one is in motion relative to the other. With instantaneous transmission of signals, we could determine, without having to wait for A and B to be reunited, which was ahead and which was behind. Since they don’t need to be reunited, neither one needs to undergo any acceleration; each clock can fix an inertial frame of reference, with a velocity vector that changes neither its direction nor its magnitude. But this violates the principle that constant-velocity motion is relative, because each clock can be considered to be at rest, in its own frame of reference. Since no experiment has ever detected any violation of the relativity of motion, we conclude that instantaneous action at a distance is impossible.

Since forces can’t be transmitted instantaneously, it becomes natural to imagine force-effects spreading outward from their source like ripples on a pond, and we then have no choice but to impute some physical reality to these ripples. We call them fields, and they have their own independent existence. Gravity is transmitted through a field called the gravitational field. Besides gravity, there are other fundamental fields of force such as electricity and magnetism (). Ripples of the electric and magnetic fields turn out to be light waves. This tells us that the speed at which electric and magnetic field ripples spread must be \( c \), and by an argument similar to the one in subsection 7.2.3 the same must hold for any other fundamental field, including the gravitational field.

Fields don’t have to wiggle; they can hold still as well. The earth’s magnetic field, for example, is nearly constant, which is why we can use it for direction-finding.

Even empty space, then, is not perfectly featureless. It has measurable properties. For example, we can drop a rock in order to measure the direction of the gravitational field, or use a magnetic compass to find the direction of the magnetic field. This concept made a deep impression on Einstein as a child. He recalled that when he was five years old, the gift of a magnetic compass convinced him that there was “something behind things, something deeply hidden.”

More evidence that fields of force are real: they carry energy.

The smoking-gun argument for this strange notion of traveling force ripples comes from the fact that they carry energy. In figure 7.1, Alice and Betty hold balls A and B at some distance from one another. These balls make a force on each other; it doesn’t really matter for the sake of our argument whether this force is gravitational, electrical, or magnetic. Let’s say it’s electrical, i.e., that the balls have the kind of electrical charge that sometimes
causes your socks to cling together when they come out of the clothes dryer. We’ll say the force is repulsive, although again it doesn’t really matter.

If Alice chooses to move her ball closer to Betty’s, \( z/2 \), Alice will have to do some mechanical work against the electrical repulsion, burning off some of the calories from that chocolate cheesecake she had at lunch. This reduction in her body’s chemical energy is offset by a corresponding increase in the electrical interaction energy. Not only that, but Alice feels the resistance stiffen as the balls get closer together and the repulsion strengthens. She has to do a little extra work, but this is all properly accounted for in the interaction energy.

But now suppose, \( z/3 \), that Betty decides to play a trick on Alice by tossing \( B \) far away just as Alice is getting ready to move \( A \). We have already established that Alice can’t feel \( B \)’s motion instantaneously, so the electric forces must actually be propagated by an electric field. Of course this experiment is utterly impractical, but suppose for the sake of argument that the time it takes the change in the electric field to propagate across the diagram is long enough so that Alice can complete her motion before she feels the effect of \( B \)’s disappearance. She is still getting stale information about \( B \)’s position. As she moves \( A \) to the right, she feels a repulsion, because the field in her region of space is still the field caused by \( B \) in its old position. She has burned some chocolate cheesecake calories, and it appears that conservation of energy has been violated, because these calories can’t be properly accounted for by any interaction with \( B \), which is long gone.

If we hope to preserve the law of conservation of energy, then the only possible conclusion is that the electric field itself carries away the cheesecake energy. In fact, this example represents an impractical method of transmitting radio waves. Alice does work on charge \( A \), and that energy goes into the radio waves. Even if \( B \) had never existed, the radio waves would still have carried energy, and Alice would still have had to do work in order to create them.

**Discussion Questions**

**A** Amy and Bill are flying on spaceships in opposite directions at such high velocities that the relativistic effect on time’s rate of flow is easily noticeable. Motion is relative, so Amy considers herself to be at rest and Bill to be in motion. She says that time is flowing normally for her, but Bill is slow. But Bill can say exactly the same thing. How can they both think the other is slow? Can they settle the disagreement by getting on the radio and seeing whose voice is normal and whose sounds slowed down and Darth-Vadery?
B The figure shows a famous thought experiment devised by Einstein. A train is moving at constant velocity to the right when bolts of lightning strike the ground near its front and back. Alice, standing on the dirt at the midpoint of the flashes, observes that the light from the two flashes arrives simultaneously, so she says the two strikes must have occurred simultaneously. Bob, meanwhile, is sitting aboard the train, at its middle. He passes by Alice at the moment when Alice later figures out that the flashes happened. Later, he receives flash 2, and then flash 1. He infers that since both flashes traveled half the length of the train, flash 2 must have occurred first. How can this be reconciled with Alice's belief that the flashes were simultaneous? Explain using a graph.

C Resolve the following paradox by drawing a spacetime diagram (i.e., a graph of $x$ versus $t$). Andy and Beth are in motion relative to one another at a significant fraction of $c$. As they pass by each other, they exchange greetings, and Beth tells Andy that she is going to blow up a stick of dynamite one hour later. One hour later by Andy's clock, she still hasn’t exploded the dynamite, and he says to himself, “She hasn’t exploded it because of time dilation. It's only been 40 minutes for her.” He now accelerates suddenly so that he's moving at the same velocity as Beth. The time dilation no longer exists. If he looks again, does he suddenly see the flash from the explosion? How can this be? Would he see her go through 20 minutes of her life in fast-motion?

D Use a graph to resolve the following relativity paradox. Relativity says that in one frame of reference, event A could happen before event B, but in someone else's frame B would come before A. How can this be? Obviously the two people could meet up at A and talk as they cruised past each other. Wouldn’t they have to agree on whether B had already happened?

E The rod in the figure is perfectly rigid. At event A, the hammer strikes one end of the rod. At event B, the other end moves. Since the rod is perfectly rigid, it can’t compress, so A and B are simultaneous. In frame 2, B happens before A. Did the motion at the right end cause the person on the left to decide to pick up the hammer and use it?

7.2.5 The light cone

Given an event P, we can now classify all the causal relationships in which P can participate. In Newtonian physics, these relationships fell into two classes: P could potentially cause any event that lay in its future, and could have been caused by any event in its past. In relativity, we have a three-way distinction rather than a two-way one. There is a third class of events that are too far away
from \( P \) in space, and too close in time, to allow any cause and effect relationship, since causality’s maximum velocity is \( c \). Since we’re working in units in which \( c = 1 \), the boundary of this set is formed by the lines with slope \( \pm 1 \) on a \((t, x)\) plot. This is referred to as the light cone, for reasons that become more visually obvious when we consider more than one spatial dimension, figure \text{ac}.

Events lying inside one another’s light cones are said to have a timelike relationship. Events outside each other’s light cones are spacelike in relation to one another, and in the case where they lie on the surfaces of each other’s light cones the term is lightlike.

7.2.6 ★ The spacetime interval

The light cone is an object of central importance in both special and general relativity. It relates the geometry of spacetime to possible cause-and-effect relationships between events. This is fundamentally how relativity works: it’s a geometrical theory of causality.

These ideas naturally lead us to ask what fruitful analogies we can form between the bizarre geometry of spacetime and the more familiar geometry of the Euclidean plane. The light cone cuts spacetime into different regions according to certain measurements of relationships between points (events). Similarly, a circle in Euclidean geometry cuts the plane into two parts, an interior and an exterior, according to the measurement of the distance from the circle’s center. A circle stays the same when we rotate the plane. A light cone stays the same when we change frames of reference. Let’s build up the analogy more explicitly.

**Measurement in Euclidean geometry**

We say that two line segments are congruent, \( AB \cong CD \), if the distance between points \( A \) and \( B \) is the same as the distance between \( C \) and \( D \), as measured by a rigid ruler.

**Measurement in spacetime**

We define \( AB \cong CD \) if:

1. \( AB \) and \( CD \) are both spacelike, and the two distances are equal as measured by a rigid ruler, in a frame where the two events touch the ruler simultaneously.

2. \( AB \) and \( CD \) are both timelike, and the two time intervals are equal as measured by clocks moving inertially.

3. \( AB \) and \( CD \) are both lightlike.

The three parts of the relativistic version each require some justification.

Case 1 has to be the way it is because space is part of spacetime. In special relativity, this space is Euclidean, so the definition of congruence has to agree with the Euclidean definition, in the case
where it is possible to apply the Euclidean definition. The spacelike relation between the points is both necessary and sufficient to make this possible. If points A and B are spacelike in relation to one another, then a frame of reference exists in which they are simultaneous, so we can use a ruler that is at rest in that frame to measure their distance. If they are lightlike or timelike, then no such frame of reference exists. For example, there is no frame of reference in which Charles VII’s restoration to the throne is simultaneous with Joan of Arc’s execution, so we can’t arrange for both of these events to touch the same ruler at the same time.

The definition in case 2 is the only sensible way to proceed if we are to respect the symmetric treatment of time and space in relativity. The timelike relation between the events is necessary and sufficient to make it possible for a clock to move from one to the other. It makes a difference that the clocks move inertially, because the twins in example 3 on p. 407 disagree on the clock time between the traveling twin’s departure and return.

Case 3 may seem strange, since it says that any two lightlike intervals are congruent. But this is the only possible definition, because this case can be obtained as a limit of the timelike one. Suppose that AB is a timelike interval, but in the planet earth’s frame of reference it would be necessary to travel at almost the speed of light in order to reach B from A. The required speed is less than c (i.e., less than 1) by some tiny amount \( \epsilon \). In the earth’s frame, the clock referred to in the definition suffers extreme time dilation. The time elapsed on the clock is very small. As \( \epsilon \) approaches zero, and the relationship between A and B approaches a lightlike one, this clock time approaches zero. In this sense, the relativistic notion of “distance” is very different from the Euclidean one. In Euclidean geometry, the distance between two points can only be zero if they are the same point.

The case splitting involved in the relativistic definition is a little ugly. Having worked out the physical interpretation, we can now consolidate the definition in a nicer way by appealing to Cartesian coordinates.

**Cartesian definition of distance in Euclidean geometry**
Given a vector \((\Delta x, \Delta y)\) from point A to point B, the square of the distance between them is defined as \(AB^2 = \Delta x^2 + \Delta y^2\).

**Definition of the interval in relativity**
Given points separated by coordinate differences \(\Delta x, \Delta y, \Delta z,\) and \(\Delta t\), the spacetime interval \(\mathcal{I}\) (cursive letter “I”) between them is defined as \(\mathcal{I} = \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2\).

This is stated in natural units, so all four terms on the right-hand side have the same units; in metric units with \(c \neq 1\), appropriate factors of \(c\) should be inserted in order to make the units of the
The interval $I$ is positive if $AB$ is timelike (regardless of which event comes first), zero if lightlike, and negative if spacelike. Since $I$ can be negative, we can’t in general take its square root and define a real number $AB$ as in the Euclidean case. When the interval is timelike, we can interpret $\sqrt{I}$ as a time, and when it’s spacelike we can take $\sqrt{-I}$ to be a distance.

The Euclidean definition of distance (i.e., the Pythagorean theorem) is useful because it gives the same answer regardless of how we rotate the plane. Although it is stated in terms of a certain coordinate system, its result is unambiguously defined because it is the same regardless of what coordinate system we arbitrarily pick. Similarly, $I$ is useful because, as proved in example 11 below, it is the same regardless of our frame of reference, i.e., regardless of our choice of coordinates.

**Example 10**

The Pioneer 10 space probe was launched in 1972, and in 1973 was the first craft to fly by the planet Jupiter. It crossed the orbit of the planet Neptune in 1983, after which telemetry data were received until 2002. The following table gives the spacecraft’s position relative to the sun at exactly midnight on January 1, 1983 and January 1, 1995. The 1983 date is taken to be $t = 0$.

<table>
<thead>
<tr>
<th>$t$ (s)</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$1.784 \times 10^{12}$ m</td>
<td>$3.951 \times 10^{12}$ m</td>
<td>$0.237 \times 10^{12}$ m</td>
</tr>
<tr>
<td>$3.7869120000 \times 10^8$ s</td>
<td>$2.420 \times 10^{12}$ m</td>
<td>$8.827 \times 10^{12}$ m</td>
<td>$0.488 \times 10^{12}$ m</td>
</tr>
</tbody>
</table>

Compare the time elapsed on the spacecraft to the time in a frame of reference tied to the sun.

We can convert these data into natural units, with the distance unit being the second (i.e., a light-second, the distance light travels in one second) and the time unit being seconds. Converting and carrying out this subtraction, we have:

<table>
<thead>
<tr>
<th>$\Delta t$ (s)</th>
<th>$\Delta x$</th>
<th>$\Delta y$</th>
<th>$\Delta z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3.7869120000 \times 10^8$ s</td>
<td>$0.2121 \times 10^4$ s</td>
<td>$1.626 \times 10^4$ s</td>
<td>$0.084 \times 10^4$ s</td>
</tr>
</tbody>
</table>

Comparing the exponents of the temporal and spatial numbers, we can see that the spacecraft was moving at a velocity on the order of $10^{-4}$ of the speed of light, so relativistic effects should be small but not completely negligible.

Since the interval is timelike, we can take its square root and interpret it as the time elapsed on the spacecraft. The result is $\sqrt{I} = 3.786911996 \times 10^8$ s. This is 0.4 s less than the time elapsed in the sun’s frame of reference.

**Example 11**

In this example we prove that the interval is the same regard-
less of what frame of reference we compute it in. This is called “Lorentz invariance.” The proof is limited to the timelike case. Given events A and B, construct the light-rectangle as defined in figure ad/1. On p. 405 we proved that the Lorentz transformation doesn’t change the area of a shape in the $x$-$t$ plane. Therefore the area of this rectangle is unchanged if we switch to the frame of reference ad/2, in which A and B occurred at the same location and were separated by a time interval $\tau$. This area equals half the interval $\mathcal{I}$ between A and B. But a straightforward calculation shows that the rectangle in ad/1 also has an area equal to half the interval calculated in that frame. Since the area in any frame equals half the interval, and the area is the same in all frames, the interval is equal in all frames as well.

ad / Light-rectangles, example 11.
1. The gray light-rectangle represents the set of all events such as P that could be visited after A and before B.
2. The rectangle becomes a square in the frame in which A and B occur at the same location in space.
3. The area of the dashed square is $\tau^2$, so the area of the gray square is $\tau^2/2$. 

ae / Example 12.
A numerical example of invariance

Figure a.e shows two frames of reference in motion relative to one another at v = 3/5. (For this velocity, the stretching and squishing of the main diagonals are both by a factor of 2.) Events are marked at coordinates that in the frame represented by the square are

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

The interval between these events is \(13^2 - 11^2 = 48\). In the frame represented by the parallelogram, the same two events lie at coordinates

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

Calculating the interval using these values, the result is \(8^2 - 4^2 = 48\), which comes out the same as in the other frame.
7.2.7  * Four-vectors and the inner product

Example 10 makes it natural that we define a type of vector with four components, the first one relating to time and the others being spatial. These are known as four-vectors. It’s clear how we should define the equivalent of a dot product in relativity:

\[ \mathbf{A} \cdot \mathbf{B} = A_t B_t - A_x B_x - A_y B_y - A_z B_z \]

The term “dot product” has connotations of referring only to three-vectors, so the operation of taking the scalar product of two four-vectors is usually referred to instead as the “inner product.” The spacetime interval can then be thought of as the inner product of a four-vector with itself. We care about the relativistic inner product for exactly the same reason we care about its Euclidean version; both are scalars, so they have a fixed value regardless of what coordinate system we choose.

Alice and Betty are identical twins. Betty goes on a space voyage at relativistic speeds, traveling away from the earth and then turning around and coming back. Meanwhile, Alice stays on earth. When Betty returns, she is younger than Alice because of relativistic time dilation (example 3, p. 407).

But isn’t it valid to say that Betty’s spaceship is standing still and the earth moving? In that description, wouldn’t Alice end up younger and Betty older? This is referred to as the “twin paradox.” It can’t really be a paradox, since it’s exactly what was observed in the Hafele-Keating experiment (p. 397).

Betty’s track in the \( x-t \) plane (her “world-line” in relativistic jargon) consists of vectors \( \mathbf{b} \) and \( \mathbf{c} \) strung end-to-end (figure af). We could adopt a frame of reference in which Betty was at rest during \( \mathbf{b} \) (i.e., \( b_x = 0 \)), but there is no frame in which \( \mathbf{b} \) and \( \mathbf{c} \) are parallel, so there is no frame in which Betty was at rest during both \( \mathbf{b} \) and \( \mathbf{c} \). This resolves the paradox.

We have already established by other methods that Betty ages less than Alice, but let’s see how this plays out in a simple numerical example. Omitting units and making up simple numbers, let’s say that the vectors in figure af are

\[ \mathbf{a} = (6, 1) \]
\[ \mathbf{b} = (3, 2) \]
\[ \mathbf{c} = (3, -1), \]

where the components are given in the order \( (t, x) \). The time experienced by Alice is then

\[ |a| = \sqrt{6^2 - 1^2} = 5.9, \]
which is greater than the Betty’s elapsed time

$$|b| + |c| = \sqrt{3^2 - 2^2} + \sqrt{3^2 - (-1)^2} = 5.1.$$
would depend on the relative velocities of three different objects: the source, the wave’s medium, and the receiver. Relativistically, things get simpler, because light isn’t a vibration of a physical medium, so the Doppler shift can only depend on a single velocity \( v \), which is the rate at which the separation between the source and the receiver is increasing.

The square in figure \( \text{aj} \) is the “graph paper” used by someone who considers the source to be at rest, while the parallelogram plays a similar role for the receiver. The figure is drawn for the case where \( v = 3/5 \) (in units where \( c = 1 \)), and in this case the stretch factor of the long diagonal is 2. To keep the area the same, the short diagonal has to be squished to half its original size. But now it’s a matter of simple geometry to show that \( OP \) equals half the width of the square, and this tells us that the Doppler shift is a factor of 1/2 in frequency. That is, the squish factor of the short diagonal is interpreted as the Doppler shift. To get this as a general equation for velocities other than 3/5, one can show by straightforward fiddling with the result of part c of problem 7 on p. 458 that the Doppler shift is

\[
D(v) = \sqrt{\frac{1 - v}{1 + v}}.
\]

Here \( v > 0 \) is the case where the source and receiver are getting farther apart, \( v < 0 \) the case where they are approaching. (This is the opposite of the sign convention used in subsection 6.1.5. It is convenient to change conventions here so that we can use positive values of \( v \) in the case of cosmological red-shifts, which are the most important application.)

Suppose that Alice stays at home on earth while her twin Betty takes off in her rocket ship at 3/5 of the speed of light. When I first learned relativity, the thing that caused me the most pain was understanding how each observer could say that the other was the one whose time was slow. It seemed to me that if I could take a pill that would speed up my mind and my body, then naturally I would see everybody else as being slow. Shouldn’t the same apply to relativity? But suppose Alice and Betty get on the radio and try to settle who is the fast one and who is the slow one. Each twin’s voice sounds sloooowed doooowwwwn to the other. If Alice claps her hands twice, at a time interval of one second by her clock, Betty hears the hand-claps coming over the radio two seconds apart, but the situation is exactly symmetric, and Alice hears the same thing if Betty claps. Each twin analyzes the situation using a diagram identical to \( \text{aj} \), and attributes her sister’s observations to a complicated combination of time distortion, the time taken by the radio signals to propagate, and the motion of her twin relative to her.
self-check B
Turn your book upside-down and reinterpret figure aj.  ▷ Answer, p. 1062

A symmetry property of the Doppler effect  example 15
Suppose that A and B are at rest relative to one another, but C is moving along the line between A and B. A transmits a signal to C, who then retransmits it to B. The signal accumulates two Doppler shifts, and the result is their product \( D(v)D(-v) \). But this product must equal 1, so we must have \( D(-v)D(v) = 1 \), which can be verified directly from the equation.

The Ives-Stilwell experiment  example 16
The result of example 15 was the basis of one of the earliest laboratory tests of special relativity, by Ives and Stilwell in 1938. They observed the light emitted by excited by a beam of \( \text{H}_2^+ \) and \( \text{H}_3^+ \) ions with speeds of a few tenths of a percent of \( c \). Measuring the light from both ahead of and behind the beams, they found that the product of the Doppler shifts \( D(v)D(-v) \) was equal to 1, as predicted by relativity. If relativity had been false, then one would have expected the product to differ from 1 by an amount that would have been detectable in their experiment. In 2003, Saathoff et al. carried out an extremely precise version of the Ives-Stilwell technique with \( \text{Li}^+ \) ions moving at 6.4% of \( c \). The frequencies observed, in units of MHz, were:

\[
\begin{align*}
    f_o & = 546466918.8 \pm 0.4 \\
    (\text{unshifted frequency}) \\
    f_o D(-v) & = 582490203.44 \pm 0.09 \\
    (\text{shifted frequency, forward}) \\
    f_o D(v) & = 512671442.9 \pm 0.5 \\
    (\text{shifted frequency, backward}) \\
    \sqrt{f_o D(-v) \cdot f_o D(v)} & = 546466918.6 \pm 0.3
\end{align*}
\]

The results show incredibly precise agreement between \( f_o \) and \( \sqrt{f_o D(-v) \cdot f_o D(v)} \), as expected relativistically because \( D(v)D(-v) \) is supposed to equal 1. The agreement extends to 9 significant figures, whereas if relativity had been false there should have been a relative disagreement of about \( v^2 = .004 \), i.e., a discrepancy in the third significant figure. The spectacular agreement with theory has made this experiment a lightning rod for anti-relativity kooks.

We saw on p. 411 that relativistic velocities should not be expected to be exactly additive, and problem 1 on p. 457 verifies this in the special case where A moves relative to B at 0.6\( c \) and B relative to C at 0.6\( c \) — the result not being 1.2\( c \). The relativistic Doppler shift provides a simple way of deriving a general equation for the relativistic combination of velocities; problem 17 on p. 461 guides you through the steps of this derivation, and the result is given on p. 1046.
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 equivalence of energy and mass, as described by the famous equation \( E = mc^2 \).

7.3.1 Momentum

Consider the following scheme for traveling faster than the speed of light. The basic idea can be demonstrated by dropping a ping-pong ball and a baseball stacked on top of each other like a snowman. They separate slightly in mid-air, and the baseball therefore has time to hit the floor and rebound before it collides with the ping-pong ball, which is still on the way down. The result is a surprise if you haven’t seen it before: the ping-pong ball flies off at high speed and hits the ceiling! A similar fact is known to people who investigate the scenes of accidents involving pedestrians. If a car moving at 90 kilometers per hour hits a pedestrian, the pedestrian flies off at nearly double that speed, 180 kilometers per hour. Now suppose the car was moving at 90 percent of the speed of light. Would the pedestrian fly off at 180% of \( c \)?

To see why not, we have to back up a little and think about where this speed-doubling result comes from. For any collision, there is a special frame of reference, the center-of-mass frame, in which the two colliding objects approach each other, collide, and rebound with their velocities reversed. In the center-of-mass frame, the total momentum of the objects is zero both before and after the collision.
An unequal collision, viewed in the center-of-mass frame, 1, and in the frame where the small ball is initially at rest, 2. The motion is shown as it would appear on the film of an old-fashioned movie camera, with an equal amount of time separating each frame from the next. Film 1 was made by a camera that tracked the center of mass, film 2 by one that was initially tracking the small ball, and kept on moving at the same speed after the collision.

Figure a/1 shows such a frame of reference for objects of very unequal mass. Before the collision, the large ball is moving relatively slowly toward the top of the page, but because of its greater mass, its momentum cancels the momentum of the smaller ball, which is moving rapidly in the opposite direction. The total momentum is zero. After the collision, the two balls just reverse their directions of motion. We know that this is the right result for the outcome of the collision because it conserves both momentum and kinetic energy, and everything not forbidden is compulsory, i.e., in any experiment, there is only one possible outcome, which is the one that obeys all the conservation laws.

**Self-check C**

How do we know that momentum and kinetic energy are conserved in figure a/1?  

Let’s make up some numbers as an example. Say the small ball has a mass of 1 kg, the big one 8 kg. In frame 1, let’s make the velocities as follows:

<table>
<thead>
<tr>
<th></th>
<th>before the collision</th>
<th>after the collision</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 0.8</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>○ 0.1</td>
<td>-0.1</td>
<td></td>
</tr>
</tbody>
</table>

Figure a/2 shows the same collision in a frame of reference where the small ball was initially at rest. To find all the velocities in this frame, we just add 0.8 to all the ones in the previous table.

<table>
<thead>
<tr>
<th></th>
<th>before the collision</th>
<th>after the collision</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 0</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>○ 0.9</td>
<td>0.7</td>
<td></td>
</tr>
</tbody>
</table>
In this frame, as expected, the small ball flies off with a velocity, 1.6, that is almost twice the initial velocity of the big ball, 0.9.

If all those velocities were in meters per second, then that’s exactly what happened. But what if all these velocities were in units of the speed of light? Now it’s no longer a good approximation just to add velocities. We need to combine them according to the relativistic rules. For instance, the technique used in problem 1 on p. 457 can be used to show that combining a velocity of 0.8 times the speed of light with another velocity of 0.8 results in 0.98, not 1.6. The results are very different:

<table>
<thead>
<tr>
<th>before the collision</th>
<th>after the collision</th>
</tr>
</thead>
<tbody>
<tr>
<td>● 0</td>
<td>0.98</td>
</tr>
<tr>
<td>○ 0.83</td>
<td>0.76</td>
</tr>
</tbody>
</table>

We can interpret this as follows. Figure a/1 is one in which the big ball is moving fairly slowly. This is very nearly the way the scene would be seen by an ant standing on the big ball. According to an observer in frame b, however, both balls are moving at nearly the speed of light after the collision. Because of this, the balls appear foreshortened, but the distance between the two balls is also shortened. To this observer, it seems that the small ball isn’t pulling away from the big ball very fast.

Now here’s what’s interesting about all this. The outcome shown in figure a/2 was supposed to be the only one possible, the only one that satisfied both conservation of energy and conservation of momentum. So how can the different result shown in figure b be possible? The answer is that relativistically, momentum must not equal mv. The old, familiar definition is only an approximation that’s valid at low speeds. If we observe the behavior of the small ball in figure b, it looks as though it somehow had some extra inertia. It’s as though a football player tried to knock another player down without realizing that the other guy had a three-hundred-pound bag full of lead shot hidden under his uniform — he just doesn’t seem to react to the collision as much as he should. As proved in section 7.3.4, this extra inertia is described by redefining momentum as

\[ p = m\gamma v. \]
Example 17.

Two early high-precision tests of the relativistic equation \( p = mv \) for momentum. Graphing \( p/m \) rather than \( p \) allows the data for electrons and protons to be placed on the same graph. Natural units are used, so that the horizontal axis is the velocity in units of \( c \), and the vertical axis is the unitless quantity \( p/mc \). The very small error bars for the data point from Zrelov are represented by the height of the black rectangle.

At very low velocities, \( \gamma \) is close to 1, and the result is very nearly \( mv \), as demanded by the correspondence principle. But at very high velocities, \( \gamma \) gets very big — the small ball in figure b has a \( \gamma \) of 5.0, and therefore has five times more inertia than we would expect nonrelativistically.

This also explains the answer to another paradox often posed by beginners at relativity. Suppose you keep on applying a steady force to an object that's already moving at 0.9999c. Why doesn't it just keep on speeding up past \( c \)? The answer is that force is the rate of change of momentum. At 0.9999c, an object already has a \( \gamma \) of 71, and therefore has already sucked up 71 times the momentum you'd expect at that speed. As its velocity gets closer and closer to \( c \), its \( \gamma \) approaches infinity. To move at \( c \), it would need an infinite momentum, which could only be caused by an infinite force.

**Push as hard as you like . . . example 17**

We don't have to depend on our imaginations to see what would happen if we kept on applying a force to an object indefinitely and tried to accelerate it past \( c \). A nice experiment of this type was done by Bertozzi in 1964. In this experiment, electrons were accelerated by an electric field \( E \) through a distance \( \ell_1 \). Applying Newton's laws gives Newtonian predictions \( a_N \) for the acceleration and \( t_N \) for the time required.

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

Figure c shows the results. According to Newton, an acceleration \( a_N \) acting for a time \( t_N \) should produce a final velocity \( a_N t_N \). The solid line in the graph shows the prediction of Newton's laws, which is that a constant force exerted steadily over time will produce a velocity that rises linearly and without limit.

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

Figure d shows experimental data confirming the relativistic equation for momentum.

---

4Newton’s second law gives \( a_N = F/m = eE/m \). The constant-acceleration equation \( \Delta x = (1/2)at^2 \) then gives \( t_N = \sqrt{2m\ell_1/eE} \).

5To make the low-energy portion of the graph legible, Bertozzi’s highest-energy data point is omitted.
7.3.2 Equivalence of mass and energy

Now we’re ready to see why mass and energy must be equivalent as claimed in the famous $E = mc^2$. So far we’ve only considered collisions in which none of the kinetic energy is converted into any other form of energy, such as heat or sound. Let’s consider what happens if a blob of putty moving at velocity $v$ hits another blob that is initially at rest, sticking to it. The nonrelativistic result is that to obey conservation of momentum the two blobs must fly off together at $v/2$. Half of the initial kinetic energy has been converted to heat.

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 $m\gamma v$, and the more rapidly moving atoms in the hot object have higher values of $\gamma$. In our collision, the final combined blob must therefore be moving a little more slowly than the expected $v/2$, since otherwise the final momentum would have been a little greater than the initial momentum. To an observer who believes in conservation of momentum and knows only about the overall motion of the objects and not about their heat content, the low velocity after the collision would seem to be the result of a magical change in the mass, as if the mass of two combined, hot blobs of putty was more than the sum of their individual masses.

Now we know that the masses of all the atoms in the blobs must be the same as they always were. The change is due to the change in $\gamma$ with heating, not to a change in mass. The heat energy, however, seems to be acting as if it was equivalent to some extra mass.

But this whole argument was based on the fact that heat is a form of kinetic energy at the atomic level. Would $E = mc^2$ apply to other forms of energy as well? Suppose a rocket ship contains some electrical energy stored in a battery. If we believed that $E = mc^2$ applied to forms of kinetic energy but not to electrical energy, then we would have to believe that the pilot of the rocket could slow the ship down by using the battery to run a heater! This would not only be strange, but it would violate the principle of relativity, because the result of the experiment would be different depending on whether the ship was at rest or not. The only logical conclusion is that all forms of energy are equivalent to mass. Running the heater then has no effect on the motion of the ship, because the total energy in the ship was unchanged; one form of energy (electrical) was simply converted to another (heat).

The equation $E = mc^2$ tells us how much energy is equivalent

---

6 A double-mass object moving at half the speed does not have the same kinetic energy. Kinetic energy depends on the square of the velocity, so cutting the velocity in half reduces the energy by a factor of $1/4$, which, multiplied by the doubled mass, makes $1/2$ the original energy.
to how much mass: the conversion factor is the square of the speed of light, \( c \). Since \( c \) a big number, you get a really really big number when you multiply it by itself to get \( c^2 \). This means that even a small amount of mass is equivalent to a very large amount of energy.

---

Gravity bending light example 18

Gravity is a universal attraction between things that have mass, and since the energy in a beam of light is equivalent to some very small amount of mass, we expect that light will be affected by gravity, although the effect should be very small. The first important experimental confirmation of relativity came in 1919 when stars next to the sun during a solar eclipse were observed to have shifted a little from their ordinary position. (If there was no eclipse, the glare of the sun would prevent the stars from being observed.) Starlight had been deflected by the sun’s gravity. Figure e 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.

Black holes example 19

A star with sufficiently strong gravity can prevent light from leaving. Quite a few black holes have been detected via their gravitational forces on neighboring stars or clouds of gas and dust.
You’ve learned about conservation of mass and conservation of energy, but now we see that they’re not even separate conservation laws. As a consequence of the theory of relativity, mass and energy are equivalent, and are not separately conserved — one can be converted into the other. Imagine that a magician waves his wand, and changes a bowl of dirt into a bowl of lettuce. You’d be impressed, because you were expecting that both dirt and lettuce would be conserved quantities. Neither one can be made to vanish, or to appear out of thin air. However, there are processes that can change one into the other. A farmer changes dirt into lettuce, and a compost heap changes lettuce into dirt. At the most fundamental level, lettuce and dirt aren’t really different things at all; they’re just collections of the same kinds of atoms — carbon, hydrogen, and so on. Because mass and energy are like two different sides of the same coin, we may speak of mass-energy, a single conserved quantity, found by adding up all the mass and energy, with the appropriate conversion factor: $E = mc^2$.

\[ \text{A rusting nail example 20} \]
\[ \text{An iron nail is left in a cup of water until it turns entirely to rust. The energy released is about 0.5 MJ. In theory, would a sufficiently precise scale register a change in mass? If so, how much?} \]
\[ \text{The energy will appear as heat, which will be lost to the environment. The total mass-energy of the cup, water, and iron will indeed be lessened by 0.5 MJ. (If it had been perfectly insulated, there would have been no change, since the heat energy would have been trapped in the cup.) The speed of light is } c = 3 \times 10^8 \text{ meters per second, so converting to mass units, we have} \]
\[ m = \frac{E}{c^2} = \frac{0.5 \times 10^6 \text{ J}}{(3 \times 10^8 \text{ m/s})^2} = 6 \times 10^{-12} \text{ kilograms.} \]

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.

\[ \text{Electron-positron annihilation example 21} \]
\[ \text{Natural radioactivity in the earth produces positrons, which are like electrons but have the opposite charge. A form of antimatter, positrons annihilate with electrons to produce gamma rays, a form of high-frequency light. Such a process would have been considered impossible before Einstein, because conservation of mass and energy were believed to be separate principles, and this process eliminates 100% of the original mass. The amount of energy produced by annihilating 1 kg of matter with 1 kg of} \]

Section 7.3  Dynamics  435
antimatter is

\[ E = mc^2 \]
\[ = (2 \text{ kg}) \left( 3.0 \times 10^8 \text{ m/s} \right)^2 \]
\[ = 2 \times 10^{17} \text{ J}, \]

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

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

One commonly hears some misinterpretations of \( E = mc^2 \), one being that the equation tells us how much kinetic energy an object would have if it was moving at the speed of light. This wouldn’t make much sense, both because the equation for kinetic energy has \( 1/2 \) in it, \( KE = (1/2)mv^2 \), and because a material object can’t be made to move at the speed of light. However, this naturally leads to the question of just how much mass-energy a moving object has. We know that when the object is at rest, it has no kinetic energy, so its mass-energy is simply equal to the energy-equivalent of its mass, \( mc^2 \),

\[ \mathcal{E} = mc^2 \text{ when } v = 0, \]

where the symbol \( \mathcal{E} \) (cursive “E”) stands for mass-energy. The point of using the new symbol is simply to remind ourselves that we’re talking about relativity, so an object at rest has \( \mathcal{E} = mc^2 \), not \( E = 0 \) as we’d assume in nonrelativistic physics.

Suppose we start accelerating the object with a constant force. A constant force means a constant rate of transfer of momentum, but \( p = m\gamma v \) approaches infinity as \( v \) approaches \( c \), so the object will only get closer and closer to the speed of light, but never reach it. Now what about the work being done by the force? The force keeps doing work and doing work, which means that we keep on using up energy. Mass-energy is conserved, so the energy being expended must equal the increase in the object’s mass-energy. We can continue this process for as long as we like, and the amount of mass-energy will increase without limit. We therefore conclude that an object’s mass-energy approaches infinity as its speed approaches the speed of light,

\[ \mathcal{E} \rightarrow \infty \text{ when } v \rightarrow c. \]

Now that we have some idea what to expect, what is the actual equation for the mass-energy? As proved in section 7.3.4, it is

\[ \mathcal{E} = m\gamma c^2. \]
**self-check D**
Verify that this equation has the two properties we wanted. ▷

Answer, p. 1062

1 KE compared to \( mc^2 \) at low speeds example 22 ▷
An object is moving at ordinary nonrelativistic speeds. Compare its kinetic energy to the energy \( mc^2 \) it has purely because of its mass.

▷ The speed of light is a very big number, so \( mc^2 \) is a huge number of joules. The object has a gigantic amount of energy because of its mass, and only a relatively small amount of additional kinetic energy because of its motion.

Another way of seeing this is that at low speeds, \( \gamma \) is only a tiny bit greater than 1, so \( \varepsilon \) is only a tiny bit greater than \( mc^2 \).

1 The correspondence principle for mass-energy example 23 ▷
Show that the equation \( \varepsilon = m\gamma c^2 \) obeys the correspondence principle.

▷ As we accelerate an object from rest, its mass-energy becomes greater than its resting value. Nonrelativistically, we interpret this excess mass-energy as the object’s kinetic energy,

\[
KE = \varepsilon(v) - \varepsilon(v = 0) = m\gamma c^2 - mc^2 = m(\gamma - 1)c^2.
\]

Expressing \( \gamma \) as \( (1 - v^2/c^2)^{-1/2} \) and making use of the approximation \( (1 + \varepsilon)^\rho \approx 1 + \rho \varepsilon \) for small \( \varepsilon \), we have \( \gamma \approx 1 + v^2/2c^2 \), so

\[
KE \approx m(1 + \frac{v^2}{2c^2} - 1)c^2 = \frac{1}{2}mv^2,
\]

which is the nonrelativistic expression. As demanded by the correspondence principle, relativity agrees with newtonian physics at speeds that are small compared to the speed of light.

7.3.3 ★ The energy-momentum four-vector
Starting from \( \varepsilon = m\gamma \) and \( p = m\gamma v \), a little algebra allows one to prove the identity

\[
m^2 = \varepsilon^2 - p^2.
\]

We can define an energy-momentum four-vector,

\[
p = (\varepsilon, p_x, p_y, p_z),
\]

and the relation \( m^2 = \varepsilon^2 - p^2 \) then arises from the inner product \( \mathbf{p} \cdot \mathbf{p} \). Since \( \varepsilon \) and \( p \) are separately conserved, the energy-momentum four-vector is also conserved.
A high-precision test of this fundamental relativistic relationship was carried out by Meyer et al. in 1963 by studying the motion of electrons in static electric and magnetic fields. They define the quantity

\[ Y^2 = \frac{\xi^2}{m^2 + p^2}, \]

which according to special relativity should equal 1. Their results, tabulated in the sidebar, show excellent agreement with theory.

### Energy and momentum of light

**example 24**

Light has \( m = 0 \) and \( \gamma = \infty \), so if we try to apply \( E = m\gamma \) and \( p = m\gamma v \) to light, or to any massless particle, we get the indeterminate form \( 0 \cdot \infty \), which can’t be evaluated without a delicate and laborious evaluation of limits as in problem 11 on p. 460.

Applying \( m^2 = \xi^2 - p^2 \) yields the same result, \( \xi = |p| \), much more easily. This example demonstrates that although we encountered the relations \( E = m\gamma \) and \( p = m\gamma v \) first, the identity \( m^2 = \xi^2 - p^2 \) is actually more fundamental.

Figure q on p. 734 shows an experiment that verified \( E = |p| \) empirically.

For the reasons given in example 24, we take \( m^2 = \xi^2 - p^2 \) to be the *definition* of mass in relativity. One thing to be careful about is that this definition is not additive. Suppose that we lump two systems together and call them one big system, adding their mass-energies and momenta. When we do this, the mass of the combination is not the same as the sum of the masses. For example, suppose we have two rays of light moving in opposite directions, with energy-momentum vectors \((\xi, 0, 0, 0)\) and \((\xi, 0, 0, 0)\). Adding these gives \((2\xi, 0, 0, 0)\), which implies a mass equal to \( 2\xi \). In fact, in the early universe, where the density of light was high, the universe’s ambient gravitational fields were mainly those caused by the light it contained.

### Mass-energy, not energy, goes in the energy-momentum four-vector

**example 25**

When we say that something is a four-vector, we mean that it behaves properly under a Lorentz transformation: we can draw such a four-vector on graph paper, and then when we change frames of reference, we should be able to measure the vector in the new frame of reference by using the new version of the graph-paper grid derived from the old one by a Lorentz transformation.

If we had used the energy \( E \) rather than the mass-energy \( \xi \) to construct the energy-momentum four-vector, we wouldn’t have gotten a valid four-vector. An easy way to see this is to consider the case where a noninteracting object is at rest in some frame of reference. Its momentum and kinetic energy are both zero. If we’d defined \( \mathbf{p} = (E, p_x, p_y, p_z) \) rather than \( \mathbf{p} = (\xi, p_x, p_y, p_z) \), we
would have had \( p = 0 \) in this frame. But when we draw a zero vector, we get a point, and a point remains a point regardless of how we distort the graph paper we use to measure it. That wouldn’t have made sense, because in other frames of reference, we have \( E \neq 0 \).

**Metric units**

The relation \( m^2 = \varepsilon^2 - p^2 \) is only valid in relativistic units. If we tried to apply it without modification to numbers expressed in metric units, we would have

\[
\text{kg}^2 = \text{kg}^2 \cdot \frac{m^4}{s^4} - \text{kg}^2 \cdot \frac{m^2}{s^2},
\]

which would be nonsense because the three terms all have different units. As usual, we need to insert factors of \( c \) to make a metric version, and these factors of \( c \) are determined by the need to fix the broken units:

\[
m^2 c^4 = \varepsilon^2 - p^2 c^2
\]

**Pair production requires matter**

Example 21 on p. 435 discussed the annihilation of an electron and a positron into two gamma rays, which is an example of turning matter into pure energy. An opposite example is pair production, a process in which a gamma ray disappears, and its energy goes into creating an electron and a positron.

Pair production cannot happen in a vacuum. For example, gamma rays from distant black holes can travel through empty space for thousands of years before being detected on earth, and they don’t turn into electron-positron pairs before they can get here. Pair production can only happen in the presence of matter. When lead is used as shielding against gamma rays, one of the ways the gamma rays can be stopped in the lead is by undergoing pair production.

To see why pair production is forbidden in a vacuum, consider the process in the frame of reference in which the electron-positron pair has zero total momentum. In this frame, the gamma ray would have had zero momentum, but a gamma ray with zero momentum must have zero energy as well (example 24). This means that conservation of four-momentum has been violated: the timelike component of the four-momentum is the mass-energy, and it has increased from 0 in the initial state to at least \( 2mc^2 \) in the final state.
7.3.4 *Proofs*

This optional section proves some results claimed earlier.

*Ultrarelativistic motion*

We start by considering the case of a particle, described as “ultrarelativistic,” that travels at very close to the speed of light. A good way of thinking about such a particle is that it’s one with a very small mass. For example, the subatomic particle called the neutrino has a very small mass, thousands of times smaller than that of the electron. Neutrinos are emitted in radioactive decay, and because the neutrino’s mass is so small, the amount of energy available in these decays is always enough to accelerate it to very close to the speed of light. Nobody has ever succeeded in observing a neutrino that was *not* ultrarelativistic. When a particle’s mass is very small, the mass becomes difficult to measure. For almost 70 years after the neutrino was discovered, its mass was thought to be zero. Similarly, we currently believe that a ray of light has no mass, but it is always possible that its mass will be found to be nonzero at some point in the future. A ray of light can be modeled as an ultrarelativistic particle.

Let’s compare ultrarelativistic particles with train cars. A single car with kinetic energy $E$ has different properties than a train of two cars each with kinetic energy $E/2$. The single car has half the mass and a speed that is greater by a factor of $\sqrt{2}$. But the same is not true for ultrarelativistic particles. Since an idealized ultrarelativistic particle has a mass too small to be detectable in any experiment, we can’t detect the difference between $m$ and $2m$. Furthermore, ultrarelativistic particles move at close to $c$, so there is no observable difference in speed. Thus we expect that a single ultrarelativistic particle with energy $E$ compared with two such particles, each with energy $E/2$, should have all the same properties as measured by a mechanical detector.

An idealized zero-mass particle also has no frame in which it can be at rest. It always travels at $c$, and no matter how fast we chase after it, we can never catch up. We can, however, observe it in different frames of reference, and we will find that its energy is different. For example, distant galaxies are receding from us at substantial fractions of $c$, and when we observe them through a telescope, they appear very dim not just because they are very far away but also because their light has less energy in our frame than in a frame at rest relative to the source. This effect must be such that changing frames of reference according to a specific Lorentz transformation always changes the energy of the particle by a fixed factor, regardless of the particle’s original energy; for if not, then the effect of a Lorentz transformation on a single particle of energy $E$ would be different from its effect on two particles of energy $E/2$. 


How does this energy-shift factor depend on the velocity $v$ of the Lorentz transformation? Rather than $v$, it becomes more convenient to express things in terms of the Doppler shift factor $D$, which multiplies when we change frames of reference. Let’s write $f(D)$ for the energy-shift factor that results from a given Lorentz transformation. Since a Lorentz transformation $D_1$ followed by a second transformation $D_2$ is equivalent to a single transformation by $D_1D_2$, we must have $f(D_1D_2) = f(D_1)f(D_2)$. This tightly constrains the form of the function $f$; it must be something like $f(D) = s^n$, where $n$ is a constant. We postpone until p. 442 the proof that $n = 1$, which is also in agreement with experiments with rays of light.

Our final result is that the energy of an ultrarelativistic particle is simply proportional to its Doppler shift factor $D$. Even in the case where the particle is truly massless, so that $D$ doesn’t have any finite value, we can still find how the energy differs according to different observers by finding the $D$ of the Lorentz transformation between the two observers’ frames of reference.

**Energy**

The following argument is due to Einstein. Suppose that a material object $O$ of mass $m$, initially at rest in a certain frame $A$, emits two rays of light, each with energy $E/2$. By conservation of energy, the object must have lost an amount of energy equal to $E$. By symmetry, $O$ remains at rest.

We now switch to a new frame of reference moving at a certain velocity $v$ in the $z$ direction relative to the original frame. We assume that $O$’s energy is different in this frame, but that the change in its energy amounts to multiplication by some unitless factor $x$, which depends only on $v$, since there is nothing else it could depend on that could allow us to form a unitless quantity. In this frame the light rays have energies $ED(v)$ and $ED(-v)$. If conservation of energy is to hold in the new frame as it did in the old, we must have $2xE = ED(v) + ED(-v)$. After some algebra, we find $x = 1/\sqrt{1 - v^2}$, which we recognize as $\gamma$. This proves that $E = m\gamma$ for a material object.

**Momentum**

We’ve seen that ultrarelativistic particles are “generic,” in the sense that they have no individual mechanical properties other than an energy and a direction of motion. Therefore the relationship between energy and momentum must be linear for ultrarelativistic particles. Indeed, experiments verify that light has momentum, and doubling the energy of a ray of light doubles its momentum rather than quadrupling it. On a graph of $p$ versus $E$, massless particles, which have $E \propto |p|$, lie on two diagonal lines that connect at the origin. If we like, we can pick units such that the slopes of these lines are plus and minus one. Material particles lie to the right of
these lines. For example, a car sitting in a parking lot has $p = 0$ and $E = mc^2$.

Now what happens to such a graph when we change to a different frame or reference that is in motion relative to the original frame? A massless particle still has to act like a massless particle, so the diagonals are simply stretched or contracted along their own lengths. In fact the transformation must be linear (p. 403), because conservation of energy and momentum involve addition, and we need these laws to be valid in all frames of reference. By the same reasoning as in figure j on p. 405, the transformation must be area-preserving. We then have the same three cases to consider as in figure g on p. 404. Case I is ruled out because it would imply that particles keep the same energy when we change frames. (This is what would happen if $c$ were infinite, so that the mass-equivalent $E/c^2$ of a given energy was zero, and therefore $E$ would be interpreted purely as the mass.) Case II can’t be right because it doesn’t preserve the $E = |p|$ diagonals. We are left with case III, which establishes the fact that the $p$-$E$ plane transforms according to exactly the same kind of Lorentz transformation as the $x$-$t$ plane. That is, $(E, p_x, p_y, p_z)$ is a four-vector.

The only remaining issue to settle is whether the choice of units that gives invariant 45-degree diagonals in the $x$-$t$ plane is the same as the choice of units that gives such diagonals in the $p$-$E$ plane. That is, we need to establish that the $c$ that applies to $x$ and $t$ is equal to the $c'$ needed for $p$ and $E$, i.e., that the velocity scales of the two graphs are matched up. This is true because in the Newtonian limit, the total mass-energy $E$ is essentially just the particle’s mass, and then $p/E \approx p/m \approx v$. This establishes that the velocity scales are matched at small velocities, which implies that they coincide for all velocities, since a large velocity, even one approaching $c$, can be built up from many small increments. (This also establishes that the exponent $n$ defined on p. 441 equals 1 as claimed.)

Since $m^2 = E^2 - p^2$, it follows that for a material particle, $p = m\gamma v$. 

442 Chapter 7 Relativity
7.4 ✽ General relativity

What you’ve learned so far about relativity is known as the special theory of relativity, which is compatible with three of the four known forces of nature: electromagnetism, the strong nuclear force, and the weak nuclear force. Gravity, however, can’t be shoehorned into the special theory. In order to make gravity work, Einstein had to generalize relativity. The resulting theory is known as the general theory of relativity.7

7.4.1 Our universe isn’t Euclidean

Euclid proved thousands of years ago that the angles in a triangle add up to 180°. But what does it really mean to “prove” this? Euclid proved it based on certain assumptions (his five postulates), listed in the margin of this page. But how do we know that the postulates are true?

Only by observation can we tell whether any of Euclid’s statements are correct characterizations of how space actually behaves in our universe. If we draw a triangle on paper with a ruler and measure its angles with a protractor, we will quickly verify to pretty good precision that the sum is close to 180°. But of course we already knew that space was at least approximately Euclidean. If there had been any gross error in Euclidean geometry, it would have been detected in Euclid’s own lifetime. The correspondence principle tells us that if there is going to be any deviation from Euclidean geometry, it must be small under ordinary conditions.

To improve the precision of the experiment, we need to make sure that our ruler is very straight. One way to check would be to sight along it by eye, which amounts to comparing its straightness to that of a ray of light. For that matter, we might as well throw the physical ruler in the trash and construct our triangle out of three laser beams. To avoid effects from the air we should do the experiment in outer space. Doing it in space also has the advantage of allowing us to make the triangle very large; as shown in figure a, the discrepancy from 180° is expected to be proportional to the area of the triangle.

But we already know that light rays are bent by gravity. We expect it based on $E = mc^2$, which tells us that the energy of a light ray is equivalent to a certain amount of mass, and furthermore it has been verified experimentally by the deflection of starlight by the sun (example 18, p. 434). We therefore know that our universe is noneuclidean, and we gain the further insight that the level of

---

7Einstein originally described the distinction between the two theories by saying that the special theory applied to nonaccelerating frames of reference, while the general one allowed any frame at all. The modern consensus is that Einstein was misinterpreting his own theory, and that special relativity actually handles accelerating frames just fine.
An Einstein’s ring. The distant object is a quasar, MG1131+0456, and the one in the middle is an unknown object, possibly a supermassive black hole. The intermediate object’s gravity focuses the rays of light from the distant one. Because the entire arrangement lacks perfect axial symmetry, the ring is nonuniform; most of its brightness is concentrated in two lumps on opposite sides.

Deviation from Euclidean behavior depends on gravity.

Since the noneuclidean effects are bigger when the system being studied is larger, we expect them to be especially important in the study of cosmology, where the distance scales are very large.

An Einstein’s ring, figure b, is formed when there is a chance alignment of a distant source with a closer gravitating body. This type of gravitational lensing is direct evidence for the noneuclidean nature of space. The two light rays are lines, and they violate Euclid’s first postulate, that two points determine a line.

One could protest that effects like these are just an imperfection of the light rays as physical models of straight lines. Maybe the noneuclidean effects would go away if we used something better and straighter than a light ray. But we don’t know of anything straighter than a light ray. Furthermore, we observe that all measuring devices, not just optical ones, report the same noneuclidean behavior.

Curvature

An example of such a non-optical measurement is the Gravity Probe B satellite, figure d, which was launched into a polar orbit in 2004 and operated until 2010. The probe carried four gyroscopes made of quartz, which were the most perfect spheres ever manufactured, varying from sphericity by no more than about 40 atoms. Each gyroscope floated weightlessly in a vacuum, so that its rotation was perfectly steady. After 5000 orbits, the gyroscopes had reoriented themselves by about $2 \times 10^{-3}\degree$ relative to the distant stars. This effect cannot be explained by Newtonian physics, since no torques acted on them. It was, however, exactly as predicted by Einstein’s theory of general relativity. It becomes easier to see why such an effect would be expected due to the noneuclidean nature of space if we characterize euclidean geometry as the geometry of a flat plane as opposed to a curved one. On a curved surface like a sphere, figure c, Euclid’s fifth postulate fails, and it’s not hard to see that we can get triangles for which the sum of the angles is not $180\degree$. By transporting a gyroscope all the way around the edges of such a triangle and back to its starting point, we change its orientation.

The triangle in figure c has angles that add up to more than $180\degree$. This type of curvature is referred to as positive. It is also possible to have negative curvature, as in figure e.

In general relativity, curvature isn’t just something caused by gravity. Gravity is curvature, and the curvature involves both space and time, as may become clearer once you get to figure k. Thus the distinction between special and general relativity is that general relativity handles curved spacetime, while special relativity is restricted
Gravity Probe B was in a polar orbit around the earth. As in the right panel of figure c, the orientation of the gyroscope changes when it is carried around a curve and back to its starting point. Because the effect was small, it was necessary to let it accumulate over the course of 5000 orbits in order to make it detectable.

A triangle in a space with negative curvature has angles that add to less than 180°.

Curvature doesn’t require higher dimensions

Although we often visualize curvature by imagining embedding a two-dimensional surface in a three-dimensional space, that’s just an aid in visualization. There is no evidence for any additional dimensions, nor is it necessary to hypothesize them in order to let spacetime be curved as described in general relativity.

Only measurements from within the plane define whether the plane is curved. It could look curved when drawn embedded in three dimensions, but nevertheless still be intrinsically flat.

Put yourself in the shoes of a two-dimensional being living in a two-dimensional space. Euclid’s postulates all refer to constructions that can be performed using a compass and an unmarked straight-edge. If this being can physically verify them all as descriptions of the space she inhabits, then she knows that her space is Euclidean, and that propositions such as the Pythagorean theorem are physically valid in her universe. But the diagram in f/1 illustrating the proof of the Pythagorean theorem in Euclid’s Elements (proposition I.47) is equally valid if the page is rolled onto a cylinder, 2, or formed into a wavy corrugated shape, 3. These types of curvature, which can be achieved without tearing or crumpling the surface, are not real to her. They are simply side-effects of visualizing her
two-dimensional universe as if it were embedded in a hypothetical third dimension — which doesn’t exist in any sense that is empirically verifiable to her. Of the curved surfaces in figure f, only the sphere, 4, has curvature that she can measure; the diagram can’t be plastered onto the sphere without folding or cutting and pasting.

So the observation of curvature doesn’t imply the existence of extra dimensions, nor does embedding a space in a higher-dimensional one so that it looks curvy always mean that there will be any curvature detectable from within the lower-dimensional space.

7.4.2 The equivalence principle

Universality of free-fall

Although light rays and gyroscopes seem to agree that space is curved in a gravitational field, it’s always conceivable that we could find something else that would disagree. For example, suppose that there is a new and improved ray called the StraightRay™. The StraightRay is like a light ray, but when we construct a triangle out of StraightRays, we always get the Euclidean result for the sum of the angles. We would then have to throw away general relativity’s whole idea of describing gravity in terms of curvature. One good way of making a StraightRay would be if we had a supply of some kind of exotic matter — call it FloatyStuff™ — that had the ordinary amount of inertia, but was completely unaffected by gravity. We could then shoot a stream of FloatyStuff particles out of a nozzle at nearly the speed of light and make a StraightRay.

Normally when we release a material object in a gravitational field, it experiences a force $mg$, and then by Newton’s second law its acceleration is $a = F/m = mg/m = g$. The $m$’s cancel, which is the reason that everything falls with the same acceleration (in the absence of other forces such as air resistance). The universality of this behavior is what allows us to interpret the gravity geometrically in general relativity. For example, the Gravity Probe B gyroscopes were made out of quartz, but if they had been made out of something else, it wouldn’t have mattered. But if we had access to some FloatyStuff, the geometrical picture of gravity would fail, because the “$m$” that described its susceptibility to gravity would be a different “$m$” than the one describing its inertia.

The question of the existence or nonexistence of such forms of matter turns out to be related to the question of what kinds of motion are relative. Let’s say that alien gangsters land in a flying saucer, kidnap you out of your back yard, konk you on the head, and take you away. When you regain consciousness, you’re locked up in a sealed cabin in their spaceship. You pull your keychain out of your pocket and release it, and you observe that it accelerates toward the floor with an acceleration that seems quite a bit slower than what you’re used to on earth, perhaps a third of a gee. There
are two possible explanations for this. One is that the aliens have taken you to some other planet, maybe Mars, where the strength of gravity is a third of what we have on earth. The other is that your keychain didn’t really accelerate at all: you’re still inside the flying saucer, which is accelerating at a third of a gee, so that it was really the deck that accelerated up and hit the keys.

There is absolutely no way to tell which of these two scenarios is actually the case — unless you happen to have a chunk of FloatyStuff in your other pocket. If you release the FloatyStuff and it hovers above the deck, then you’re on another planet and experiencing genuine gravity; your keychain responded to the gravity, but the FloatyStuff didn’t. But if you release the FloatyStuff and see it hit the deck, then the flying saucer is accelerating through outer space.

The nonexistence of FloatyStuff in our universe is called the equivalence principle. If the equivalence principle holds, then an acceleration (such as the acceleration of the flying saucer) is always equivalent to a gravitational field, and no observation can ever tell the difference without reference to something external. (And suppose you did have some external reference point — how would you know whether it was accelerating?)

The artificial horizon example 29

The pilot of an airplane cannot always easily tell which way is up. The horizon may not be level simply because the ground has an actual slope, and in any case the horizon may not be visible if the weather is foggy. One might imagine that the problem could be solved simply by hanging a pendulum and observing which way it pointed, but by the equivalence principle the pendulum cannot tell the difference between a gravitational field and an acceleration of the aircraft relative to the ground — nor can any other accelerometer, such as the pilot’s inner ear. For example, when the plane is turning to the right, accelerometers will be tricked into believing that “down” is down and to the left. To get around this problem, airplanes use a device called an artificial horizon, which is essentially a gyroscope. The gyroscope has to be initialized when the plane is known to be oriented in a horizontal plane. No gyroscope is perfect, so over time it will drift. For this reason the instrument also contains an accelerometer, and the gyroscope is always forced into agreement with the accelerometer’s average output over the preceding several minutes. If the plane is flown in circles for several minutes, the artificial horizon will be fooled into indicating that the wrong direction is vertical.

Gravitational Doppler shifts and time dilation

An interesting application of the equivalence principle is the explanation of gravitational time dilation. As described on p. 400, experiments show that a clock at the top of a mountain runs faster
A ray of light is emitted upward from the floor of the elevator. The elevator accelerates upward. By the time the light is detected at the ceiling, the elevator has changed its velocity, so the light is detected with a Doppler shift.

Pound and Rebka at the top and bottom of the tower.
you have no way of determining from local observations, inside the saucer, whether you are actually weightless in deep space, or simply free-falling in apparent weightlessness, like the astronauts aboard the International Space Station. That means that locally, we can always adopt a free-falling frame of reference in which there is no gravitational field at all. If there is no gravity, then special relativity is valid, and we can treat our local region of spacetime as being approximately flat.

In figure k, an apple falls out of a tree. Its path is a “straight” line in spacetime, in the same sense that the equator is a “straight” line on the earth’s surface.

**Inertial frames**

In Newtonian mechanics, we have a distinction between inertial and noninertial frames of reference. An inertial frame according to Newton is one that has a constant velocity vector relative to the stars. But what if the stars themselves are accelerating due to a gravitational force from the rest of the galaxy? We could then take the galaxy’s center of mass as defining an inertial frame, but what if something else is acting on the galaxy?

If we had some FloatyStuff, we could resolve the whole question. FloatyStuff isn’t affected by gravity, so if we release a sample of it in mid-air, it will continue on a trajectory that defines a perfect Newtonian inertial frame. (We’d better have it on a tether, because otherwise the earth’s rotation will carry the earth out from under it.) But if the equivalence principle holds, then Newton’s definition of an inertial frame is fundamentally flawed.

There is a different definition of an inertial frame that works better in relativity. A Newtonian inertial frame was defined by an object that isn’t subject to any forces, gravitational or otherwise. In general relativity, we instead define an inertial frame using an
object that that isn’t influenced by anything other than gravity. By this definition, a free-falling rock defines an inertial frame, but this book sitting on your desk does not.

7.4.3 Black holes

The observations described so far showed only small effects from curvature. To get a big effect, we should look at regions of space in which there are strong gravitational fields. The prime example is a black hole. The best studied examples are two objects in our own galaxy: Cygnus X-1, which is believed to be a black hole with about ten times the mass of our sun, and Sagittarius A*, an object near the center of our galaxy with about four million solar masses.

Although a black hole is a relativistic object, we can gain some insight into how it works by applying Newtonian physics. A spherical body of mass $M$ has an escape velocity $v = \sqrt{\frac{2GM}{r}}$, which is the minimum velocity that we would need to give to a projectile shot from a distance $r$ so that it would never fall back down. If $r$ is small enough, the escape velocity will be greater than $c$, so that even a ray of light can never escape.

We can now make an educated guess as to what this means without having to study all the mathematics of general relativity. In relativity, $c$ isn’t really the speed of light, it’s really to be thought of as a restriction on how fast cause and effect can propagate through space. This suggests the correct interpretation, which is that for an object compact enough to be a black hole, there is no way for an event at a distance closer than $r$ to have an effect on an event far away. There is an invisible, spherical boundary with radius $r$, called the event horizon, and the region within that boundary is cut off from the rest of the universe in terms of cause and effect. If you wanted to explore that region, you could drop into it while wearing a space-suit — but it would be a one-way trip, because you could never get back out to report on what you had seen.

In the Newtonian description of a black hole, matter could be lifted out of a black hole, m. Would this be possible with a real-world black hole, which is relativistic rather than Newtonian? No, because the bucket is causally separated from the outside universe. No rope would be strong enough for this job (problem 12, p. 460).

One misleading aspect of the Newtonian analysis is that it encourages us to imagine that a light ray trying to escape from a black hole will slow down, stop, and then fall back in. This can’t be right, because we know that any observer who sees a light ray flying by always measures its speed to be $c$. This was true in special relativity, and by the equivalence principle we can be assured that the same is true locally in general relativity. Figure n shows what would really happen.