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 aa.

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 space-
time 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, \( \overline{AB} \cong \overline{CD} \), 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 \( \overline{AB} \cong \overline{CD} \) if:

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

3. \( \overline{AB} \) and \( \overline{CD} \) are both lightlike.

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

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

The definition in case 2 is the only sensible way to proceed if we are to respect the symmetric treatment of time and space in relativity. The timelike relation between the events is necessary and sufficient to make it possible for a clock to move from one to the other. It makes a difference that the clocks move inertially, because the twins in example 1 on p. 391 disagree on the clock time between the traveling twin’s departure and return.

Case 3 may seem strange, since it says that any two lightlike intervals are congruent. But this is the only possible definition,
because this case can be obtained as a limit of the timelike one. Suppose that AB is a timelike interval, but in the planet earth’s frame of reference it would be necessary to travel at almost the speed of light in order to reach B from A. The required speed is less than \( c \) (i.e., less than 1) by some tiny amount \( \epsilon \). 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 \( I \) (cursive letter “I”) between them is defined as \( I = \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2 \).

This is stated in natural units, so all four terms on the right-hand side have the same units; in metric units with \( c \neq 1 \), appropriate factors of \( c \) should be inserted in order to make the units of the terms agree. The interval \( I \) 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 8 below, it is the same regardless of our frame of reference, i.e., regardless of our choice of coordinates.

**Pioneer 10**

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

\[
\begin{array}{cccc}
 t (s) & x & y & z \\
0 & 1.784 \times 10^{12} \text{ m} & 3.951 \times 10^{12} \text{ m} & 0.237 \times 10^{12} \text{ m} \\
3.786912000 \times 10^8 \text{ s} & 2.420 \times 10^{12} \text{ m} & 8.827 \times 10^{12} \text{ m} & 0.488 \times 10^{12} \text{ m}
\end{array}
\]

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

\( \triangle \) 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:

\[
\begin{array}{c}
\Delta t (s) \\
3.786912000 \times 10^8 \text{ s}
\end{array}
\begin{array}{cccc}
\Delta x & \Delta y & \Delta z \\
0.2121 \times 10^4 \text{ s} & 1.626 \times 10^4 \text{ s} & 0.084 \times 10^4 \text{ s}
\end{array}
\]

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

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

\begin{itemize}
\item \text{1. The gray light-rectangle represents the set of all events such as} \( P \) \text{ that could be visited after} \( A \) \text{ and before} \( B \).
\item \text{2. The rectangle becomes a square in the frame in which} \( A \) \text{ and} \( B \) \text{ occur at the same location in space}.
\item \text{3. The area of the dashed square is} \( \tau^2 \), \text{ so the area of the gray square is} \( \tau^2/2 \).
\end{itemize}

\textit{Invariance of the interval example 8}

In this example we prove that the interval is the same regardless of what frame of reference we compute it in. This is called “Lorentz invariance.” The proof is limited to the timelike case. Given events \( A \) and \( B \), construct the light-rectangle as defined in figure ab/1. On p. 389 we proved that the Lorentz transformation doesn’t change the area of a shape in the \( x-t \) plane. Therefore the area of this rectangle is unchanged if we switch to the frame of reference ab/2, in which \( A \) and \( B \) occurred at the same location and were separated by a time interval \( \tau \). This area equals half the interval \( \mathcal{I} \) between \( A \) and \( B \). But a straightforward calculation shows that the rectangle in ab/1 also has an area equal to half the interval calculated in \textit{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.

Example 9. A numerical example of invariance

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

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

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.
Example 10. Four-vectors and the inner product

Example 7 makes it natural that we define a type of vector with four components, the first one relating to time and the others being spatial. These are known as four-vectors. It’s clear how we should define the equivalent of a dot product in relativity:

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

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

The twin paradox example 10

Alice and Betty are identical twins. Betty goes on a space voyage at relativistic speeds, traveling away from the earth and then turning around and coming back. Meanwhile, Alice stays on earth. When Betty returns, she is younger than Alice because of relativistic time dilation (example 1, p. 391).

But isn’t it valid to say that Betty’s spaceship is standing still and the earth moving? In that description, wouldn’t Alice end up younger and Betty older? This is referred to as the “twin paradox.” It can’t really be a paradox, since it’s exactly what was observed in the Hafele-Keating experiment (p. 381).

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

We have already established by other methods that Betty ages less 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 ad are

\[
\begin{align*}
\mathbf{a} &= (6, 1) \\
\mathbf{b} &= (3, 2) \\
\mathbf{c} &= (3, -1)
\end{align*}
\]

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

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

$$|b| + |c| = \sqrt{3^2 - 2^2} + \sqrt{3^2 - (-1)^2} = 5.1.$$ 

**Simultaneity using inner products**

Suppose that an observer $O$ moves inertially along a vector $\mathbf{o}$, and let the vector separating two events $P$ and $Q$ be $\mathbf{s}$. $O$ judges these events to be simultaneous if $\mathbf{o} \cdot \mathbf{s} = 0$. To see why this is true, suppose we pick a coordinate system as defined by $O$. In this coordinate system, $O$ considers herself to be at rest, so she says her vector has only a time component, $\mathbf{o} = (\Delta t, 0, 0, 0)$. If she considers $P$ and $Q$ to be simultaneous, then the vector from $P$ to $Q$ is of the form $(0, \Delta x, \Delta y, \Delta z)$. The inner product is then zero, since each of the four terms vanishes. Since the inner product is independent of the choice of coordinate system, it doesn’t matter that we chose one tied to $O$ herself. Any other observer $O'$ can look at $O$’s motion, note that $\mathbf{o} \cdot \mathbf{s} = 0$, and infer that $O$ must consider $P$ and $Q$ to be simultaneous, even if $O'$ says they weren’t.

### 7.2.8 *Doppler shifts of light and addition of velocities*

When Doppler shifts happen to ripples on a pond or the sound waves from an airplane, they can depend on the relative motion of three different objects: the source, the receiver, and the medium. But light waves don’t have a medium. Therefore Doppler shifts of light can only depend on the relative motion of the source and observer.

One simple case is the one in which the relative motion of the source and the receiver is perpendicular to the line connecting them. That is, the motion is transverse. Nonrelativistic Doppler shifts happen because the distance between the source and receiver is changing, so in nonrelativistic physics we don’t expect any Doppler shift at all when the motion is transverse, and this is what is in fact observed to high precision. For example, the photo shows shortened and lengthened wavelengths to the right and left, along the source’s line of motion, but an observer above or below the source measures just the normal, unshifted wavelength and frequency. But relativistically, we have a time dilation effect, so for light waves emitted transversely, there is a Doppler shift of $1/\gamma$ in frequency (or $\gamma$ in wavelength).

The other simple case is the one in which the relative motion of the source and receiver is longitudinal, i.e., they are either approaching or receding from one another. For example, distant galaxies are receding from our galaxy due to the expansion of the universe, and this expansion was originally detected because Doppler shifts toward the red (low-frequency) end of the spectrum were observed.

Nonrelativistically, we would expect the light from such a galaxy to be Doppler shifted down in frequency by some factor, which
would depend on the relative velocities of three different objects: the source, the wave’s medium, and the receiver. Relativistically, things get simpler, because light isn’t a vibration of a physical medium, so the Doppler shift can only depend on a single velocity \( v \), which is the rate at which the separation between the source and the receiver is increasing.

The square in figure ah 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. 440 that the Doppler shift is

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

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

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

Example 12
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 13
The result of example 12 was the basis of one of the earliest laboratory tests of special relativity, by Ives and Stilwell in 1938. They observed the light emitted by excited by a beam of $H_2^+$ and $H_3^+$ ions with speeds of a few tenths of a percent of $c$. Measuring the light from both ahead of and behind the beams, they found that the product of the Doppler shifts $D(v)D(-v)$ was equal to 1, as predicted by relativity. If relativity had been false, then one would have expected the product to differ from 1 by an amount that would have been detectable in their experiment. In 2003, Saathoff et al. carried out an extremely precise version of the Ives-Stilwell technique with Li$^+$ ions moving at 6.4% of $c$. The frequencies observed, in units of MHz, were:

- $f_0 = 546466918.8 \pm 0.4$ (unshifted frequency)
- $f_0D(-v) = 582490203.44 \pm 0.09$ (shifted frequency, forward)
- $f_0D(v) = 512671442.9 \pm 0.5$ (shifted frequency, backward)
- $\sqrt{f_0D(-v) \cdot f_0D(v)} = 546466918.6 \pm 0.3$

The results show incredibly precise agreement between $f_0$ and $\sqrt{f_0D(-v) \cdot f_0D(v)}$, as expected relativistically because $D(v)D(-v)$ is supposed to equal 1. The agreement extends to 9 significant figures, whereas if relativity had been false there should have been a relative disagreement of about $v^2 = .004$, i.e., a discrepancy in the third significant figure. The spectacular agreement with theory has made this experiment a lightning rod for anti-relativity kooks.

We saw on p. 394 that relativistic velocities should not be expected to be exactly additive, and problem 1 on p. 439 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. 443 guides you through the steps of this derivation, and the result is given on p. 940.
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>•</td>
<td>-0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>○</td>
<td>0.1</td>
<td>-0.1</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>•</td>
<td>0</td>
<td>1.6</td>
</tr>
<tr>
<td>○</td>
<td>0.9</td>
<td>0.7</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. 439 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></th>
<th>before the collision</th>
<th>after the collision</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>0</td>
<td>0.98</td>
</tr>
<tr>
<td>○</td>
<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

$$p = m\gamma v$$

b/ An 8-kg ball moving at 83% of the speed of light hits a 1-kg ball. The balls appear foreshortened due to the relativistic distortion of space.
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 0.9999$c$. Why doesn’t it just keep on speeding up past $c$? The answer is that force is the rate of change of momentum. At 0.9999$c$, 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 14**

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.\(^4\)

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

Figure c shows the results.\(^5\) According to Newton, an acceleration $a_N$ acting for a time $t_N$ should produce a final velocity $a_N t_N$. The solid line in the graph shows the prediction of Newton’s laws, which is that a constant force exerted steadily over time will produce a velocity that rises linearly and without limit.

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

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

---

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

\(^5\)To make the low-energy portion of the graph legible, Bertozzi’s highest-energy data point is omitted.
happens if a blob of putty moving at velocity $v$ hits another blob that is initially at rest, sticking to it. The nonrelativistic result is that to obey conservation of momentum the two blobs must fly off together at $v/2$. Half of the initial kinetic energy has been converted to heat.\(^6\)

Relativistically, however, an interesting thing happens. A hot object has more momentum than a cold object! This is because the relativistically correct expression for momentum is $m\gamma v$, 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 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.

\(^6\)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.
Gravity bending light

Example 15

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 d 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 then, appeared at positions slightly different than their normal ones.

Black holes

Example 16

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: \( E + mc^2 \).

\[ A \text{ rusting nail} \quad \text{example 17} \]

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

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

\[
m = \frac{E}{c^2} = \frac{0.5 \times 10^6 \text{ J}}{(3 \times 10^8 \text{ m/s})^2} = 6 \times 10^{-12} \text{ kilograms.}
\]

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.

\[ \text{Electron-positron annihilation} \quad \text{example 18} \]

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

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

$$E = mc^2 \text{ when } v = 0$$

where the symbol $E$ (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 $E = mc^2$, 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 $p = mγv$ 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,

$$E \to \infty \text{ when } v \to c .$$

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

$$E = mγc^2 .$$

**self-check D**

Verify that this equation has the two properties we wanted.

Answer, p. 928
KE compared to $mc^2$ at low speeds  

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 $\mathcal{E}$ is only a tiny bit greater than $mc^2$.

The correspondence principle for mass-energy  

Show that the equation $\mathcal{E} = m\gamma c^2$ obeys the correspondence principle.

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

$$KE = \mathcal{E}(v) - \mathcal{E}(v = 0) = m\gamma c^2 - mc^2 = m(\gamma - 1)c^2.$$  

Expressing $\gamma$ as $(1 - v^2/c^2)^{-1/2}$ and making use of the approximation $(1 + \epsilon)^p \approx 1 + p\epsilon$ for small $\epsilon$, 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 classical expression. As demanded by the correspondence principle, relativity agrees with classical physics at speeds that are small compared to the speed of light.

7.3.3 The energy-momentum four-vector

Starting from $\mathcal{E} = m\gamma$ and $p = m\gamma v$, a little algebra allows one to prove the identity

$$m^2 = \mathcal{E}^2 - p^2.$$  

We can define an energy-momentum four-vector,

$$\mathbf{p} = (\mathcal{E}, p_x, p_y, p_z),$$

and the relation $m^2 = \mathcal{E}^2 - p^2$ then arises from the inner product $\mathbf{p} \cdot \mathbf{p}$. Since $\mathcal{E}$ and $p$ are separately conserved, the energy-momentum four-vector is also conserved.
Energy and momentum of light example 21

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. 441.

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

Mass-energy, not energy, goes in the energy-momentum four-vector example 22

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 $\varepsilon$ 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 $\mathbf{p} = (\varepsilon, p_x, p_y, p_z)$, we would have had $\mathbf{p} = 0$ in this frame. But when we draw a zero vector, we get a point, and a point remains a point regardless of how we distort the graph paper we use to measure it. That wouldn't have made sense, because in other frames of reference, we have $E \neq 0$.

Metric units example 23

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

$$kg^2 = kg^2 \cdot \frac{m^4}{s^4} - kg^2 \cdot \frac{m^2}{s^2},$$

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 24

Example 18 on p. 417 discussed the annihilation of an electron and a positron into two gamma rays, which is an example of turn-
ing matter into pure energy. An opposite example is pair production, a process in which a gamma ray disappears, and its energy goes into creating an electron and a positron.

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

To see why pair production is forbidden in a vacuum, consider the process in the frame of reference in which the electron-positron pair has zero total momentum. In this frame, the gamma ray would have to have had zero momentum, but a gamma ray with zero momentum must have zero energy as well (example 21). 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 \( D_1 D_2 \), we must have \( f(D_1 D_2) = f(D_1) f(D_2) \). 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. 424 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 \( 2x E = ED(v) + ED(-v) \). 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 the \( p-E \) plane, massless particles lie on the two diagonals, while particles with mass lie to the right.
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. 387), 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. 389, the transformation must be area-preserving. We then have the same three cases to consider as in figure g on p. 388. 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. 423 equals 1 as claimed.)

Since \( m^2 = E^2 - p^2 \), it follows that for a material particle, \( p = m\gamma v \).
Postulates of Euclidean geometry:
1. Two points determine a line.
2. Line segments can be extended.
3. A unique circle can be constructed given any point as its center and any line segment as its radius.
4. All right angles are equal to one another.
5. Given a line and a point not on the line, no more than one line can be drawn through the point and parallel to the given line.

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 15, p. 416). We therefore know that our universe is noneuclidean, and we gain the further insight that the level of \(^7\)Einstein originally described the distinction between the two theories by saying that the special theory applied to nonaccelerating frames of reference, while the general one allowed any frame at all. The modern consensus is that Einstein was misinterpreting his own theory, and that special relativity actually handles accelerating frames just fine.
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.

**Einstein’s ring example 25**

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 $2 \times 10^{-3}$ 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°. 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°. 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 $180^\circ$.

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 straightedge. 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 to the case where spacetime is flat.

f / 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.

d / 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.

e / A triangle in a space with negative curvature has angles that add to less than $180^\circ$. 

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 straightedge. 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 to the case where spacetime is flat.

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.

f / 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.

d / 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.

e / A triangle in a space with negative curvature has angles that add to less than $180^\circ$. 

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 straightedge. 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 to the case where spacetime is flat.

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.

f / 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.

d / 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.

e / A triangle in a space with negative curvature has angles that add to less than $180^\circ$. 

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 straightedge. 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 to the case where spacetime is flat.

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.

f / 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.
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 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. 384, experiments show that a clock at the top of a mountain runs faster
1. A ray of light is emitted upward from the floor of the elevator. The elevator accelerates upward. 2. 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.

In the mountain-valley experiment, the frequency of the clock in the valley therefore appears to be running too slowly by a factor of \(1 - gh/c^2\) when it is compared via radio with the clock at the top of the mountain. We conclude that time runs more slowly when one is lower down in a gravitational field, and the slow-down factor between two points is given by \(1 - gh/c^2\), where \(h\) is the difference in height.

We have built up a picture of light rays interacting with gravity. To confirm that this makes sense, recall that we have already observed in subsection 7.3.3 and in problem 11 on p. 441 that light has momentum. The equivalence principle says that whatever has inertia must also participate in gravitational interactions. Therefore light waves must have weight, and must lose energy when they rise through a gravitational field.

**Local flatness**

The noneuclidean nature of spacetime produces effects that grow in proportion to the area of the region being considered. Interpreting such effects as evidence of curvature, we see that this connects naturally to the idea that curvature is undetectable from close up. For example, the curvature of the earth’s surface is not normally noticeable to us in everyday life. Locally, the earth’s surface is flat, and the same is true for spacetime.

Local flatness turns out to be another way of stating the equivalence principle. In a variation on the alien-abduction story, suppose that you regain consciousness aboard the flying saucer and find yourself weightless. If the equivalence principle holds, then
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?

Wouldn’t it be nice if we could define the meaning of a Newtonian inertial frame of reference? Newton makes it sound easy: to define an inertial frame, just find some object that is not accelerating because it is not being acted on by any external forces. But what object would we use? The earth? The “fixed stars”? Our galaxy? Our supercluster of galaxies? All of these are accelerating — relative to something.

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 \( v = \sqrt{2GM/r} \), 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. 442).

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 7.4 shows what would really happen.
The equivalence principle tells us that spacetime locally has the same structure as in special relativity, so we can draw the familiar parallelogram of $x - t$ coordinates at each point near the black hole. Superimposed on each little grid is a pair of lines representing motion at the speed of light in both directions, inward and outward. Because spacetime is curved, these lines do not appear to be at 45-degree angles, but to an observer in that region, they would appear to be. When light rays are emitted inward and outward from a point outside the event horizon, one escapes and one plunges into the black hole. On this diagram, they look like they are decelerating and accelerating, but local observers comparing them to their own coordinate grids would always see them as moving at exactly $c$. When rays are emitted from a point inside the event horizon, neither escapes; the distortion is so severe that “outward” is really inward.

Although the light rays in figure n don’t speed up or slow down, they do experience gravitational Doppler shifts. If a light ray is emitted from just above the event horizon, then it will escape to an infinite distance, but it will suffer an extreme Doppler shift toward low frequencies. A distant observer also has the option of interpreting this as a gravitational time dilation that greatly lowers the frequency of the oscillating electric charges that produced the ray. If the point of emission is made closer and closer to the horizon, the frequency and energy measured by a distant observer approach zero, making the ray impossible to observe.

*Information paradox*

Black holes have some disturbing implications for the kind of universe that in the Age of the Enlightenment was imagined to have been set in motion initially and then left to run forever like clockwork.

Newton’s laws have built into them the implicit assumption that omniscience is possible, at least in principle. For example, Newton’s definition of an inertial frame of reference leads to an infinite regress, as described on p. 431. For Newton this isn’t a problem, because in principle an omniscient observer can know the location of every mass in the universe. In this conception of the cosmos, there are no theoretical limits on human knowledge, only practical ones; if we could gather sufficiently precise data about the state of the universe at one time, and if we could carry out all the calculations to extrapolate into the future, then we could know everything that would ever happen. (See the famous quote by Laplace on p. 16.)

But the existence of event horizons surrounding black holes makes
In Newtonian contexts, physicists and astronomers had a correct intuition that it’s hard for things to collapse gravitationally. This star cluster has been around for about 15 billion years, but it hasn’t collapsed into a black hole. If any individual star happens to head toward the center, conservation of angular momentum tends to cause it to swing past and fly back out. The Penrose singularity theorem tells us that this Newtonian intuition is wrong when applied to an object that has collapsed past a certain point.

Furthermore, a black hole has at its center an infinitely dense point, called a singularity, containing all its mass, and this implies that information can be destroyed and made inaccessible to any observer at all. For example, suppose that astronaut Alice goes on a suicide mission to explore a black hole, free-falling in through the event horizon. She has a certain amount of time to collect data and satisfy her intellectual curiosity, but then she impacts the singularity and is compacted into a mathematical point. Now astronaut Betty decides that she will never be satisfied unless the secrets revealed to Alice are known to her as well — and besides, she was Alice’s best friend, and she wants to know whether Alice had any last words. Betty can jump through the horizon, but she can never know Alice’s last words, nor can any other observer who jumps in after Alice does.

This destruction of information is known as the black hole information paradox, and it’s referred to as a paradox because quantum physics (ch. 13) has built into its DNA the requirement that information is never lost in this sense.

**Formation**

Around 1960, as black holes and their strange properties began to be better understood and more widely discussed, many physicists who found these issues distressing comforted themselves with the belief that black holes would never really form from realistic initial conditions, such as the collapse of a massive star. Their skepticism was not entirely unreasonable, since it is usually very hard in astronomy to hit a gravitating target, the reason being that conservation of angular momentum tends to make the projectile swing past. (See problem 13 on p. 289 for a quantitative analysis.) For example, if we wanted to drop a space probe into the sun, we would have to extremely precisely stop its sideways orbital motion so that it would drop almost exactly straight in. Once a star started to collapse, the theory went, and became relatively compact, it would be such a small target that further infalling material would be unlikely to hit it, and the process of collapse would halt. According to this point of view, theorists who had calculated the collapse of a star into a black hole had been oversimplifying by assuming a star that was initially perfectly spherical and nonrotating. Remove the unrealistically perfect symmetry of the initial conditions, and a black hole would never actually form.

But Roger Penrose proved in 1964 that this was wrong. In fact, once an object collapses to a certain density, the Penrose singularity theorem guarantees mathematically that it will collapse further until a singularity is formed, and this singularity is surrounded by an event horizon. Since the brightness of an object like Sagittarius A* is far too low to be explained unless it has an event horizon (the
interstellar gas flowing into it would glow due to frictional heating), we can be certain that there really is a singularity at its core.

7.4.4 Cosmology

The Big Bang

Subsection 6.1.5 presented the evidence, discovered by Hubble, that the universe is expanding in the aftermath of the Big Bang: when we observe the light from distant galaxies, it is always Doppler-shifted toward the red end of the spectrum, indicating that no matter what direction we look in the sky, everything is rushing away from us. This seems to go against the modern attitude, originated by Copernicus, that we and our planet do not occupy a special place in the universe. Why is everything rushing away from our planet in particular? But general relativity shows that this anti-Copernican conclusion is wrong. General relativity describes space not as a rigidly defined background but as something that can curve and stretch, like a sheet of rubber. We imagine all the galaxies as existing on the surface of such a sheet, which then expands uniformly. The space between the galaxies (but not the galaxies themselves) grows at a steady rate, so that any observer, inhabiting any galaxy, will see every other galaxy as receding. There is therefore no privileged or special location in the universe.

We might think that there would be another kind of special place, which would be the one at which the Big Bang happened. Maybe someone has put a brass plaque there? But general relativity doesn’t describe the Big Bang as an explosion that suddenly occurred in a preexisting background of time and space. According to general relativity, space itself came into existence at the Big Bang, and the hot, dense matter of the early universe was uniformly distributed everywhere. The Big Bang happened everywhere at once.

Observations show that the universe is very uniform on large scales, and for ease of calculation, the first physical models of the expanding universe were constructed with perfect uniformity. In these models, the Big Bang was a singularity. This singularity can’t even be included as an event in spacetime, so that time itself only exists after the Big Bang. A Big Bang singularity also creates an even more acute version of the black hole information paradox. Whereas matter and information disappear into a black hole singularity, stuff pops out of a Big Bang singularity, and there is no physical principle that could predict what it would be.

As with black holes, there was considerable skepticism about whether the existence of an initial singularity in these models was an artifact of the unrealistically perfect uniformity assumed in the models. Perhaps in the real universe, extrapolation of all the paths of the galaxies backward in time would show them missing each other by millions of light-years. But in 1972 Stephen Hawking proved
An expanding universe with positive spatial curvature can be imagined as a balloon being blown up. Every galaxy's distance from every other galaxy increases, but no galaxy is the center of the expansion.

A variant on the Penrose singularity theorem that applied to Big Bang singularities. By the Hawking singularity theorem, the level of uniformity we see in the present-day universe is more than sufficient to prove that a Big Bang singularity must have existed.

The cosmic censorship hypothesis

It might not be too much of a philosophical jolt to imagine that information was spontaneously created in the Big Bang. Setting up the initial conditions of the entire universe is traditionally the prerogative of God, not the laws of physics. But there is nothing fundamental in general relativity that forbids the existence of other singularities that act like the Big Bang, being information producers rather than information consumers. As John Earman of the University of Pittsburgh puts it, anything could pop out of such a singularity, including green slime or your lost socks. This would eliminate any hope of finding a universal set of laws of physics that would be able to make a prediction given any initial situation.

That would be such a devastating defeat for the enterprise of physics that in 1969 Penrose proposed an alternative, humorously named the “cosmic censorship hypothesis,” which states that every singularity in our universe, other than the Big Bang, is hidden behind an event horizon. Therefore if green slime spontaneously pops out of one, there is limited impact on the predictive ability of physics, since the slime can never have any causal effect on the outside world. A singularity that is not modestly cloaked behind an event horizon is referred to as a naked singularity. Nobody has yet been able to prove the cosmic censorship hypothesis.

The advent of high-precision cosmology

We expect that if there is matter in the universe, it should have gravitational fields, and in the rubber-sheet analogy this should be represented as a curvature of the sheet. Instead of a flat sheet, we can have a spherical balloon, so that cosmological expansion is like inflating it with more and more air. It is also possible to have negative curvature, as in figure e on p. 427. All three of these are valid, possible cosmologies according to relativity. The positive-curvature type happens if the average density of matter in the universe is above a certain critical level, the negative-curvature one if the density is below that value.

To find out which type of universe we inhabit, we could try to take a survey of the matter in the universe and determine its average density. Historically, it has been very difficult to do this, even to within an order of magnitude. Most of the matter in the universe probably doesn’t emit light, making it difficult to detect. Astronomical distance scales are also very poorly calibrated against absolute units such as the SI.

Instead, we measure the universe's curvature, and infer the den-
The angular scale of fluctuations in the cosmic microwave background can be used to infer the curvature of the universe. It turns out that we can do this by observing the cosmic microwave background (CMB) radiation, which is the light left over from the brightly glowing early universe, which was dense and hot. As the universe has expanded, light waves that were in flight have expanded their wavelengths along with it. This afterglow of the big bang was originally visible light, but after billions of years of expansion it has shifted into the microwave radio part of the electromagnetic spectrum. The CMB is not perfectly uniform, and this turns out to give us a way to measure the universe’s curvature. Since the CMB was emitted when the universe was only about 400,000 years old, any vibrations or disturbances in the hot hydrogen and helium gas that filled space in that era would only have had time to travel a certain distance, limited by the speed of sound. We therefore expect that no feature in the CMB should be bigger than a certain known size. In a universe with negative spatial curvature, the sum of the interior angles of a triangle is less than the Euclidean value of 180 degrees. Therefore if we observe a variation in the CMB over some angle, the distance between two points on the sky is actually greater than would have been inferred from Euclidean geometry. The opposite happens if the curvature is positive.

This observation was done by the 1989-1993 COBE probe, and its 2001-2009 successor, the Wilkinson Microwave Anisotropy Probe. The result is that the angular sizes are almost exactly equal to what they should be according to Euclidean geometry. We therefore infer that the universe is very close to having zero average spatial curvature on the cosmological scale, and this tells us that its average density must be within about 0.5% of the critical value. The years since COBE and WMAP mark the advent of an era in which cosmology has gone from being a field of estimates and rough guesses to a high-precision science.

If one is inclined to be skeptical about the seemingly precise answers to the mysteries of the cosmos, there are consistency checks that can be carried out. In the bad old days of low-precision cosmology, estimates of the age of the universe ranged from 10 billion to 20 billion years, and the low end was inconsistent with the age of the oldest star clusters. This was believed to be a problem either for observational cosmology or for the astrophysical models used to estimate the ages of the clusters: “You can’t be older than your ma.” Current data have shown that the low estimates of the age were incorrect, so consistency is restored. (The best figure for the age of the universe is currently 13.8 ± 0.1 billion years.)

Dark energy and dark matter

Not everything works out so smoothly, however. One surprises is that the universe’s expansion is not currently slowing down, as had been expected due to the gravitational attraction of all the matter...
in it. Instead, it is currently speeding up. This is attributed to a variable in Einstein's equations, long assumed to be zero, which represents a universal gravitational repulsion of space itself, occurring even when there is no matter present. The current name for this is “dark energy,” although the fancy name is just a label for our ignorance about what causes it.

Another surprise comes from attempts to model the formation of the elements during the era shortly after the Big Bang, before the formation of the first stars. The observed relative abundances of hydrogen, helium, and deuterium \((^2H)\) cannot be reconciled with the density of low-velocity matter inferred from the observational data. If the inferred mass density were entirely due to normal matter (i.e., matter whose mass consisted mostly of protons and neutrons), then nuclear reactions in the dense early universe should have proceeded relatively efficiently, leading to a much higher ratio of helium to hydrogen, and a much lower abundance of deuterium. The conclusion is that most of the matter in the universe must be made of an unknown type of exotic matter, known as “dark matter.” We are in the ironic position of knowing that precisely 96% of the universe is something other than atoms, but knowing nothing about what that something is. As of 2013, there have been several experiments that have been carried out to attempt the direct detection of dark matter particles. These are carried out at the bottom of mineshafts to eliminate background radiation. Early claims of success appear to have been statistical flukes, and the most sensitive experiments have not detected anything.\(^8\)

\(^8\text{arxiv.org/abs/1310.8214}\)
Problems

The symbols \(\sqrt{\ \ }, =, \) etc. are explained on page 445.

1. The figure illustrates a Lorentz transformation using the conventions employed in section 7.2. For simplicity, the transformation chosen is one that lengthens one diagonal by a factor of 2. Since Lorentz transformations preserve area, the other diagonal is shortened by a factor of 2. Let the original frame of reference, depicted with the square, be A, and the new one B. (a) By measuring with a ruler on the figure, show that the velocity of frame B relative to frame A is 0.6c. (b) Print out a copy of the page. With a ruler, draw a third parallelogram that represents a second successive Lorentz transformation, one that lengthens the long diagonal by another factor of 2. Call this third frame C. Use measurements with a ruler to determine frame C’s velocity relative to frame A. Does it equal double the velocity found in part a? Explain why it should be expected to turn out the way it does.
2. Astronauts in three different spaceships are communicating with each other. Those aboard ships A and B agree on the rate at which time is passing, but they disagree with the ones on ship C.
(a) Alice is aboard ship A. How does she describe the motion of her own ship, in its frame of reference?
(b) Describe the motion of the other two ships according to Alice.
(c) Give the description according to Betty, whose frame of reference is ship B.
(d) Do the same for Cathy, aboard ship C.

3. What happens in the equation for $\gamma$ when you put in a negative number for $v$? Explain what this means physically, and why it makes sense.

4. The Voyager 1 space probe, launched in 1977, is moving faster relative to the earth than any other human-made object, at 17,000 meters per second.
(a) Calculate the probe’s $\gamma$.
(b) Over the course of one year on earth, slightly less than one year passes on the probe. How much less? (There are 31 million seconds in a year.)

5. In example 2 on page 391, I remarked that accelerating a macroscopic (i.e., not microscopic) object to close to the speed of light would require an unreasonable amount of energy. Suppose that the starship Enterprise from Star Trek has a mass of $8.0 \times 10^7$ kg, about the same as the Queen Elizabeth 2. Compute the kinetic energy it would have to have if it was moving at half the speed of light. Compare with the total energy content of the world’s nuclear arsenals, which is about $10^{21}$ J.

6. The earth is orbiting the sun, and therefore is contracted relativistically in the direction of its motion. Compute the amount by which its diameter shrinks in this direction.

7. In this homework problem, you’ll fill in the steps of the algebra required in order to find the equation for $\gamma$ on page 389. To keep the algebra simple, let the time $t$ in figure k equal 1, as suggested in the figure accompanying this homework problem. The original square then has an area of 1, and the transformed parallelogram must also have an area of 1. (a) Prove that point P is at $x = v\gamma$, so that its $(t, x)$ coordinates are $(\gamma, v\gamma)$. (b) Find the $(t, x)$ coordinates of point Q. (c) Find the length of the short diagonal connecting P and Q. (d) Average the coordinates of P and Q to find the coordinates of the midpoint C of the parallelogram, and then find distance OC. (e) Find the area of the parallelogram by computing twice the area of triangle PQO. [Hint: You can take PQ to be the base of the triangle.] (f) Set this area equal to 1 and solve for $\gamma$ to prove $\gamma = 1/\sqrt{1 - v^2}$.
(a) A free neutron (as opposed to a neutron bound into an atomic nucleus) is unstable, and undergoes beta decay (which you may want to review). The masses of the particles involved are as follows:

<table>
<thead>
<tr>
<th>Particle</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>neutron</td>
<td>$1.67495 \times 10^{-27}$ kg</td>
</tr>
<tr>
<td>proton</td>
<td>$1.67265 \times 10^{-27}$ kg</td>
</tr>
<tr>
<td>electron</td>
<td>$0.00091 \times 10^{-27}$ kg</td>
</tr>
<tr>
<td>antineutrino</td>
<td>$&lt; 10^{-35}$ kg</td>
</tr>
</tbody>
</table>

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

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

If there are processes that can convert neutrons into protons, we might imagine that there could also be proton-to-neutron conversions, and indeed such a process does occur sometimes in nuclei that contain both neutrons and protons: a proton can decay into a neutron, a positron, and a neutrino. A positron is a particle with the same properties as an electron, except that its electrical charge is positive (see chapter 7). A neutrino, like an antineutrino, has negligible mass.

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

(a) Find a relativistic equation for the velocity of an object in terms of its mass and momentum (eliminating $\gamma$).

(b) Show that your result is approximately the same as the classical value, $p/m$, at low velocities.

(c) Show that very large momenta result in speeds close to the speed of light.

(a) Show that for $v = (3/5)c$, $\gamma$ comes out to be a simple fraction.

(b) Find another value of $v$ for which $\gamma$ is a simple fraction.

An object moving at a speed very close to the speed of light is referred to as ultrarelativistic. Ordinarily (luckily) the only ultrarelativistic objects in our universe are subatomic particles, such as cosmic rays or particles that have been accelerated in a particle accelerator.

(a) What kind of number is $\gamma$ for an ultrarelativistic particle?

(b) Repeat example 19 on page 419, but instead of very low, non-relativistic speeds, consider ultrarelativistic speeds.

(c) Find an equation for the ratio $E/p$. The speed may be relativistic, but don’t assume that it’s ultrarelativistic.
(d) Simplify your answer to part c for the case where the speed is ultrarelativistic.

(e) We can think of a beam of light as an ultrarelativistic object — it certainly moves at a speed that’s sufficiently close to the speed of light! Suppose you turn on a one-watt flashlight, leave it on for one second, and then turn it off. Compute the momentum of the recoiling flashlight, in units of kg·m/s.

(f) Discuss how your answer in part e relates to the correspondence principle.

12 As discussed in chapter 6, the speed at which a disturbance travels along a string under tension is given by \( v = \sqrt{T/\mu} \), where \( \mu \) is the mass per unit length, and \( T \) is the tension.

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

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

(c) A black hole is a star that has collapsed and become very dense, so that its gravity is too strong for anything ever to escape from it. For instance, the escape velocity from a black hole is greater than \( c \), so a projectile can’t be shot out of it. Many people, when they hear this description of a black hole in terms of an escape velocity greater than \( c \), wonder why it still wouldn’t be possible to extract an object from a black hole by other means than launching it out as a projectile. For example, suppose we lower an astronaut into a black hole on a rope, and then pull him back out again. Why might this not work?

13 (a) A charged particle is surrounded by a uniform electric field. Starting from rest, it is accelerated by the field to speed \( v \) after traveling a distance \( d \). Now it is allowed to continue for a further distance \( 3d \), for a total displacement from the start of \( 4d \). What speed will it reach, assuming classical physics?

(b) Find the relativistic result for the case of \( v = c/2 \).

14 Problem 14 has been deleted.
15 Expand the equation \( K = m(\gamma - 1) \) in a Taylor series, and find the first two nonvanishing terms. Show that the first term is the nonrelativistic expression for kinetic energy.

16 Expand the relativistic equation for momentum in a Taylor series, and find the first two nonvanishing terms. Show that the first term is the classical expression.

17 As promised in subsection 7.2.8, this problem will lead you through the steps of finding an equation for the combination of velocities in relativity, generalizing the numerical result found in problem 1. Suppose that A moves relative to B at velocity \( u \), and B relative to C at \( v \). We want to find A’s velocity \( w \) relative to C, in terms of \( u \) and \( v \). Suppose that A emits light with a certain frequency. This will be observed by B with a Doppler shift \( D(u) \). C detects a further shift of \( D(v) \) relative to B. We therefore expect the Doppler shifts to multiply, \( D(w) = D(u)D(v) \), and this provides an implicit rule for determining \( w \) if \( u \) and \( v \) are known. (a) Using the expression for \( D \) given in section 7.2.8, write down an equation relating \( u \), \( v \), and \( w \). (b) Solve for \( w \) in terms of \( u \) and \( v \). (c) Show that your answer to part b satisfies the correspondence principle.

\( \triangledown \) Solution, p. 940

18 The figure shows seven four-vectors, represented in a two-dimensional plot of \( x \) versus \( t \). All the vectors have \( y \) and \( z \) components that are zero. Which of these vectors are congruent to others, i.e., which represent spacetime intervals that are equal to one another?

19 Four-vectors can be timelike, lightlike, or spacelike. What can you say about the inherent properties of particles whose momentum four-vectors fall in these various categories?

20 The following are the three most common ways in which gamma rays interact with matter:

\textit{Photoelectric effect}: The gamma ray hits an electron, is annihilated, and gives all of its energy to the electron.

\textit{Compton scattering}: The gamma ray bounces off of an electron, exiting in some direction with some amount of energy.

\textit{Pair production}: The gamma ray is annihilated, creating an electron and a positron.

Example 24 on p. 420 shows that pair production can’t occur in a vacuum due to conservation of the energy-momentum four-vector. What about the other two processes? Can the photoelectric effect occur without the presence of some third particle such as an atomic nucleus? Can Compton scattering happen without a third particle?
21 Expand the relativistic equation for the longitudinal Doppler shift of light $D(v)$ in a Taylor series, and find the first two nonvanishing terms. Show that these two terms agree with the nonrelativistic expression, so that any relativistic effect is of higher order in $v$.

22 Prove, as claimed in the caption of figure a on p. 425, that $S - 180^\circ = 4(s - 180^\circ)$, where $S$ is the sum of the angles of the large equilateral triangle and $s$ is the corresponding sum for one of the four small ones.

23 If a two-dimensional being lived on the surface of a cone, would it say that its space was curved, or not?

24 (a) Verify that the equation $1 - gh/c^2$ for the gravitational Doppler shift and gravitational time dilation has units that make sense. (b) Does this equation satisfy the correspondence principle?

25 (a) Calculate the Doppler shift to be expected in the Pound-Rebka experiment described on p. 430. (b) In the 1978 Iijima mountain-valley experiment (p. 384), analysis was complicated by the clock’s sensitivity to pressure, humidity, and temperature. A cleaner version of the experiment was done in 2005 by hobbyist Tom Van Baak. He put his kids and three of his atomic clocks in a minivan and drove from Bellevue, Washington to a lodge on Mount Rainier, 1340 meters higher in elevation. At home, he compared the clocks to others that had stayed at his house. Verify that the effect shown in the graph is as predicted by general relativity.

26 The International Space Station orbits at an altitude of about 350 km and a speed of about 8000 m/s relative to the ground. Compare the gravitational and kinematic time dilations. Over all, does time run faster on the ISS than on the ground, or more slowly?

27 Section 7.4.3 presented a Newtonian estimate of how compact an object would have to be in order to be a black hole. Although this estimate is not really right, it turns out to give the right answer to within about a factor of 2. To roughly what size would the earth have to be compressed in order to become a black hole?

28 Clock A sits on a desk. Clock B is tossed up in the air from the same height as the desk and then comes back down. Compare the elapsed times.
The angular defect \( d \) of a triangle (measured in radians) is defined as \( s - \pi \), where \( s \) is the sum of the interior angles. The angular defect is proportional to the area \( A \) of the triangle. Consider the geometry measured by a two-dimensional being who lives on the surface of a sphere of radius \( R \). First find some triangle on the sphere whose area and angular defect are easy to calculate. Then determine the general equation for \( d \) in terms of \( A \) and \( R \).

Key to symbols:

- easy
- typical
- challenging
- difficult
- very difficult

\( \checkmark \) An answer check is available at www.lightandmatter.com.
In this exercise you will analyze the Michelson-Morley experiment, and find what the results should have been according to Galilean relativity and Einstein’s theory of relativity. A beam of light coming from the west (not shown) comes to the half-silvered mirror A. Half the light goes through to the east, is reflected by mirror C, and comes back to A. The other half is reflected north by A, is reflected by B, and also comes back to A. When the beams reunite at A, part of each ends up going south, and these parts interfere with one another. If the time taken for a round trip differs by, for example, half the period of the wave, there will be destructive interference.

The point of the experiment was to search for a difference in the experimental results between the daytime, when the laboratory was moving west relative to the sun, and the nighttime, when the laboratory was moving east relative to the sun. Galilean relativity and Einstein’s theory of relativity make different predictions about the results. According to Galilean relativity, the speed of light cannot be the same in all reference frames, so it is assumed that there is one special reference frame, perhaps the sun’s, in which light travels at the same speed in all directions; in other frames, Galilean relativity predicts that the speed of light will be different in different directions, e.g., slower if the observer is chasing a beam of light. There are four different ways to analyze the experiment:

- **Laboratory’s frame of reference, Galilean relativity.** This is not a useful way to analyze the experiment, since one does not know how fast light will travel in various directions.

- **Sun’s frame of reference, Galilean relativity.** We assume that in this special frame of reference, the speed of light is the same in all directions: we call this speed \( c \). In this frame, the laboratory moves with velocity \( v \), and mirrors A, B, and C move while the light beam is in flight.

- **Laboratory’s frame of reference, Einstein’s theory of relativity.** The analysis is extremely simple. Let the length of each arm be \( L \). Then the time required to get from A to either mirror is \( L/c \), so each beam’s round-trip time is \( 2L/c \).

- **Sun’s frame of reference, Einstein’s theory of relativity.** We analyze this case by starting with the laboratory’s frame of reference and then transforming to the sun’s frame.
Groups 1-4 work in the sun’s frame of reference according to Galilean relativity.

Group 1 finds time AC. Group 2 finds time CA. Group 3 finds time AB. Group 4 finds time BA.

Groups 5 and 6 transform the lab-frame results into the sun’s frame according to Einstein’s theory.

Group 5 transforms the $x$ and $t$ when ray ACA gets back to A into the sun’s frame of reference, and group 6 does the same for ray ABA.

Discussion:

Michelson and Morley found no change in the interference of the waves between day and night. Which version of relativity is consistent with their results?

What does each theory predict if $v$ approaches $c$?

What if the arms are not exactly equal in length?

Does it matter if the “special” frame is some frame other than the sun’s?
Exercise 7B: Sports in Slowlightland

In Slowlightland, the speed of light is 20 mi/hr \(\approx 32\) km/hr \(\approx 9\) m/s. Think of an example of how relativistic effects would work in sports. Things can get very complex very quickly, so try to think of a simple example that focuses on just one of the following effects:

- relativistic momentum
- relativistic kinetic energy
- relativistic addition of velocities
- time dilation and length contraction
- Doppler shifts of light
- equivalence of mass and energy
- time it takes for light to get to an athlete’s eye
- deflection of light rays by gravity
Exercise 7C: Events and Spacetime

Bell's spaceship paradox

A difficult philosophical question is whether the time dilation and length contractions predicted by relativity are "real." This depends, of course, on what one means by "real." They are frame-dependent, i.e., observers in different frames of reference disagree about them. But this doesn't tell us much about their reality, since velocities are frame-dependent in Newtonian mechanics, but nobody worries about whether velocities are real. John Bell (1928-1990) proposed the following thought experiment to physicists in the CERN cafeteria, and found that nearly all of them got it wrong. He took this as evidence that their intuitions had been misguided by the standard way of approaching this question of the reality of Lorentz contractions.

Let spaceships A and B accelerate as shown in the figure, along a straight line. Observer C, whose frame of reference is indicated by the square, does not accelerate. The accelerations, as judged by C, are constant, and equal for the two ships. Each ship is equipped with a yard-arm, and a thread is tied between the two arms. Does the thread break, due to length contraction?

Use the spacetime diagram below to find the answer.

What does C say about the distance between A and B as time goes by? You may want to use a ruler to verify that the tracks in the diagram do correspond properly to the statement of the paradox.

At what time does the parallelogram correspond to A's frame of reference? B's?

What does an observer in the parallelogram frame say about the distance between A and B as time goes by?

What would the result be if, instead of the Lorentz transformation, we used the Galilean version?
The following paper-and-pencil exercises involve graphs of position versus time. These are often referred to as spacetime diagrams, and tracks on them are called world-lines. As in the other examples in this book, a time axis is always closer to horizontal, and a position axis closer to vertical, and the units are such that the universal speed is a slope of $+1$ or $-1$.

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

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

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