When we obtained Faraday’s law from Maxwell’s equations and Stokes’s theorem, we assumed implicitly that once we picked a surface, it stayed the same. The only change over time was the change in the magnetic field. But clearly if we viewed the situation in another frame of reference, the surface would be moving, and we expect our laws of physics to be valid regardless of the frame of reference. It is OK for the surface to move, and it can even change size or shape. This is intuitively appealing if we think of the “cutting” interpretation.

Faraday arrived at what we now call Faraday’s law completely empirically in 1831. It’s fascinating to read about the original trail of evidence that he followed. Figure g is a simplified drawing of a crucial experiment, as described in his original paper: “Two hundred and three feet of copper wire ... were passed round a large block of wood; [another] two hundred and three feet of similar wire were interposed as a spiral between the turns of the first, and metallic contact everywhere prevented by twine [insulation]. One of these [coils] was connected with a galvanometer [voltmeter], and the other with a battery... When the contact was made, there was a sudden and very slight effect at the galvanometer, and there was also a similar slight effect when the contact with the battery was broken. But whilst the ... current was continuing to pass through the one [coil], no ... effect ... upon the other [coil] could be perceived, although the active power of the battery was proved to be great, by its heating the whole of its own coil [through ordinary resistive heating] ...”

From Faraday’s notes and publications, it appears that the situation in figure g/3 was a surprise to him, and he probably thought it would be a surprise to his readers, as well. That’s why he offered evidence that the current was still flowing: to show that the battery hadn’t just died. The induction effect occurred during the short time it took for the black coil’s magnetic field to be established, g/2. Even more counterintuitively, we get an effect, equally strong but in the opposite direction, when the circuit is broken, g/4. The effect occurs only when the magnetic field is changing, and it appears to be proportional to the rate of change of the magnetic flux through the block (or actually through the corkscrew-shaped surface formed by the white wire), which has one sign when the field is being established, and in the opposite direction when it collapses.

Although Faraday’s discovery was empirical, Faraday’s law has close logical interrelationships with other principles of physics. In example 5 we see that induction is necessary based on conservation of energy, and in example 6 that it is necessary based on the fact that motion is relative.
Shorting across an inductor, example 5

Panel 2 shows what is observed immediately after the resistances are changed.

Conservation of energy, example 5

Figure h/1 shows a solenoid with a current being driven through it by a battery. The wire that the solenoid is made of has some finite resistance, so this is not a short circuit, but current is flowing. There is a magnetic field in and around the solenoid. The magnetic compass near the mouth of the solenoid is nearly aligned with the solenoid’s axis, showing that this field is much stronger than any ambient field such as the earth’s.

So far there has been no obvious reason for having the two variable resistors. The one set to $R = 0$ might as well be a piece of wire, and the one set to $R = \infty$ could just be air. But now suppose that we rapidly change the resistors so that they have the values shown in h/2. In fact, I have a power supply in my lab that seems to do essentially this when I flip its switch to the “off” position.

If our intuition is based solely on experience with DC circuits, then we would expect that the current would instantly cease. The resistance that is now infinite is an open circuit, which means that the power supply is disconnected from the circuit. We have shorted across the inductor, and we know that if we short across a light-bulb, it just winks out.

But that is not at all what happens. The compass stays aligned with the solenoid, showing that the field still exists. Only very slowly does it relax to alignment with the direction of the ambient field.

Although this is a little surprising, it becomes easier to understand when we consider that the field in circuit 1 had energy. Therefore the field can’t just go poof. It has to have some mechanism for transforming that energy into some other form. The only mechanism available for doing that is resistive heating in the coil. But this will take some time, as measured by the RL time constant of circuit 2. During this time, the field and the current just gradually die out.

To push this current through the circuit, which has some resistance, we will need an electric field, even though the battery has been taken out of action. This electric field exists due to Faraday’s law, and the minus sign in Faraday’s law is interpreted as saying that the field is in the direction that tends to resist the change in the magnetic field.

Frames of reference, example 6

In figure i, flea 1 doesn’t believe in this modern foolishness about induction. She’s sitting on the bar magnet, which to her is obviously at rest. As the square wire loop is dragged away from her and the magnet, its protons experience a force out of the page, because the cross product $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$ is out of the page. The electrons, which are negatively charged, feel a force into the page.
page. The conduction electrons are free to move, but the protons aren’t. In the front and back sides of the loop, this force is perpendicular to the wire. In the right and left sides, however, the electrons are free to respond to the force. Note that the magnetic field is weaker on the right side. It’s as though we had two pumps in a loop of pipe, with the weaker pump trying to push in the opposite direction; the weaker pump loses the argument. We get a current that circulates around the loop. There is no induction going on in this frame of reference; the forces that cause the current are just the ordinary magnetic forces experienced by any charged particle moving through a magnetic field.

Flea 2 is sitting on the loop, which she considers to be at rest. In her frame of reference, it’s the bar magnet that is moving. Like flea 1, she observes a current circulating around the loop, but unlike flea 1, she cannot use magnetic forces to explain this current. As far as she is concerned, the electrons were initially at rest. Magnetic forces are forces between moving charges and other moving charges, so a magnetic field can never accelerate a charged particle starting from rest. A force that accelerates a charge from rest can only be an electric force, so she is forced to conclude that there is an electric field in her region of space. This field drives electrons around and around in circles, so it is apparently violating the loop rule — it is a curly field. What reason can flea 2 offer for the existence of this electric field pattern? Well, she’s been noticing that the magnetic field in her region of space has been changing, possibly because that bar magnet over there has been getting farther away. She observes that a changing magnetic field creates a curly electric field.

We therefore conclude that induction effects must exist based on the fact that motion is relative. If we didn’t want to admit induction effects, we would have to outlaw flea 2’s frame of reference, but the whole idea of relative motion is that all frames of reference are created equal, and there is no way to determine which one is really at rest.

---

1If the pump analogy makes you uneasy, consider what would happen if all the electrons moved into the page on both sides of the loop. We’d end up with a net negative charge at the back side, and a net positive charge on the front. This actually would happen in the first nanosecond after the loop was set in motion. This buildup of charge would start to quench both currents due to electrical forces, but the current in the right side of the wire, which is driven by the weaker magnetic field, would be the first to stop. Eventually, an equilibrium will be reached in which the same amount of current is flowing at every point around the loop, and no more charge is being piled up.

2The wire is not a perfect conductor, so this current produces heat. The energy required to produce this heat comes from the hands, which are doing mechanical work as they separate the magnet from the loop.
Discussion question

A 1. The figure shows a line of charges moving to the right, creating a current $I$. An Ampèrian surface in the form of a disk has been superimposed. Use Maxwell’s equations to find the field $B$ at point $P$.

2. A tiny gap is chopped out of the line of charge. What happens when this gap is directly underneath the point $P$?
Problems

Key
√ A computerized answer check is available online.
★ A difficult problem.

1 The equation \( B = \left(\frac{k}{c^2}\right) \cdot \frac{2I}{r} \) for the magnetic field of a long, straight wire was derived in in examples 2, p. 117, (the form of the equation) and 6, p. 179 (the factor of 2). Derive the equation using Ampère’s law. (Cf. also problem 9, p. 279, using the Biot-Savart law.)

2 The figure shows a sheet of current coming out of the page. Such a sheet can be characterized by a linear current density \( \eta \), which has units of A/m. (The letter \( \eta \) is Greek eta, which makes the “ee” sound in modern Greek.) The figure shows magnetic field vectors in the \( \pm x \) directions, with equal magnitudes above and below the sheet. In general, however, this symmetry need not exist. We could always add a constant magnetic field to the whole field pattern and get another equally valid solution of Maxwell’s equations. The only thing we can actually determine is \( \Delta B_x \), the difference in the horizontal field between the top and bottom of the sheet.
   (a) Find \( \Delta B_x \) in terms of \( \eta \).
   (b) In the symmetric case shown in the figure, what can you say about the pressure or tension experienced by the sheet, using the visual modes of reasoning from sec. 5.2.2, p. 122? If the sheet has no structural strength, what will it do?

3 A U-shaped wire makes electrical contact with a second, straight wire, of length \( \ell \), which rolls along it to the right at velocity \( v \), as shown in the figure. The whole thing is immersed in a uniform magnetic field \( B \), which is perpendicular to the plane of the circuit. The resistance \( R \) of the rolling wire is much greater than that of the U.
   (a) Use Faraday’s law to find the amount of current through the wire, and its direction.
   (b) Use conservation of energy to find the direction of the force on the wire.
   (c) Verify the direction of the force using right-hand rules.
   (d) Find the magnitude of the force acting on the wire.
   (e) Consider how the answer to part a would have changed if the direction of the field had been reversed, and also do the case where the direction of the rolling wire’s motion is reversed. Verify that this is in agreement with your answer to part b.
4 The circular parallel-plate capacitor shown in the figure is being charged up over time, with the voltage difference across the plates varying as $V = st$, where $s$ is a constant. The plates have radius $b$, and the distance between them is $d$. We assume $d \ll b$, so that the electric field between the plates is uniform, and parallel to the axis. Find the induced magnetic field at a point between the plates, at a distance $R$ from the axis. $\triangleright$ Hint, p. 423 $\checkmark$

Problems 5-6 require enough knowledge of vector calculus to evaluate line and surface integrals with integrands that aren’t constant.

5 A wire loop of resistance $R$ and area $A$, lying in the $y-z$ plane, falls through a nonuniform magnetic field $B = k z \hat{x}$, where $k$ is a constant. The $z$ axis is vertical.
(a) Find the direction of the force on the wire based on conservation of energy.
(b) Verify the direction of the force using right-hand rules.
(c) Find the magnetic force on the wire. $\checkmark$

6 Verify Ampère’s law in the case shown in the figure, assuming the known equation for the field of a wire. A wire carrying current $I$ passes perpendicularly through the center of the rectangular Ampérien surface. The length of the rectangle is infinite, so it’s not necessary to compute the contributions of the ends.

7 A certain electrical transmission line, shown in cross-section, consists of two hollow, coaxial pipes. Let $r$ be the distance from the axis. The inner pipe is at $r = a$ and the outer at $r = b$. The inner pipe carries current $I$, and the outer $-I$, i.e., the transmission line is part of a complete circuit, and the current flows out through one conductor and back through the other. Find the magnitude of the magnetic field (a) for $r < a$, (b) for $a < r < b$, and (c) for $r > b$.

8 A cylindrical wire carries a current that is not uniformly distributed across its cross-section. The current density is a function $j(r)$ of the distance from the axis. Show that if the magnetic field is known as a function $B(r)$, then the current density can be determined.
Electromagnetic properties of materials
Chapter 16
Electromagnetic properties of materials

Different types of matter have a variety of useful electrical and magnetic properties. Some are conductors, and some are insulators. Some, like iron and nickel, can be magnetized, while others have useful electrical properties, e.g., dielectrics, discussed qualitatively in the discussion question on page 311, which allow us to make capacitors with much higher values of capacitance than would otherwise be possible. We need to organize our knowledge about the properties that materials can possess, and see whether this knowledge allows us to calculate anything useful with Maxwell’s equations.

16.1 Conductors

A perfect conductor, such as a superconductor, has no DC electrical resistance. It is not possible to have a static electric field inside it, because then charges would move in response to that field, and the motion of the charges would tend to reduce the field, contrary to the assumption that the field was static. Things are a little different at the surface of a perfect conductor than on the interior. We expect that any net charges that exist on the conductor will spread out under the influence of their mutual repulsion, and settle on the surface. Gauss’s law requires that the fields on the two sides of a sheet of charge have \(|E_{\perp,1} - E_{\perp,2}|\) proportional to the surface charge density, and since the field inside the conductor is zero, we infer that there can be a field on or immediately outside the conductor, with a nonvanishing component perpendicular to the surface. The component of the field parallel to the surface must vanish, however, since otherwise it would cause the charges to move along the surface.

On a hot summer day, the reason the sun feels warm on your skin is that the oscillating fields of the light waves excite currents in your skin, and these currents dissipate energy by ohmic heating. In a perfect conductor, however, this could never happen, because there is no such thing as ohmic heating. Since electric fields can’t penetrate a perfect conductor, we also know that an electromagnetic wave can never pass into one. By conservation of energy, we know that the wave can’t just vanish, and if the energy can’t be dissipated as heat, then the only remaining possibility is that all of the wave’s energy is reflected. This is why metals, which are good
A capacitor with a dielectric between the plates.

electrical conductors, are also highly reflective. They are not perfect electrical conductors, however, so they are not perfectly reflective. The wave enters the conductor, but immediately excites oscillating currents, and these oscillating currents dissipate the energy both by ohmic heating and by reradiating the reflected wave. Since the parts of Maxwell’s equations describing radiation have time derivatives in them, the efficiency of this reradiation process depends strongly on frequency. When the frequency is high and the material is a good conductor, reflection predominates, and is so efficient that the wave only penetrates to a very small depth, called the skin depth. In the limit of poor conduction and low frequencies, absorption predominates, and the skin depth becomes much greater. In a high-frequency AC circuit, the skin depth in a copper wire is very small, and therefore the signals in such a circuit are propagated entirely at the surfaces of the wires. In the limit of low frequencies, i.e., DC, the skin depth approaches infinity, so currents are carried uniformly over the wires’ cross-sections.

We can quantify how well a particular material conducts electricity. We know that the resistance of a wire is proportional to its length, and inversely proportional to its cross-sectional area. The constant of proportionality is $1/\sigma$, where $\sigma$ (not the same $\sigma$ as the surface charge density) is called the electrical conductivity. Exposed to an electric field $E$, a conductor responds with a current per unit cross-sectional area $J = \sigma E$. The skin depth is proportional to $1/\sqrt{f\sigma}$, where $f$ is the frequency of the wave.

16.2 Dielectrics

A material with a very low conductivity is an insulator. Such materials are usually composed of atoms or molecules whose electrons are strongly bound to them; since the atoms or molecules have zero total charge, their motion cannot create an electric current. But even though they have zero charge, they may not have zero dipole moment. Imagine such a substance filling in the space between the plates of a capacitor, as in figure a. For simplicity, we assume that the molecules are oriented randomly at first, a/1, and then become completely aligned when a field is applied, a/2. The effect has been to take all of the negatively charged black ends of the molecules and shift them upward, and the opposite for the positively charged white ends. Where the black and white charges overlap, there is still zero net charge, but we have a strip of negative charge at the top, and a strip of positive charge at the bottom, a/3. The effect has been to cancel out part of the charge that was deposited on the plates of the capacitor. Now this is very subtle, because Maxwell’s equations treat these charges on an equal basis, but in terms of practical measurements, they are completely different. The charge on the plates can be measured be inserting an ammeter in the circuit, and inte-
grating the current over time. But the charges in the layers at the
top and bottom of the dielectric never flowed through any wires, and
cannot be detected by an ammeter. In other words, the total charge,
$q$, appearing in Maxwell’s equations is actually $q = q_{\text{free}} - q_{\text{bound}}$, where $q_{\text{free}}$ is the charge that moves freely through wires, and can be detected in an ammeter, while $q_{\text{bound}}$ is the charge bound onto the individual molecules, which can’t. We will, however, detect the presence of the bound charges via their electric fields. Since their electric fields partially cancel the fields of the free charges, a voltmeter will register a smaller than expected voltage difference between the plates. If we measure $q_{\text{free}}/V$, we have a result that is larger than the capacitance we would have expected.

Although the relationship $E \leftrightarrow q$ between electric fields and their sources is unalterably locked in by Gauss’s law, that’s not what we see in practical measurements. In this example, we can measure the voltage difference between the plates of the capacitor and divide by the distance between them to find $E$, and then integrate an ammeter reading to find $q_{\text{free}}$, and we will find that Gauss’s law appears not to hold. We have $E \leftrightarrow q_{\text{free}}/(\text{constant})$, where the constant fudge factor is greater than one. This constant is a property of the dielectric material, and tells us how many dipoles there are, how strong they are, and how easily they can be reoriented. The conventional notation is to incorporate this fudge factor into Gauss’s law by defining an altered version of the electric field,

$$D = \varepsilon E,$$

and to rewrite Gauss’s law as

$$\Phi_D = q_{\text{in, free}}.$$

The constant $\varepsilon$ is a property of the material, known as its permittivity. In a vacuum, $\varepsilon$ takes on a value known as $\varepsilon_0$, defined as $1/(4\pi k)$. In a dielectric, $\varepsilon$ is greater than $\varepsilon_0$. When a dielectric is present between the plates of a capacitor, its capacitance is proportional to $\varepsilon$. The following table gives some sample values of the permittivities of a few substances.

<table>
<thead>
<tr>
<th>substance</th>
<th>$\varepsilon/\varepsilon_0$ at zero frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>vacuum</td>
<td>1</td>
</tr>
<tr>
<td>air</td>
<td>1.00054</td>
</tr>
<tr>
<td>water</td>
<td>80</td>
</tr>
<tr>
<td>barium titanate</td>
<td>1250</td>
</tr>
</tbody>
</table>

A capacitor with a very high capacitance is potentially a superior replacement for a battery, but until the 1990’s this was impractical because capacitors with high enough values couldn’t be made, even with dielectrics having the largest known permittivities. Such supercapacitors, some with values in the kilofarad range, are now available. Most of them do not use dielectric at all; the very high
b/ A stud finder is used to locate the wooden beams, or studs, that form the frame behind the wallboard. It is a capacitor whose capacitance changes when it is brought close to a substance with a particular permittivity. Although the wall is external to the capacitor, a change in capacitance is still observed, because the capacitor has “fringing fields” that extend outside the region between its plates.

Although figure a/2 shows the dipoles in the dielectric being completely aligned, this is not a situation commonly encountered in practice. In such a situation, the material would be as polarized as it could possibly be, and if the field was increased further, it would not respond. In reality, a capacitor, for example, would normally be operated with fields that produced quite a small amount of alignment, and it would be under these conditions that the linear relationship $D = \varepsilon E$ would actually be a good approximation. Before a material’s maximum polarization is reached, it may actually spark or burn up.

**self-check A**

Suppose a parallel-plate capacitor is built so that a slab of dielectric material can be slid in or out. (This is similar to the way the stud finder in figure b works.) We insert the dielectric, hook the capacitor up to a battery to charge it, and then use an ammeter and a voltmeter to observe what happens when the dielectric is withdrawn. Predict the changes observed on the meters, and correlate them with the expected change in capacitance. Discuss the energy transformations involved, and determine whether positive or negative work is done in removing the dielectric. ▷ Answer, p. 431

c/ The magnetic version of figure a. A magnetically permeable material is placed at the center of a solenoid.

16.3 Magnetic materials

16.3.1 Magnetic permeability

Atoms and molecules may have magnetic dipole moments as well as electric dipole moments. Just as an electric dipole contains bound charges, a magnetic dipole has bound currents, which come from the motion of the electrons as they orbit the nucleus, c/1. Such a substance, subjected to a magnetic field, tends to align itself, c/2,
so that a sheet of current circulates around the externally applied field. Figure c/3 is closely analogous to figure a/3; in the central gray area, the atomic currents cancel out, but the atoms at the outer surface form a sheet of bound current. However, whereas like charges repel and opposite charges attract, it works the other way around for currents: currents in the same direction attract, and currents in opposite directions repel. Therefore the bound currents in a material inserted inside a solenoid tend to reinforce the free currents, and the result is to strengthen the field. The total current is \( I = I_{\text{free}} + I_{\text{bound}} \), and we define an altered version of the magnetic field,

\[
\mathbf{H} = \frac{\mathbf{B}}{\mu},
\]

and rewrite Ampère’s law as

\[
\Gamma H = I_{\text{through, free}}.
\]

The constant \( \mu \) is the permeability, with a vacuum value of \( \mu_0 = 4\pi k/c^2 \). Here are the magnetic permeabilities of some substances:

<table>
<thead>
<tr>
<th>Substance</th>
<th>( \mu/\mu_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>vacuum</td>
<td>1</td>
</tr>
<tr>
<td>aluminum</td>
<td>1.00002</td>
</tr>
<tr>
<td>steel</td>
<td>700</td>
</tr>
<tr>
<td>transformer iron</td>
<td>4,000</td>
</tr>
<tr>
<td>mu-metal</td>
<td>20,000</td>
</tr>
</tbody>
</table>

An iron-core electromagnet

A solenoid has 1000 turns of wire wound along a cylindrical core with a length of 10 cm. If a current of 1.0 A is used, find the magnetic field inside the solenoid if the core is air, and if the core is made of iron with \( \mu/\mu_0 = 4,000 \).

Air has essentially the same permeability as vacuum, so using the result of example 4 on page 348, we find that the field is 0.013 T.

We now consider the case where the core is filled with iron. The original derivation in example 4 started from Ampère’s law, which we now rewrite as \( \Gamma H = I_{\text{through, free}} \). As argued previously, the only significant contributions to the circulation come from line segment AB. This segment lies inside the iron, where \( \mathbf{H} = \mathbf{B}/\mu \). The \( \mathbf{H} \) field is the same as in the air-core case, since the new form of Ampère’s law only relates \( \mathbf{H} \) to the current in the wires (the free current). This means that \( \mathbf{B} = \mu \mathbf{H} \) is greater by a factor of 4,000 than in the air-core case, or 52 T. This is an extremely intense field — so intense, in fact, that the iron’s magnetic polarization would probably become saturated before we could actually get the field that high.

The electromagnet of example 1 could also be used as an inductor, and its inductance would be proportional to the permittivity...
A transformer with a laminated iron core. The input and output coils are inside the paper wrapper. The iron core is the black part that passes through the coils at the center, and also wraps around them on the outside.

A transformer or inductor with a permeable core does have some disadvantages, however, in certain applications. The oscillating magnetic field induces an electric field, and because the core is typically a metal, these currents dissipate energy strongly as heat. This behaves like a fairly large resistance in series with the coil. Figure f shows a method for reducing this effect. The iron core of this transformer has been constructed out of laminated layers, which has the effect of blocking the conduction of the eddy currents.

Cables designed to carry audio signals are typically made with two adjacent conductors, such that the current flowing out through one conductor comes back through the other one. Computer cables are similar, but usually have several such pairs bundled inside the insulator. This paired arrangement is known as differential mode, and has the advantage of cutting down on the reception and transmission of interference. In terms of transmission, the magnetic field created by the outgoing current is almost exactly canceled by the field from the return current, so electromagnetic waves are only weakly induced. In reception, both conductors are bathed in the same electric and magnetic fields, so an emf that adds current on one side subtracts current from the other side, resulting in cancellation.

The opposite of differential mode is called common mode. In common mode, all conductors have currents flowing in the same direction. Even when a circuit is designed to operate in differential mode, it may not have exactly equal currents in the two conductors with \( I_1 + I_2 = 0 \), meaning that current is leaking off to ground at one end of the circuit or the other. Although paired cables are relatively immune to differential-mode interference, they do not have any automatic protection from common-mode interference.

Figure g shows a device for reducing common-mode interference called a ferrite bead, which surrounds the cable like a bead on a string. Ferrite is a magnetically permeable alloy. In this application, the ohmic properties of the ferrite actually turn out to be advantageous.

Let's consider common-mode transmission of interference. The bare cable has some DC resistance, but is also surrounded by a magnetic field, so it has inductance as well. This means that it behaves like a series L-R circuit, with an impedance that varies as \( R + i\omega L \), where both \( R \) and \( L \) are very small. When we add the ferrite bead, the inductance is increased by orders of magnitude, but so is the resistance. Neither \( R \) nor \( L \) is actually constant with respect to frequency, but both are much greater than for the bare
Suppose, for example, that a signal is being transmitted from a digital camera to a computer via a USB cable. The camera has an internal impedance that is on the order of 10 Ω, the computer’s input also has a ~ 10 Ω impedance, and in differential mode the ferrite bead has no effect, so the cable’s impedance has its low, designed value (probably also about 10 Ω, for good impedance matching). The signal is transmitted unattenuated from the camera to the computer, and there is almost no radiation from the cable.

But in reality there will be a certain amount of common-mode current as well. With respect to common mode, the ferrite bead has a large impedance, with the exact value depending on frequency, but typically on the order of 100 Ω for frequencies in the MHz range. We now have a series circuit consisting of three impedances: 10, 100, and 10 Ω. For a given emf applied by an external radio wave, the current induced in the circuit has been attenuated by an order of magnitude, relative to its value without the ferrite bead.

Why is the ferrite necessary at all? Why not just insert ordinary air-core inductors in the circuit? We could, for example, have two solenoidal coils, one in the outgoing line and one in the return line, interwound with one another with their windings oriented so that their differential-mode fields would cancel. There are two good reasons to prefer the ferrite bead design. One is that it allows a clip-on device like the one in the top panel of figure g, which can be added without breaking the circuit. The other is that our circuit will inevitably have some stray capacitance, and will therefore act like an LRC circuit, with a resonance at some frequency. At frequencies close to the resonant frequency, the circuit would absorb and transmit common-mode interference very strongly, which is exactly the opposite of the effect we were hoping to produce. The resonance peak could be made low and broad by adding resistance in series, but this extra resistance would attenuate the differential-mode signals as well as the common-mode ones. The ferrite’s resistance, however, is actually a purely magnetic effect, so it vanishes in differential mode.

Surprisingly, some materials have magnetic permeabilities less than \( \mu_0 \). This cannot be accounted for in the model above, and although there are semiclassical arguments that can explain it to some extent, it is fundamentally a quantum mechanical effect. Materials with \( \mu > \mu_0 \) are called paramagnetic, while those with \( \mu < \mu_0 \) are referred to as diamagnetic. Diamagnetism is generally a much weaker effect than paramagnetism, and is easily masked if there is any trace of contamination from a paramagnetic material. Diamagnetic materials have the interesting property that they are repelled...
Summary of auxiliary equations and definitions

\[ H = \frac{B}{\mu} \]
\[ D = \varepsilon E \]
\[ \mu_0 = \frac{4\pi k}{c^2} \]
\[ \varepsilon_0 = \frac{1}{4\pi k} \]

At a boundary between two substances with \( \mu_2 > \mu_1 \), the \( H \) field has a continuous component parallel to the surface, which implies a discontinuity in the parallel component of the magnetic field \( B \).

Variables that are continuous at a boundary

\( E_\parallel \quad D_\perp \)
\( H_\parallel \quad B_\perp \)

from regions of strong magnetic field, and it is therefore possible to levitate a diamagnetic object above a magnet, as in figure h.

A complete statement of Maxwell’s equations in the presence of electric and magnetic materials is as follows:

\[
\begin{align*}
\text{div} \, D &= \rho \\
\text{div} \, B &= 0 \\
\text{curl} \, E &= -\frac{\partial B}{\partial t} \\
\text{curl} \, H &= \frac{\partial D}{\partial t} + j.
\end{align*}
\]

Suppose we have a boundary between two substances. By constructing a Gaussian or Ampèrian surface that extends across the boundary, we can arrive at various constraints on how the fields must behave as move from one substance into the other, when there are no free currents or charges present, and the fields are static. An interesting example is the application of Faraday’s law, \( \Gamma_H = 0 \), to the case where one medium — let’s say it’s air — has a low permeability, while the other one has a very high one. We will violate Faraday’s law unless the component of the \( H \) field parallel to the boundary is a continuous function, \( H_\parallel,1 = H_\parallel,2 \). This means that if \( \mu/\mu_0 \) is very high, the component of \( B = \mu H \) parallel to the surface will have an abrupt discontinuity, being much stronger inside the high-permeability material. The result is that when a magnetic field enters a high-permeability material, it tends to twist abruptly to one side, and the pattern of the field tends to be channeled through the material like water through a funnel. In a transformer, a permeable core functions to channel more of the magnetic flux from the input coil to the output coil. Figure j shows another example, in which the effect is to shield the interior of the sphere from the externally imposed field. Special high-permeability alloys, with trade names like Mu-Metal, are sold for this purpose.

16.3.2 Ferromagnetism

The very last magnetic phenomenon we’ll discuss is probably the very first experience you ever had of magnetism. Ferromagnetism is a phenomenon in which a material tends to organize itself so that it has a nonvanishing magnetic field. It is exhibited strongly by iron and nickel, which explains the origin of the name.

Figure k/1 is a simple one-dimensional model of ferromagnetism. Each magnetic compass needle represents an atom. The compasses in the chain are stable when aligned with one another, because each
j. A hollow sphere with $\mu/\mu_0 = 10$, is immersed in a uniform, externally imposed magnetic field. The interior of the sphere is shielded from the field. The arrows map the magnetic field $\mathbf{B}$. (See homework problem nn, page nnn.)

k. A model of ferromagnetism.

one’s north end is attracted to its neighbor’s south end. The chain can be turned around, $k/2$, without disrupting its organization, and the compasses do not realign themselves with the Earth’s field, because their torques on one another are stronger than the Earth’s torques on them. The system has a memory. For example, if I want to remind myself that my friend’s address is 137 Coupling Ct., I can align the chain at an angle of 137 degrees. The model fails, however, as an explanation of real ferromagnetism, because in two or more dimensions, the most stable arrangement of a set of interacting
magnetic dipoles is something more like $k/3$, in which alternating rows point in opposite directions. In this two-dimensional pattern, every compass is aligned in the most stable way with all four of its neighbors. This shows that ferromagnetism, like diamagnetism, has no purely classical explanation; a full explanation requires quantum mechanics.

Because ferromagnetic substances “remember” the history of how they were prepared, they are commonly used to store information in computers. Figure 1 shows 16 bits from an ancient (ca. 1970) 4-kilobyte random-access memory, in which each doughnut-shaped iron “core” can be magnetized in one of two possible directions, so that it stores one bit of information. Today, RAM is made of transistors rather than magnetic cores, but a remnant of the old technology remains in the term “core dump,” meaning “memory dump,” as in “my girlfriend gave me a total core dump about her mom’s divorce.” Most computer hard drives today do store their information on rotating magnetic platters, but the platter technology may be obsoleted by flash memory in the near future.

The memory property of ferromagnets can be depicted on the type of graph shown in figure m, known as a hysteresis curve. The y axis is the magnetization of a sample of the material — a measure of the extent to which its atomic dipoles are aligned with one another. If the sample is initially unmagnetized, 1, and a field $H$ is externally applied, the magnetization increases, 2, but eventually becomes saturated, 3, so that higher fields do not result in any further magnetization, 4. The external field can then be reduced, 5, and even eliminated completely, but the material will retain its magnetization. It is a permanent magnet. To eliminate its magnetization completely, a substantial field must be applied in the opposite direction. If this reversed field is made stronger, then the substance will eventually become magnetized just as strongly in the opposite direction. Since the hysteresis curve is nonlinear, and is not a function (it has more than one value of $M$ for a particular value of $B$), a ferromagnetic material does not have a single, well-defined value of the permeability $\mu$; a value like 4,000 for transformer iron represents some kind of a rough average.

The fluxgate compass example 3

The fluxgate compass is a type of magnetic compass without moving parts, commonly used on ships and aircraft. An AC current is applied in a coil wound around a ferromagnetic core, driving the core repeatedly around a hysteresis loop. Because the hysteresis curve is highly nonlinear, the addition of an external field such as the Earth’s alters the core’s behavior. Suppose, for example, that the axis of the coil is aligned with the magnetic north-south. The core will reach saturation more quickly when the coil’s field is in the same direction as the Earth’s, but will not saturate as early in the next half-cycle, when the two fields are
in opposite directions. With the use of multiple coils, the com-
ponents of the Earth’s field can be measured along two or three
axes, permitting the compass’s orientation to be determined in
two or (for aircraft) three dimensions.

**Sharp magnet poles**  
*Example 4*  
Although a ferromagnetic material does not really have a single
value of the magnetic permeability, there is still a strong tendency
to have $B_\parallel \approx 0$ just outside the magnet’s surface, for the same
reasons as discussed above for high-permeability substances in
general. For example, if we have a cylindrical bar magnet about
the size and shape of your finger, magnetized lengthwise, then
the field near the ends is nearly perpendicular to the surfaces,
while the field near the sides, although it may be oriented nearly
parallel to the surface, is very weak, so that we still have $B_\parallel \approx
0$. This is in close analogy to the situation for the electric field
near the surface of a conductor in equilibrium, for which $E_\parallel = 0$.
This analogy is close enough so that we can recycle much of our
knowledge about electrostatics.

For example, we saw in example 2, p. 91, and problem 17, p. 108,
that charge tends to collect on the most highly curved portions
of a conductor, and therefore becomes especially dense near a
corner or knife-edge. This gives us a way of making especially
intense magnetic fields. Most people would imagine that a very
intense field could be made simply by using a very large and bulky
permanent magnet, but this doesn’t actually work very well, be-
cause magnetic dipole fields fall off as $1/r^3$, so that at a point near
the surface, nearly all the field is contributed by atoms near the
surface. Our analogy with electrostatics suggests that we should
instead construct a permanent magnet with a sharp edge.

Figure o shows the cross-sectional shapes of two magnet poles
used in the historic Stern-Gerlach experiment that discovered the
spin of the electron. The external magnetic field is represented
using field lines. The field lines enter and exit the surfaces per-
pendicularly, and they are particularly dense near the corner of
the upper pole, indicating a strong field. The spreading of the
field lines indicates that the field is strongly nonuniform, becom-
ing much weaker toward the bottom of the gap between the poles.
This strong nonuniformity was crucial for the experiment, in which
the magnets were used as part of a dipole spectrometer. See fig-
ure u on p. 132 for an electric version of such a spectrometer.

### 16.4 Electromagnetic waves in matter

In example 8, p. 300, we gave an explanation of why electromagnetic
waves traveling through matter are dispersive (sec. 6.4), i.e., their
speed depends on their frequency. The concept is that when an
electromagnetic wave enters a material such as a glass windowpane, charges inside the material oscillate in response to the wave. The charges have a resonant frequency (or, in real-world materials, several different resonant frequencies), so the amplitude and phase of their oscillation depends on the frequency of the driving wave. The oscillating charges in turn re-emit their own wave, which superposes on top of the original wave. The superposition may either lead or lag behind the original wave, so its crests arrive early or late. The effect is identical to what we would expect if the speed of the wave had some other value than $c$.

In that explanation, we focused on the phases and the trend of the effect with changing frequency, ignoring real constants. Fortunately, it turns out that the effect of all those ignored constants can be summarized in a simple way in terms of the bulk properties of the material. If we compare Maxwell’s equations in matter with their vacuum version, we see that the speed of an electromagnetic wave moving through a substance described by permittivity and permeability $\epsilon$ and $\mu$ is $1/\sqrt{\epsilon \mu}$.

For most substances, we observe that $\epsilon$ is highly frequency-dependent, and this is well explained by our earlier analysis in terms of the excitation of oscillating charges, where the behavior changes dramatically as the frequency passes through a resonant frequency.

The possibility of $\mu \neq \mu_0$ corresponds microscopically to a picture in which dipoles flip their orientation back and forth in response to the wave. For most substances we find that this isn’t a significant effect, and $\mu \approx \mu_0$.

---

**Color interference example 5**

The colorful image on the cover of this book was created by taking two polarizing films, as in example 1, p. 153, and minilab 6, p. 170, and placing between them a calcium sulfate crystal. At the atomic level, the crystal is a lattice of atoms, and the lattice is asymmetric, so that it has two distinguishable axes. Let’s call these axes $x$ and $y$, and describe the polarization of electromagnetic waves by the direction of the electric fields. If a wave has a polarization in the $x$ direction, the permittivity has some value $\epsilon_x$, but a wave with its polarization in the $y$ direction experiences some other $\epsilon_y$. Both of these depend on frequency.

Let’s analyze the simplest possible example that elucidates the physics. Suppose that we orient the first polarizing film at a $45^\circ$ angle with respect to the axes, so that the light entering the crystal is (ignoring unitful constant factors)

$$E_1 = \hat{x} + \hat{y}.$$  

As the wave emerges from the crystal, the difference in velocity for the two components will put them out of phase. Let’s say that the result is to reverse the phase of the $y$ component compared
to the \( x \),

\[ \mathbf{E}_2 = \hat{x} - \hat{y}. \]

This is a 90-degree rotation compared to \( \mathbf{E}_1 \). Now suppose that the second polarizing film is oriented in the same direction as the first film. Then although \( \mathbf{E}_1 \) was in exactly the right direction to pass through, the direction of \( \mathbf{E}_2 \) is precisely wrong. The light is completely blocked.

But this phase relationship depends not just on the thickness of the crystal and the values of \( \epsilon_x \) and \( \epsilon_y \), but also on frequency. For some other frequency of light, the phase of the \( x \) and \( y \) polarizations in \( \mathbf{E}_2 \) could end up the same as in \( \mathbf{E}_1 \), in which case the light would be entirely transmitted through the second filter. Of course all of the intermediate possibilities occur as well. For this reason, some colors are more strongly transmitted through this setup and some more strongly absorbed.
Problems

Key
✓ A computerized answer check is available online.
★ A difficult problem.
Relativity
Chapter 17

☆Relativity (optional stand-alone chapter)

This optional chapter is a stand-alone presentation of special relativity. It can be read before, during, or after the rest of the book.

17.1 Time is not absolute

When Einstein first began to develop the theory of relativity, around 1905, the only real-world observations he could draw on were ambiguous and indirect. Today, the evidence is part of everyday life. For example, every time you use a GPS receiver, you’re using Einstein’s theory of relativity. Somewhere between 1905 and today, technology became good enough to allow conceptually simple experiments that students in the early 20th century could only discuss in terms like “Imagine that we could…”

A good jumping-on point is 1971. In that year, J.C. Hafele and R.E. Keating brought atomic clocks aboard commercial airliners, and went around the world, once from east to west and once from west to east. Hafele and Keating observed that there was a discrepancy between the times measured by the traveling clocks and the times measured by similar clocks that stayed home at the U.S. Naval Observatory in Washington. The east-going clock lost time, ending up off by $-59 \pm 10$ nanoseconds, while the west-going one gained $273 \pm 7$ ns.

17.1.1 The correspondence principle

This establishes that time doesn’t work the way Newton believed it did when he wrote that “Absolute, true, and mathematical time, of itself, and from its own nature flows equably without regard to anything external…” We are used to thinking of time as absolute and universal, so it is disturbing to find that it can flow at a different rate for observers in different frames of reference. Nevertheless, the effects that Hafele and Keating observed were small. This makes sense: Newton’s laws have already been thoroughly tested by experiments under a wide variety of conditions, so a new theory like relativity must agree with Newton’s to a good approximation, within the Newtonian theory’s realm of applicability. This requirement of backward-compatibility is known as the correspondence principle.
17.1.2 Causality

It’s also reassuring that the effects on time were small compared to the three-day lengths of the plane trips. There was therefore no opportunity for paradoxical scenarios such as one in which the east-going experimenter arrived back in Washington before he left and then convinced himself not to take the trip. A theory that maintains this kind of orderly relationship between cause and effect is said to satisfy causality.

Causality is like a water-hungry front-yard lawn in Los Angeles: we know we want it, but it’s not easy to explain why. Even in plain old Newtonian physics, there is no clear distinction between past and future. In figure c, number 18 throws the football to number 25, and the ball obeys Newton’s laws of motion. If we took a video of the pass and played it backward, we would see the ball flying from 25 to 18, and Newton’s laws would still be satisfied. Nevertheless, we have a strong psychological impression that there is a forward arrow of time. I can remember what the stock market did last year, but I can’t remember what it will do next year. Joan of Arc’s military victories against England caused the English to burn her at the stake; it’s hard to accept that Newton’s laws provide an equally good description of a process in which her execution in 1431 caused her to win a battle in 1429. There is no consensus at this point among physicists on the origin and significance of time’s arrow, and for our present purposes we don’t need to solve this mystery. Instead, we merely note the empirical fact that, regardless of what causality really means and where it really comes from, its behavior is consistent. Specifically, experiments show that if an observer in a certain frame of reference observes that event A causes event B, then observers in other frames agree that A causes B, not the other way around. This is merely a generalization about a large body of experimental results, not a logically necessary assumption. If Keating had gone around the world and arrived back in Washington before he left, it would have disproved this statement about causality.

17.1.3 Time distortion arising from motion and gravity

Hafele and Keating were testing specific quantitative predictions of relativity, and they verified them to within their experiment’s error bars. Let’s work backward instead, and inspect the empirical results for clues as to how time works.

The two traveling clocks experienced effects in opposite directions, and this suggests that the rate at which time flows depends on the motion of the observer. The east-going clock was moving in the same direction as the earth’s rotation, so its velocity relative to the earth’s center was greater than that of the clock that remained in Washington, while the west-going clock’s velocity was correspondingly reduced. The fact that the east-going clock fell behind, and the west-going one got ahead, shows that the effect of motion is to
make time go more slowly. This effect of motion on time was predicted by Einstein in his original 1905 paper on relativity, written when he was 26.

If this had been the only effect in the Hafele-Keating experiment, then we would have expected to see effects on the two flying clocks that were equal in size. Making up some simple numbers to keep the arithmetic transparent, suppose that the earth rotates from west to east at 1000 km/hr, and that the planes fly at 300 km/hr. Then the speed of the clock on the ground is 1000 km/hr, the speed of the clock on the east-going plane is 1300 km/hr, and that of the west-going clock 700 km/hr. Since the speeds of 700, 1000, and 1300 km/hr have equal spacing on either side of 1000, we would expect the discrepancies of the moving clocks relative to the one in the lab to be equal in size but opposite in sign.

In fact, the two effects are unequal in size: −59 ns and 273 ns. This implies that there is a second effect involved, simply due to the planes’ being up in the air. This was verified more directly in a 1978 experiment by Iijima and Fujiwara, figure e, in which identical atomic clocks were kept at rest at the top and bottom of a
The correspondence principle requires that the relativistic distortion of time become small for small velocities. This experiment, unlike the Hafele-Keating one, isolates one effect on time, the gravitational one: time’s rate of flow increases with height in a gravitational field. Einstein didn’t figure out how to incorporate gravity into relativity until 1915, after much frustration and many false starts. The simpler version of the theory without gravity is known as special relativity, the full version as general relativity. We’ll restrict ourselves to special relativity, and that means that what we want to focus on right now is the distortion of time due to motion, not gravity.

We can now see in more detail how to apply the correspondence principle. The behavior of the three clocks in the Hafele-Keating experiment shows that the amount of time distortion increases as the speed of the clock’s motion increases. Newton lived in an era when the fastest mode of transportation was a galloping horse, and the best pendulum clocks would accumulate errors of perhaps a minute over the course of several days. A horse is much slower than a jet plane, so the distortion of time would have had a relative size of only $\sim 10^{-15}$ — much smaller than the clocks were capable of detecting. At the speed of a passenger jet, the effect is about $10^{-12}$, and state-of-the-art atomic clocks in 1971 were capable of measuring that. A GPS satellite travels much faster than a jet airplane, and the effect on the satellite turns out to be $\sim 10^{-10}$. The general idea here is that all physical laws are approximations, and approximations aren’t simply right or wrong in different situations. Approximations are better or worse in different situations, and the question is whether a particular approximation is good enough in a given situation to serve a particular purpose. The faster the motion, the worse the Newtonian approximation of absolute time. Whether the approximation is good enough depends on what you’re trying to accomplish. The correspondence principle says that the approximation must have been good enough to explain all the experiments done in the centuries before Einstein came up with relativity.

By the way, don’t get an inflated idea of the importance of the Hafele-Keating experiment. Special relativity had already been confirmed by a vast and varied body of experiments decades before 1971. The only reason I’m giving such a prominent role to this experiment, which was actually more important as a test of general relativity, is that it is conceptually very direct.

### 17.2 Distortion of space and time

#### 17.2.1 The Lorentz transformation

Relativity says that when two observers are in different frames of reference, each observer considers the other one’s perception of time to be distorted. We’ll also see that something similar happens to their observations of distances, so both space and time are distorted.
What exactly is this distortion? How do we even conceptualize it?

The idea isn’t really as radical as it might seem at first. We can visualize the structure of space and time using a graph with position and time on its axes. These graphs are familiar by now, but we’re going to look at them in a slightly different way. Before, we used them to describe the motion of objects. The grid underlying the graph was merely the stage on which the actors played their parts. Now the background comes to the foreground: it’s time and space themselves that we’re studying. We don’t necessarily need to have a line or a curve drawn on top of the grid to represent a particular object. We may, for example, just want to talk about events, depicted as points on the graph as in figure g. A distortion of the Cartesian grid underlying the graph can arise for perfectly ordinary reasons that Isaac Newton would have readily accepted. For example, we can simply change the units used to measure time and position, as in figure h.

We’re going to have quite a few examples of this type, so I’ll adopt the convention shown in figure i for depicting them. Figure i summarizes the relationship between figures g and h in a more compact form. The gray rectangle represents the original coordinate grid of figure g, while the grid of black lines represents the new version from figure h. Omitting the grid from the gray rectangle makes the diagram easier to decode visually.

Our goal of unraveling the mysteries of special relativity amounts to nothing more than finding out how to draw a diagram like i in the case where the two different sets of coordinates represent measurements of time and space made by two different observers, each in motion relative to the other. Galileo and Newton thought they knew the answer to this question, but their answer turned out to be only approximately right. To avoid repeating the same mistakes, we need to clearly spell out what we think are the basic properties of time and space that will be a reliable foundation for our reasoning. I want to emphasize that there is no purely logical way of deciding on this list of properties. The ones I’ll list are simply a summary of the patterns observed in the results from a large body of experiments. Furthermore, some of them are only approximate. For example, property 1 below is only a good approximation when the gravitational field is weak, so it is a property that applies to special relativity, not to general relativity.

Experiments show that:

1. No point in time or space has properties that make it different from any other point.

2. Likewise, all directions in space have the same properties.

3. Motion is relative, i.e., all inertial frames of reference are
equally valid.

4. Causality holds, in the sense described on page 376.

5. Time depends on the state of motion of the observer.

Most of these are not very subversive. Properties 1 and 2 date back to the time when Galileo and Newton started applying the same universal laws of motion to the solar system and to the earth; this contradicted Aristotle, who believed that, for example, a rock would naturally want to move in a certain special direction (down) in order to reach a certain special location (the earth’s surface). Property 3 is the reason that Einstein called his theory “relativity,” but Galileo and Newton believed exactly the same thing to be true, as dramatized by Galileo’s run-in with the Church over the question of whether the earth could really be in motion around the sun. Property 4 would probably surprise most people only because it asserts in such a weak and specialized way something that they feel deeply must be true. The only really strange item on the list is 5, but the Hafele-Keating experiment forces it upon us.

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

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

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

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

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

The examples in figures k and l show that the transformation we’re looking for must be linear, meaning that it must transform lines into lines, and furthermore that it has to take parallel lines to parallel lines. Einstein wrote in his 1905 paper that “…on account of the property of homogeneity [property 1] which we ascribe to time and space, the [transformation] must be linear.” Applying this to our diagrams, the original gray rectangle, which is a special type of parallelogram containing right angles, must be transformed into another parallelogram. There are three types of transformations, figure m, that have this property. Case I is the Galilean transformation of figure j on page 380, which we’ve already ruled out.

---

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

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

Three types of transformations that preserve parallelism. Their distinguishing feature is what they do to simultaneity, as shown by what happens to the left edge of the original rectangle. In I, the left edge remains vertical, so simultaneous events remain simultaneous. In II, the left edge turns counterclockwise. In III, it turns clockwise.

Case II can also be discarded. Here every point on the grid rotates counterclockwise. What physical parameter would determine the amount of rotation? The only thing that could be relevant would be $v$, the relative velocity of the motion of the two frames of reference with respect to one another. But if the angle of rotation was proportional to $v$, then for large enough velocities the grid would have left and right reversed, and this would violate property 4, causality: one observer would say that event A caused a later event B, but another observer would say that B came first and caused A.

The only remaining possibility is case III, which I’ve redrawn in figure n with a couple of changes. This is the one that Einstein predicted in 1905. The transformation is known as the Lorentz transformation, after Hendrik Lorentz (1853-1928), who partially anticipated Einstein’s work, without arriving at the correct interpretation. The distortion is a kind of smooshing and stretching, as suggested by the hands. Also, we’ve already seen in figures g-i on page 379 that we’re free to stretch or compress everything as much as we like in the horizontal and vertical directions, because this simply corresponds to choosing different units of measurement for time and distance. In figure n I’ve chosen units that give the whole drawing a convenient symmetry about a 45-degree diagonal line. Ordinarily it wouldn’t make sense to talk about a 45-degree angle on a graph whose axes had different units. But in relativity, the symmetric appearance of the transformation tells us that space and time ought to be treated on the same footing, and measured in the same units.

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

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

\[ p \]

Proof that Lorentz transformations don’t change area: We first subject a square to a transformation with velocity \( v \), and this increases its area by a factor \( R(v) \), which we want to prove equals 1. We chop the resulting parallelogram up into little squares and finally apply a \(-v\) transformation; this changes each little square’s area by a factor \( R(-v) \), so the whole figure’s area is also scaled by \( R(-v) \). The final result is to restore the square to its original shape and area, so \( R(v)R(-v) = 1 \). But \( R(v) = R(-v) \) by property 2 of spacetime on page 379, which states that all directions in space have the same properties, so \( R(v) = 1 \).

17.2.2 The \( \gamma \) factor

With a little algebra and geometry (homework problem 3, page 419), one can use the equal-area property to show that the factor \( \gamma \) (Greek letter gamma) defined in figure q is given by the equation

\[
\gamma = \frac{1}{\sqrt{1 - v^2}}.
\]
If you’ve had good training in physics, the first thing you probably think when you look at this equation is that it must be nonsense, because its units don’t make sense. How can we take something with units of velocity squared, and subtract it from a unitless 1? But remember that this is expressed in our special relativistic units, in which the same units are used for distance and time. We refer to these as natural units. In this system, velocities are always unitless. This sort of thing happens frequently in physics. For instance, before James Joule discovered conservation of energy, nobody knew that heat and mechanical energy were different forms of the same thing, so instead of measuring them both in units of joules as we would do now, they measured heat in one unit (such as calories) and mechanical energy in another (such as foot-pounds). In ordinary metric units, we just need an extra conversion factor \( c \), and the equation becomes

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

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

**Time dilation**

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

As proved in figures r and s, lengths are also distorted:

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

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

**self-check A**

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

Figure t shows the behavior of \( \gamma \) as a function of \( v \).

---

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

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

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

---

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

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

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

---

**An interstellar road trip**  
*example 3*

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

Betty experiences time dilation. At this speed, her \( \gamma \) is 2.0, so that
Muons accelerated to nearly \( c \) undergo radioactive decay much more slowly than they would according to an observer at rest with respect to the muons. The first two data-points (unfilled circles) were subject to large systematic errors.

The correspondence principle requires that \( \gamma \) be close to 1 for the velocities much less than \( c \) encountered in everyday life. In natural units, \( \gamma = (1 - v^2)^{-1/2} \). For small values of \( \epsilon \), the approximation \( (1 + \epsilon)^p \approx 1 + p\epsilon \) holds (see p. 434). Applying this approximation, we find \( \gamma \approx 1 + v^2/2 \).

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

In problem 16 on p. 421 we rewrite this in SI units.

Figure t on p. 385 shows that the approximation is not valid for large values of \( v/c \). In fact, \( \gamma \) blows up to infinity as \( v \) gets closer and closer to \( c \).

The time dilation effect in the Hafele-Keating experiment was very small. If we want to see a large time dilation effect, we can’t do it with something the size of the atomic clocks they used; the kinetic energy would be greater than the total megatonnage of all the world’s nuclear arsenals. We can, however, accelerate subatomic particles to speeds at which \( \gamma \) is large. For experimental particle physicists, relativity is something you do all day before heading home and stopping off at the store for milk. An early, low-precision experiment of this kind was performed by Rossi and Hall in 1941, using naturally occurring cosmic rays. Figure w shows a 1974 experiment\(^2\) of a similar type which verified the time dilation predicted by relativity to a precision of about one part per thousand.

Particles called muons (named after the Greek letter \( \mu \), “myoo”) were produced by an accelerator at CERN, near Geneva. A muon is essentially a heavier version of the electron. Muons undergo radioactive decay, lasting an average of only \( 2.197 \mu s \) before they

\(^2\)Bailey et al., Nucl. Phys. B150(1979) 1
evaporate into an electron and two neutrinos. The 1974 experiment was actually built in order to measure the magnetic properties of muons, but it produced a high-precision test of time dilation as a byproduct. Because muons have the same electric charge as electrons, they can be trapped using magnetic fields. Muons were injected into the ring shown in figure w, circling around it until they underwent radioactive decay. At the speed at which these muons were traveling, they had \( \gamma = 29.33 \), so on the average they lasted 29.33 times longer than the normal lifetime. In other words, they were like tiny alarm clocks that self-destructed at a randomly selected time. Figure v shows the number of radioactive decays counted, as a function of the time elapsed after a given stream of muons was injected into the storage ring. The two dashed lines show the rates of decay predicted with and without relativity. The relativistic line is the one that agrees with experiment.
An example of length contraction

Figure x shows an artist's rendering of the length contraction for the collision of two gold nuclei at relativistic speeds in the RHIC accelerator in Long Island, New York. The gold nuclei would appear nearly spherical (or just slightly lengthened like an American football) in frames moving along with them, but in the laboratory's frame, they both appear drastically foreshortened as they approach the point of collision. The later pictures show the nuclei merging to form a hot soup, observed at RHIC in 2010, in which the quarks are no longer confined inside the protons and neutrons.

Example 7: In the garage's frame of reference, the bus is moving, and can fit in the garage due to its length contraction. In the bus's frame of reference, the garage is moving, and can’t hold the bus due to its length contraction.

The garage paradox

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

The paradox is resolved when we recognize that the concept of
fitting the bus in the garage “all at once” contains a hidden assumption, the assumption that it makes sense to ask whether the front and back of the bus can simultaneously be in the garage. Observers in different frames of reference moving at high relative speeds do not necessarily agree on whether things happen simultaneously. As shown in figure y, the person in the garage’s frame can shut the door at an instant B he perceives to be simultaneous with the front bumper’s arrival A at the back wall of the garage, but the driver would not agree about the simultaneity of these two events, and would perceive the door as having shut long after she plowed through the back wall.

17.2.3 The universal speed c

Let’s think a little more about the role of the 45-degree diagonal in the Lorentz transformation. Slopes on these graphs are interpreted as velocities. This line has a slope of 1 in relativistic units, but that slope corresponds to c in ordinary metric units. We already know that the relativistic distance unit must be extremely large compared to the relativistic time unit, so c must be extremely large. Now note what happens when we perform a Lorentz transformation: this particular line gets stretched, but the new version of the line lies right on top of the old one, and its slope stays the same. In other words, if one observer says that something has a velocity equal to c, every other observer will agree on that velocity as well. (The same thing happens with −c.)

Velocities don’t simply add and subtract.

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

A universal speed limit

For example, suppose Janet takes a trip in a spaceship, and accelerates until she is moving at 0.6 c relative to the earth. She then launches a space probe in the forward direction at a speed relative to her ship of 0.6c. We might think that the probe was then moving at a velocity of 1.2c, but in fact the answer is still less than c (problem 1, page 418). This is an example of a more general fact about relativity, which is that c represents a universal speed limit. This is required by causality, as shown in figure z.
The Michelson-Morley experiment, shown in photographs, and drawings from the original 1887 paper. 1. A simplified drawing of the apparatus. A beam of light from the source, s, is partially reflected and partially transmitted by the half-silvered mirror $h_1$. The two half-intensity parts of the beam are reflected by the mirrors at a and b, reunited, and observed in the telescope, t. If the earth's surface was supposed to be moving through the ether, then the times taken by the two light waves to pass through the moving ether would be unequal, and the resulting time lag would be detectable by observing the interference between the waves when they were reunited. 2. In the real apparatus, the light beams were reflected multiple times. The effective length of each arm was increased to 11 meters, which greatly improved its sensitivity to the small expected difference in the speed of light. 3. In an earlier version of the experiment, they had run into problems with its “extreme sensitiveness to vibration,” which was “so great that it was impossible to see the interference fringes except at brief intervals … even at two o'clock in the morning.” They therefore mounted the whole thing on a massive stone floating in a pool of mercury, which also made it possible to rotate it easily. 4. A photo of the apparatus.

Light travels at $c$.

Now consider a beam of light. We're used to talking casually about the “speed of light,” but what does that really mean? Motion is relative, so normally if we want to talk about a velocity, we have to specify what it’s measured relative to. A sound wave has a certain speed relative to the air, and a water wave has its own speed relative
to the water. If we want to measure the speed of an ocean wave, for example, we should make sure to measure it in a frame of reference at rest relative to the water. But light isn’t a vibration of a physical medium; it can propagate through the near-perfect vacuum of outer space, as when rays of sunlight travel to earth. This seems like a paradox: light is supposed to have a specific speed, but there is no way to decide what frame of reference to measure it in. The way out of the paradox is that light must travel at a velocity equal to $c$. Since all observers agree on a velocity of $c$, regardless of their frame of reference, everything is consistent.

**The Michelson-Morley experiment**

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

---

*The ring laser gyroscope*  
Example 8

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

---

*No frequency-dependence*  
Example 9

Relativity has only one universal speed, so it requires that all light waves travel at the same speed, regardless of their frequency and wavelength. Presently the best experimental tests of the invariance of the speed of light with respect to wavelength come from astronomical observations of gamma-ray bursts, which are sudden outpourings of high-frequency light, believed to originate from a supernova explosion in another galaxy. One such obser-
Discussion question B

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

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

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

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

E Figure x from page 387, depicting the collision of two nuclei at the RHIC accelerator, is reproduced below. What would the shapes of the two nuclei look like to a microscopic observer riding on the left-hand nucleus? To an observer riding on the right-hand one? Can they agree on what is happening? If not, why not — after all, shouldn’t they see the same thing if they both compare the two nuclei side-by-side at the same instant in time?

Discussion question E: colliding nuclei show relativistic length contraction.

F If you stick a piece of foam rubber out the window of your car while driving down the freeway, the wind may compress it a little. Does it make sense to interpret the relativistic length contraction as a type of strain that pushes an object’s atoms together like this? How does this relate to discussion question E?

G The machine-gunner in the figure sends out a spray of bullets.

\(^3\)http://arxiv.org/abs/0908.1832
Suppose that the bullets are being shot into outer space, and that the distances traveled are trillions of miles (so that the human figure in the diagram is not to scale). After a long time, the bullets reach the points shown with dots which are all equally far from the gun. Their arrivals at those points are events A through E, which happen at different times. Sketch these events on a position-time graph. The chain of impacts extends across space at a speed greater than $c$. Does this violate special relativity?

Discussion question G.
17.3 No action at a distance

17.3.1 The Newtonian picture

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

17.3.2 Time delays in forces exerted at a distance

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

Since forces can’t be transmitted instantaneously, it becomes natural to imagine force-effects spreading outward from their source...
like ripples on a pond, and we then have no choice but to impute some physical reality to these ripples. We call them fields, and they have their own independent existence. Gravity is transmitted through a field called the gravitational field. Besides gravity, there are other fundamental fields of force such as electricity and magnetism (). Ripples of the electric and magnetic fields turn out to be light waves. This tells us that the speed at which electric and magnetic field ripples spread must be \( c \), and by an argument similar to the one in subsection 17.2.3 the same must hold for any other fundamental field, including the gravitational field.

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

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

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

The smoking-gun argument for this strange notion of traveling force ripples comes from the fact that they carry energy. In figure ae/1, Alice and Betty hold balls A and B at some distance from one another. These balls make a force on each other; it doesn’t really matter for the sake of our argument whether this force is gravitational, electrical, or magnetic. Let’s say it’s electrical, i.e., that the balls have the kind of electrical charge that sometimes causes your socks to cling together when they come out of the clothes dryer. We’ll say the force is repulsive, although again it doesn’t really matter.

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

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

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

Discussion questions

A  Amy and Bill are flying on spaceships in opposite directions at such high velocities that the relativistic effect on time’s rate of flow is easily noticeable. Motion is relative, so Amy considers herself to be at rest and Bill to be in motion. She says that time is flowing normally for her, but Bill is slow. But Bill can say exactly the same thing. How can they both think the other is slow? Can they settle the disagreement by getting on the radio and seeing whose voice is normal and whose sounds slowed down and Darth-Vadery?

B  The figure shows a famous thought experiment devised by Einstein. A train is moving at constant velocity to the right when bolts of lightning strike the ground near its front and back. Alice, standing on the dirt at the midpoint of the flashes, observes that the light from the two flashes arrives simultaneously, so she says the two strikes must have occurred simultaneously. Bob, meanwhile, is sitting aboard the train, at its middle. He passes by Alice at the moment when Alice later figures out that the flashes happened. Later, he receives flash 2, and then flash 1. He infers that since both flashes traveled half the length of the train, flash 2 must have occurred first. How can this be reconciled with Alice’s belief that the flashes were simultaneous? Explain using a graph.

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

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

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

17.4 The light cone

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

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

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

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

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

Measurement in spacetime
We define \( AB \cong CD \) if:

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

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

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

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

Case 1 has to be the way it is because space is part of spacetime. In special relativity, this space is Euclidean, so the definition of congruence has to agree with the Euclidean definition, in the case where it is possible to apply the Euclidean definition. The spacelike relation between the points is both necessary and sufficient to make this possible. If points \( A \) and \( B \) are spacelike in relation to one another, then a frame of reference exists in which they are simultaneous, so we can use a ruler that is at rest in that frame to measure their distance. If they are lightlike or timelike, then no such frame of reference exists. For example, there is no frame of reference in which Charles VII’s restoration to the throne is simultaneous with Joan of Arc’s execution, so we can’t arrange for both of these events to touch the same ruler at the same time.
The definition in case 2 is the only sensible way to proceed if we are to respect the symmetric treatment of time and space in relativity. The timelike relation between the events is necessary and sufficient to make it possible for a clock to move from one to the other. It makes a difference that the clocks move inertially, because the twins in example 3 on p. 385 disagree on the clock time between the traveling twin’s departure and return.

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

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

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

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

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

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

### Pioneer 10 example 10

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

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

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

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

\[
\Delta t (\text{s}) \quad \Delta x \quad \Delta y \quad \Delta z
\]

\[
3.7869120000 \times 10^8 \text{ s} \quad 0.2121 \times 10^4 \text{ s} \quad 1.626 \times 10^4 \text{ s} \quad 0.084 \times 10^4 \text{ s}
\]

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

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

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