meters per second, so converting to mass units, we have

$$\begin{aligned} m &= \frac{E}{c^2} \\ &= \frac{0.5 \times 10^6 \text{ J}}{\left(3 \times 10^8 \text{ m/s}\right)^2} \\ &= 6 \times 10^{-12} \text{ kilograms.} \end{aligned}$$

The change in mass is too small to measure with any practical technique. This is because the square of the speed of light is such a large number.

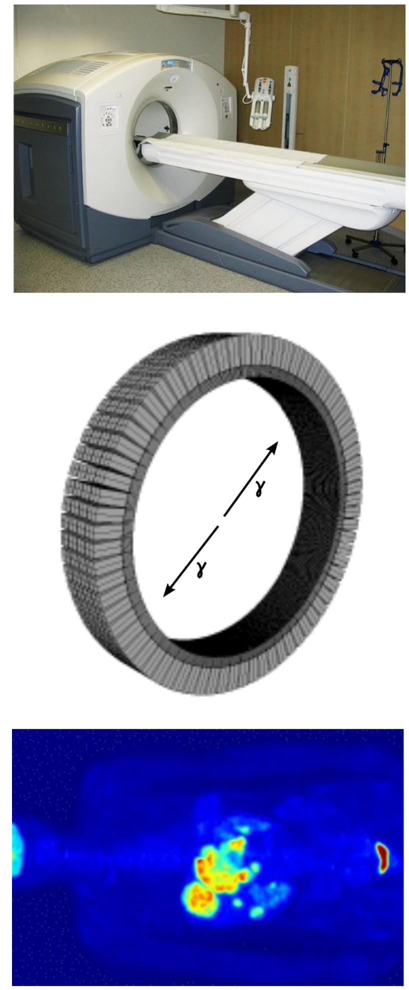
Electron-positron annihilation *example 11*

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

$$\begin{aligned} E &= mc^2 \\ &= (2 \text{ kg}) \left(3.0 \times 10^8 \text{ m/s}\right)^2 \\ &= 2 \times 10^{17} \text{ J,} \end{aligned}$$

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

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



ar / Top: A PET scanner. Middle: Each positron annihilates with an electron, producing two gamma-rays that fly off back-to-back. When two gamma rays are observed simultaneously in the ring of detectors, they are assumed to come from the same annihilation event, and the point at which they were emitted must lie on the line connecting the two detectors. Bottom: A scan of a person's torso. The body has concentrated the radioactive tracer around the stomach, indicating an abnormal medical condition.

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

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

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

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

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

Now that we have some idea what to expect, what is the actual equation for the mass-energy? As proved in my book *Simple Nature*, it is

$$\mathcal{E} = m\gamma c^2.$$

self-check E

Verify that this equation has the two properties we wanted. ▷ Answer, p. 1044

KE compared to mc^2 at low speeds *example 12*

▷ An object is moving at ordinary nonrelativistic speeds. Compare its kinetic energy to the energy mc^2 it has purely because of its mass.

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

Another way of seeing this is that at low speeds, γ is only a tiny bit greater than 1, so $\mathcal{E}$ is only a tiny bit greater than mc^2.

The correspondence principle for mass-energy *example 13*

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

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

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

In example 4 on page 683, we found $\gamma \approx 1 + v^2/2c^2$, so

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

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

26.6 ⋆ Proofs

In section 26.5 I gave physical arguments to the effect that relativistic momentum should be greater than mv and that an energy E should be equivalent relativistically to some amount of mass m. In this section I'll prove that the relativistic equations are as claimed: $p = m\gamma v$ and $E = mc^2$. The structure of the proofs is essentially the same as in two famous 1905 papers by Einstein, "On the electrodynamics of moving bodies" and "Does the inertia of a body depend upon its energy content?" If you're interested in reading these arguments as Einstein originally wrote them, you can find English translations at `www.fourmilab.ch`. We start off by proving two preliminary results relating to Doppler shifts.

Transformation of the fields in a light wave

On p. 716 I showed that when a light wave is observed in two different frames in different states of motion parallel to the wave's direction of motion, the frequency is observed to be Doppler-shifted by a factor $D(v) = \sqrt{(1-v)/(1+v)}$, where $c = 1$ and v is the relative velocity of the two frames. But a change in frequency is not the only change we expect. We also expect the *intensity* of the wave to change, since a combination of electric and magnetic fields observed in one frame of reference becomes some other set of fields in a different frame (p. 689). There are equations that express this transformation from $\mathbf{E}$ and $\mathbf{B}$ to $\mathbf{E}'$ and $\mathbf{B}'$, but they're a little complicated, so instead we'll just determine what happens in the special case of an electromagnetic wave.

Since the transformation of $\mathbf{E}$ and $\mathbf{B}$ to $\mathbf{E}'$ and $\mathbf{B}'$ is a universal thing, we're free to imagine that the wave was created in any way we wish. Suppose that it was created by a uniform sheet of charge in the x-y plane, oscillating in the y direction with amplitude A and frequency f. This will clearly produce electromagnetic waves propagating in the $+z$ and $-z$ directions, and by an argument similar to that of problem 7 on p. 662, we know that these waves' intensity will not fall off at all with distance from the sheet. Since magnetic fields are produced by currents, and the currents produced by the motion of the sheet are proportional to Af, the amplitude of the magnetic field in the wave is proportional to Af. The oscillating magnetic field induces an electric field, and since electromagnetic waves always have $E = Bc$, the oscillating part of the electric field is also proportional of Af.

An observer moving away from the sheet sees a sheet that is both oscillating more slowly (f is Doppler-shifted to fD) and receding. But the recession has no effect, because the fields don't fall off with distance. Also, A stays the same, because the Lorentz transformation has no effect on lengths perpendicular to the relative motion of the two frames. Since the fields are proportional to Af, the fields seen by the receding observer are attenuated by a factor of D.

Transformation of the volume of a light wave

Since the fields in an electromagnetic wave are changed by a factor of D when we change frames, we might expect that the wave's energy would change by a factor of D^2. But the square of the field only gives the energy per unit volume, and the volume changes as well. The following argument shows that the volume increases by a factor of $1/D$.

If an electromagnetic wave-train has duration Δt, we already know that its duration changes by a factor of $1/D$ when we change to a different frame of reference. But the speed of light is the same for all observers, so if the length of the wave-train is Δz, all observers must agree on the value of $\Delta z/\Delta t$, and $1/D$ must also be the factor by which Δz scales up.[10] Since the Lorentz transformation doesn't change Δx or Δy, the volume of the wave-train is also increased by a factor of $1/D$.

Transformation of the energy of a light wave

Combining the two preceding results, we find that when we change frames of reference, the energy *density* (per unit volume) of a light wave changes by a factor of D^2, but the volume changes by $1/D$, so the result is that the wave's energy changes by a factor of D. In Einstein's words, "It is remarkable that the energy and the frequency of a [wave-train] vary with the state of motion of the observer in accordance with the same law," i.e., that both scale by the same factor D. Einstein had a reason to be especially interested in this fact. In the same "miracle year" of 1905, he also published a paper in which he hypothesized that light had both particle and wave properties, with the energy E of a light-particle related to the frequency f of the corresponding light-wave by $E = hf$, where h was a constant.

[10]At first glance, one might think that this length-scaling factor would simply be γ, and that the volume would be reduced rather than increased. But γ is only the scale-down factor for the length of a thing compared to that thing's length in a frame where it is at rest. Light waves don't have a frame in which they're at rest. One can also see this from the geometry of figure x on p. 717. The diagram is completely symmetric with respect to its treatment of time and space, so if we flip it across its diagonal, interchanging the roles of z and t, we obtain the same result for the wave-train's spatial extent Δz.

(More about this in ch. 34.) If E and f had not both scaled by the same factor, then the relation $E = hf$ could not have held in all frames of reference.

$E = mc^2$

Suppose that a material object O, initially at rest, emits two light rays, each with energy E, in the $+z$ and $-z$ directions. O could be a lantern with windows on opposite sides, or it could be an electron and an antielectron annihilating each other to produce a pair of gamma rays. In this frame, O loses energy $2E$ and the light rays gain $2E$, so energy is conserved.

We now switch to a new frame of reference moving at a certain velocity v in the z direction relative to the original frame. We assume that O's energy is different in this frame, but that the change in its energy amounts to multiplication by some unitless factor x, which depends only on v, since there is nothing else it could depend on that could allow us to form a unitless quantity. In this frame the light rays have energies $ED(v)$ and $ED(-v)$. If conservation of energy is to hold in the new frame as it did in the old, we must have $2xE = ED(v) + ED(-v)$. After some algebra, we find $x = 1/\sqrt{1 - v^2}$. In other words, an object with energy E in its rest frame has energy γE in a frame moving at velocity v relative to the first one. Since γ is never zero, it follows that even an object at rest has some nonzero energy. We define this energy-at-rest as its mass, i.e., $E = m$ in units where $c = 1$.

$P = m\gamma v$

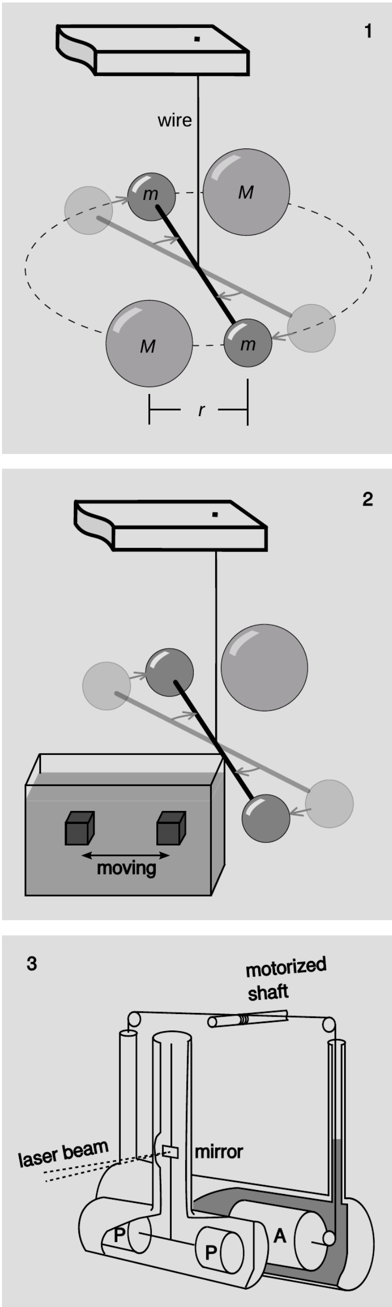
Defining an object's energy-at-rest as its mass only works if this same mass is also a valid measure of inertia. More specifically, we should be able to use this mass to construct a self-consistent logical system in which (1) momentum is conserved, (2) conservation of momentum holds in all frames of reference, and (3) $p \approx mv$ for $v << c$, satisfying the correspondence principle.

Let a material object P, at rest and having mass $2E$, be completely annihilated, creating two beams of light, each with energy E, flying off in opposite directions. A real-world example would be if P consisted of an electron and an antielectron. As shown on p. 715, light has momentum. Because beams of light can be split up or recombined without violating conservation of momentum, a light wave's momentum must be proportional to its energy, $|p| = yE$, where the constant of proportionality y is found in problem 12 on p. 813 but not needed here. Let the momentum of a material object be mvx, where our goal is to prove $x = \gamma$. In this frame of reference, $v = 0$, and conservation of momentum follows by symmetry.

We now change to a new frame of reference, moving at some speed v along the line of emission of the two light rays. In this frame, conservation of momentum requires $2Evx = yE/D - yED$. We therefore have $vx/y = (1/D - D)/2$, which can be shown with a little algebra to equal $v\gamma$. Since only x can depend on v, not y, and the correspondence principle requires $x \approx 1$ for $v << c$, we find that $x = \gamma$, as claimed.

Problem 15 on p. 814 checks that this result also works correctly for a system consisting of material particles.

26.7 ⋆ Two tests of Newton's third law



as / 1. A balance that measures the gravitational attraction between masses M and m. (See section 10.5 for a more detailed description.) When the two masses M are inserted, the fiber twists. 2. A simplified diagram of Kreuzer's modification. The moving teflon mass is submerged in a liquid with nearly the same density. 3. Kreuzer's actual apparatus.

$E = mc^2$ states that a certain amount of energy E is equivalent to a certain amount of mass m. But mass pops up in physics in several different guises: the mass measured by an object's inertia, the "active" gravitational mass m_a that determines the gravitational forces it makes on other objects, and the "passive" gravitational mass m_p that measures how strongly it feels gravity. Einstein's reason for predicting the same behavior for m_a and m_p was that anything else would have violated Newton's third law for gravitational forces.

Suppose instead that an object's energy content contributes only to m_p, not to m_a. Atomic nuclei get something like 1% of their mass from the energy of the electric fields inside their nuclei, but this percentage varies with the number of protons, so if we have objects m and M with different chemical compositions, it follows that in this theory m_p/m_a will not be the same as M_p/M_a, and in this non-Einsteinian version of relativity, Newton's third law is violated.

This was tested in a Princeton PhD-thesis experiment by Kreuzer[11] in 1966. Kreuzer carried out an experiment, figure as, using masses made of two different substances. The first substance was teflon. The second substance was a mixture of the liquids trichloroethylene and dibromoethane, with the proportions chosen so as to give a passive-mass density as close as possible to that of teflon, as determined by the neutral buoyancy of the teflon masses suspended inside the liquid. If the active-mass densities of these substances are not strictly proportional to their passive-mass densities, then moving the chunk of teflon back and forth in figure as/2 would change the gravitational force acting on the nearby small sphere. No such change was observed, and the results verified $m_p/m_a = M_p/M_a$ to within one part in 10^6, in agreement with Einstein and Newton. If electrical energy had not contributed at all to active mass, then a violation of the third law would have been detected at the level of about one part in 10^2.

The Kreuzer result was improved in 1986 by Bartlett and van Buren[12] using data gathered by bouncing laser beams off of a mirror left behind on the moon by the Apollo astronauts, as described p. 277. Since the moon has an asymmetrical distribution of iron

[11] Kreuzer, Phys. Rev. 169 (1968) 1007
[12] Phys. Rev. Lett. 57 (1986) 21

and aluminum, a theory with $m_p/m_a \neq M_p/M_a$ would cause it to have an anomalous acceleration along a certain line. The lack of any such observed acceleration limits violations of Newton's third law to about one part in 10^{10}.

Summary

Selected vocabulary

alpha particle . .	a form of radioactivity consisting of helium nuclei
beta particle . . .	a form of radioactivity consisting of electrons
gamma ray	a form of radioactivity consisting of a very high-frequency form of light
proton	a positively charged particle, one of the types that nuclei are made of
neutron	an uncharged particle, the other types that nuclei are made of
isotope	one of the possible varieties of atoms of a given element, having a certain number of neutrons
atomic number .	the number of protons in an atom's nucleus; determines what element it is
atomic mass . . .	the mass of an atom
mass number . .	the number of protons plus the number of neutrons in a nucleus; approximately proportional to its atomic mass
strong nuclear force	the force that holds nuclei together against electrical repulsion
weak nuclear force	the force responsible for beta decay
beta decay	the radioactive decay of a nucleus via the reaction $n \rightarrow p + e^- + \bar{\nu}$ or $p \rightarrow n + e^+ + \nu$; so called because an electron or antielectron is also known as a beta particle
alpha decay . . .	the radioactive decay of a nucleus via emission of an alpha particle
fission	the radioactive decay of a nucleus by splitting into two parts
fusion	a nuclear reaction in which two nuclei stick together to form one bigger nucleus
μSv	a unit for measuring a person's exposure to radioactivity

Notation

e^-	an electron
e^+	an antielectron; just like an electron, but with positive charge
n	a neutron
p	a proton
ν	a neutrino
$\bar{\nu}$	an antineutrino
$\mathcal{E}$	mass-energy

Other terminology and notation

Z	atomic number (number of protons in a nucleus)
N	number of neutrons in a nucleus
A	mass number ($N + Z$)

Summary

Quantization of charge: Millikan's oil drop experiment showed that the total charge of an object could only be an integer multiple of a basic unit of charge (e). This supported the idea the the "flow" of electrical charge was the motion of tiny particles rather than the motion of some sort of mysterious electrical fluid.

Einstein's analysis of Brownian motion was the first definitive proof of the existence of atoms. Thomson's experiments with vacuum tubes demonstrated the existence of a new type of microscopic particle with a very small ratio of mass to charge. Thomson correctly interpreted these as building blocks of matter even smaller than atoms: the first discovery of subatomic particles. These particles are called electrons.

The above experimental evidence led to the first useful model of the interior structure of atoms, called the raisin cookie model. In the raisin cookie model, an atom consists of a relatively large, massive, positively charged sphere with a certain number of negatively charged electrons embedded in it.

Rutherford and Marsden observed that some alpha particles from a beam striking a thin gold foil came back at angles up to 180 degrees. This could not be explained in the then-favored raisin-cookie model of the atom, and led to the adoption of the planetary model of the atom, in which the electrons orbit a tiny, positively-charged nucleus. Further experiments showed that the nucleus itself was a cluster of positively-charged protons and uncharged neutrons.

Radioactive nuclei are those that can release energy. The most common types of radioactivity are alpha decay (the emission of a helium nucleus), beta decay (the transformation of a neutron into a proton or vice-versa), and gamma decay (the emission of a type of very-high-frequency light). Stars are powered by nuclear fusion reactions, in which two light nuclei collide and form a bigger nucleus, with the release of energy.

Human exposure to ionizing radiation is measured in units of microsieverts (μSv). The typical person is exposed to about 2000 μSv worth of natural background radiation per year.

Exploring further

The First Three Minutes, Steven Weinberg. This book describes the first three minutes of the universe's existence.

Problems

Key

- √ A computerized answer check is available online.
- ∫ A problem that requires calculus.
- ⋆ A difficult problem.

1 Use the nutritional information on some packaged food to make an order-of-magnitude estimate of the amount of chemical energy stored in one atom of food, in units of joules. Assume that a typical atom has a mass of 10^{-26} kg. This constitutes a rough estimate of the amounts of energy there are on the atomic scale. [See chapter 1 for help on how to do order-of-magnitude estimates. Note that a nutritional "calorie" is really a kilocalorie.] √

2 The nuclear process of beta decay by electron capture is described parenthetically on p. 783. The reaction is $\mathrm{p}+\mathrm{e}^- \rightarrow \mathrm{n}+\nu$.
(a) Show that charge is conserved in this reaction.
(b) Explain why electron capture doesn't occur in hydrogen atoms. (If it did, matter wouldn't exist!) ▷ Solution, p. 1035

3 ^{241}Pu decays either by electron decay or by alpha decay. (A given ^{241}Pu nucleus may do either one; it's random.) What are the isotopes created as products of these two modes of decay?

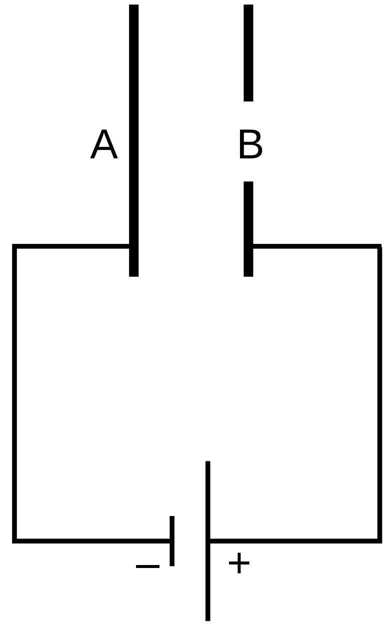
4 (a) Recall that the gravitational energy of two gravitationally interacting spheres is given by $PE_g = -Gm_1m_2/r$, where r is the center-to-center distance. What would be the analogous equation for two electrically interacting spheres? Justify your choice of a plus or minus sign on physical grounds, considering attraction and repulsion. √
(b) Use this expression to estimate the energy required to pull apart a raisin-cookie atom of the one-electron type, assuming a radius of 10^{-10} m. √
(c) Compare this with the result of problem 1.

5 A neon light consists of a long glass tube full of neon, with metal caps on the ends. Positive charge is placed on one end of the tube, negative on the other. The electric forces generated can be strong enough to strip electrons off of a certain number of neon atoms. Assume for simplicity that only one electron is ever stripped off of any neon atom. When an electron is stripped off of an atom, both the electron and the neon atom (now an ion) have electric charge, and they are accelerated by the forces exerted by the charged ends of the tube. (They do not feel any significant forces from the other ions and electrons within the tube, because only a tiny minority of neon atoms ever gets ionized.) Light is finally produced when ions are reunited with electrons. Give a numerical comparison of the magnitudes and directions of the accelerations of the electrons and ions. [You may need some data from page 1062.] √

6 If you put two hydrogen atoms near each other, they will feel an attractive force, and they will pull together to form a molecule.

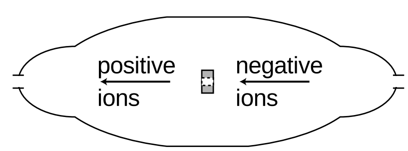
(Molecules consisting of two hydrogen atoms are the normal form of hydrogen gas.) How is this possible, since each is electrically neutral? Shouldn't the attractive and repulsive forces all cancel out exactly? Use the raisin cookie model. (Students who have taken chemistry often try to use fancier models to explain this, but if you can't explain it using a simple model, you probably don't understand the fancy model as well as you thought you did!) It's not so easy to prove that the force should actually be attractive rather than repulsive, so just concentrate on explaining why it doesn't necessarily have to vanish completely.

7 The figure shows a simplified diagram of an electron gun such as the one used in an old-fashioned TV tube. Electrons that spontaneously emerge from the negative electrode (cathode) are then accelerated to the positive electrode, which has a hole in it. (Once they emerge through the hole, they will slow down. However, if the two electrodes are fairly close together, this slowing down is a small effect, because the attractive and repulsive forces experienced by the electron tend to cancel.) (a) If the voltage difference between the electrodes is ΔV, what is the velocity of an electron as it emerges at B? (Assume its initial velocity, at A, is negiligible.) (b) Evaluate your expression numerically for the case where $\Delta V = 10$ kV, and compare to the speed of light. ▷ Solution, p. 1035



Problem 7.

8 The figure shows a simplified diagram of a device called a tandem accelerator, used for accelerating beams of ions up to speeds on the order of 1-10% of the speed of light. (Since these velocities are not too big compared to c, you can use nonrelativistic physics throughout this problem.) The nuclei of these ions collide with the nuclei of atoms in a target, producing nuclear reactions for experiments studying the structure of nuclei. The outer shell of the accelerator is a conductor at zero voltage (i.e., the same voltage as the Earth). The electrode at the center, known as the "terminal," is at a high positive voltage, perhaps millions of volts. Negative ions with a charge of −1 unit (i.e., atoms with one extra electron) are produced offstage on the right, typically by chemical reactions with cesium, which is a chemical element that has a strong tendency to give away electrons. Relatively weak electric and magnetic forces are used to transport these −1 ions into the accelerator, where they are attracted to the terminal. Although the center of the terminal has a hole in it to let the ions pass through, there is a very thin carbon foil there that they must physically penetrate. Passing through the foil strips off some number of electrons, changing the atom into a positive ion, with a charge of $+n$ times the fundamental charge. Now that the atom is positive, it is repelled by the terminal, and accelerates some more on its way out of the accelerator.



Problem 8.

(a) Find the velocity, v, of the emerging beam of positive ions, in terms of n, their mass m, the terminal voltage V, and fundamental constants. Neglect the small change in mass caused by the loss of electrons in the stripper foil. ✓

(b) To fuse protons with protons, a minimum beam velocity of about 11% of the speed of light is required. What terminal voltage would be needed in this case? ✓

(c) In the setup described in part b, we need a target containing atoms whose nuclei are single protons, i.e., a target made of hydrogen. Since hydrogen is a gas, and we want a foil for our target, we have to use a hydrogen compound, such as a plastic. Discuss what effect this would have on the experiment.

9 In example 6 on page 683, I remarked that accelerating a macroscopic (i.e., not microscopic) object to close to the speed of light would require an unreasonable amount of energy. Suppose that the starship Enterprise from Star Trek has a mass of 8.0×10^7 kg, about the same as the Queen Elizabeth 2. Compute the kinetic energy it would have to have if it was moving at half the speed of light. Compare with the total energy content of the world's nuclear arsenals, which is about 10^{21} J. ✓

10 (a) A free neutron (as opposed to a neutron bound into an atomic nucleus) is unstable, and undergoes beta decay (which you may want to review). The masses of the particles involved are as follows:

neutron	1.67495×10^{-27} kg
proton	1.67265×10^{-27} kg
electron	0.00091×10^{-27} kg
antineutrino	$< 10^{-35}$ kg

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

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

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

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

11 (a) Find a relativistic equation for the velocity of an object in terms of its mass and momentum (eliminating γ). Use natural units (i.e., discard factors of c) throughout. √
(b) Show that your result is approximately the same as the nonrelativistic value, p/m, at low velocities.
(c) Show that very large momenta result in speeds close to the speed of light.
(d) Insert factors of c to make your result from part a usable in SI units. √
⋆

12 An object moving at a speed very close to the speed of light is referred to as ultrarelativistic. Ordinarily (luckily) the only ultrarelativistic objects in our universe are subatomic particles, such as cosmic rays or particles that have been accelerated in a particle accelerator.
(a) What kind of number is γ for an ultrarelativistic particle?
(b) Repeat example 12 on page 802, but instead of very low, nonrelativistic speeds, consider ultrarelativistic speeds.
(c) Find an equation for the ratio $\mathcal{E}/p$. The speed may be relativistic, but don't assume that it's ultrarelativistic. √
(d) Simplify your answer to part c for the case where the speed is ultrarelativistic. √
(e) We can think of a beam of light as an ultrarelativistic object — it certainly moves at a speed that's sufficiently close to the speed of light! Suppose you turn on a one-watt flashlight, leave it on for one second, and then turn it off. Compute the momentum of the recoiling flashlight, in units of kg·m/s. (Cf. p. 715.) √
(f) Discuss how your answer in part e relates to the correspondence principle.

13 As discussed in section 19.2, the speed at which a disturbance travels along a string under tension is given by $v = \sqrt{T/\mu}$, where μ is the mass per unit length, and T is the tension.
(a) Suppose a string has a density ρ, and a cross-sectional area A. Find an expression for the maximum tension that could possibly exist in the string without producing $v > c$, which is impossible according to relativity. Express your answer in terms of ρ, A, and c. The interpretation is that relativity puts a limit on how strong any material can be. √
(b) Every substance has a tensile strength, defined as the force per unit area required to break it by pulling it apart. The tensile strength is measured in units of $\mathrm{N/m^2}$, which is the same as the pascal (Pa), the mks unit of pressure. Make a numerical estimate of the maximum tensile strength allowed by relativity in the case where the rope is made out of ordinary matter, with a density on the same order of magnitude as that of water. (For comparison, kevlar has a tensile strength of about 4×10^9 Pa, and there is speculation that fibers made from carbon nanotubes could have values as high as 6×10^{10} Pa.) √

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

14 (a) A charged particle is surrounded by a uniform electric field. Starting from rest, it is accelerated by the field to speed v after traveling a distance d. Now it is allowed to continue for a further distance $3d$, for a total displacement from the start of $4d$. What speed will it reach, assuming newtonian physics?
(b) Find the relativistic result for the case of $v = c/2$.

15 Problem 15 on p. 390 (with the solution given in the back of the book) demonstrates that in Newtonian mechanics, conservation of momentum is the necessary and sufficient condition if conservation of energy is to hold in all frames of reference. The purpose of this problem is to generalize that idea to relativity (in one dimension).

Let a system contain two interacting particles, each with unit mass. Then if energy is conserved in a particular frame, we must have $\gamma_1 + \gamma_2 = \gamma_1' + \gamma_2'$, where the primes indicate the quantities after interaction. Suppose that we now change to a new frame, in motion relative to the first one at a velocity ϵ that is much less than 1 (in units where $c = 1$). The velocities all change according to the result of problem 21 on p. 727. Show that energy is conserved in the new frame if and only if momentum is conserved.

Hints: (1) Since ϵ is small, you can take $1/(1+\epsilon) \approx 1-\epsilon$. (2) Rather than directly using the result of problem 21, it is easier to eliminate the velocities in favor of the corresponding Doppler-shift factors D, which simply multiply when the velocities are combined. (3) The identity $v\gamma = (1/D - D)/2$ is handy here. ⋆

16 (a) Let L be the diameter of our galaxy. Suppose that a person in a spaceship of mass m wants to travel across the galaxy at constant speed, taking proper time τ. Find the kinetic energy of the spaceship. (b) Your friend is impatient, and wants to make the voyage in an hour. For $L = 10^5$ light years, estimate the energy in units of megatons of TNT (1 megaton=4×10^9 J).

Exercise 26A: Sports in Slowlightland

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

- relativistic momentum
- relativistic kinetic energy
- relativistic addition of velocities (See problem 21, with the answer given on p. 1034.)
- time dilation and length contraction
- Doppler shifts of light (See section 24.7.)
- equivalence of mass and energy
- time it takes for light to get to an athlete's eye
- deflection of light rays by gravity

Exercise 26B: Nuclear decay

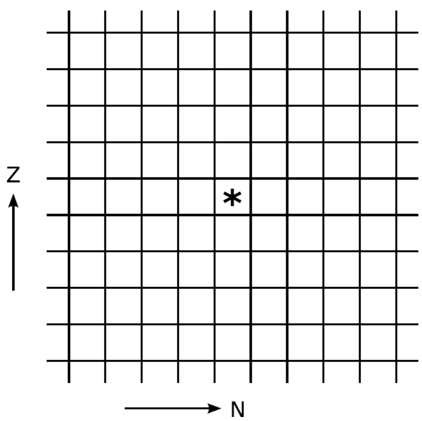
1. Consulting a periodic table, find the N, Z, and A of the following:

	N	Z	A
^{4}He			
^{244}Pu			

2. Consider the following five decay processes:

- α decay
- γ decay
- $p \rightarrow n + e^{+} + \nu$ (β^{+} decay)
- $n \rightarrow p + e^{-} + \bar{\nu}$ (β^{-} decay)
- $p + e^{-} \rightarrow n + \nu$ (electron capture)

What would be the action of each of these on the chart of the nuclei? The * represents the original nucleus.



3. (a) Suppose that ^{244}Pu undergoes perfectly symmetric fission, and also emits two neutrons. Find the daughter isotope.

(b) Is the daughter stable, or is it neutron-rich or -poor relative to the line of stability? (To estimate what's stable, you can use a large chart of the nuclei, or, if you don't have one handy, consult a periodic table and use the average atomic mass as an approximation to the stable value of A.)

(c) Consulting the chart of the nuclei (fig. ac on p. 788), explain why it turns out this way.

(d) If the daughter is unstable, which process from question #2 would you expect it to decay by?

Exercise 26C: Misconceptions about relativity

The following is a list of common misconceptions about relativity. The class will be split up into random groups, and each group will cooperate on developing an explanation of the misconception, and then the groups will present their explanations to the class. There may be multiple rounds, with students assigned to different randomly chosen groups in successive rounds.

1. How can light have momentum if it has zero mass?
2. What does the world look like in a frame of reference moving at c?
3. Alice observes Betty coming toward her from the left at $c/2$, and Carol from the right at $c/2$. Therefore Betty is moving at the speed of light relative to Carol.
4. Are relativistic effects such as length contraction and time dilation real, or do they just seem to be that way?
5. Special relativity only matters if you're moving close to the speed of light.
6. Special relativity says that everything is relative.
7. There is a common misconception that relativistic length contraction is what we would actually *see*. Refute this by drawing a spacetime diagram for an object approaching an observer, and tracing rays of light emitted from the object's front and back that both reach the observer's eye at the same time.
8. When you travel close to the speed of light, your time slows down.
9. Is a light wave's wavelength relativistically length contracted by a factor of gamma?
10. Accelerate a baseball to ultrarelativistic speeds. Does it become a black hole?
11. Where did the Big Bang happen?
12. The universe can't be infinite in size, because it's only had a finite amount of time to expand from the point where the Big Bang happened.

Chapter 27
General relativity

What you've learned so far about relativity is known as the special theory of relativity, which is compatible with three of the four known forces of nature: electromagnetism, the strong nuclear force, and the weak nuclear force. Gravity, however, can't be shoehorned into the special theory. In order to make gravity work, Einstein had to generalize relativity. The resulting theory is known as the general theory of relativity.[1]

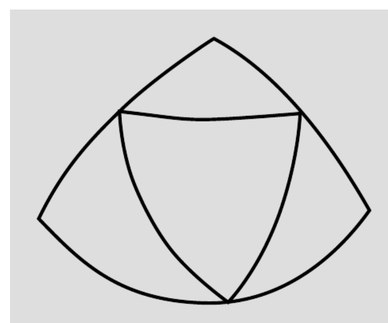
27.1 Our universe isn't Euclidean

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

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

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

Postulates of Euclidean geometry:
1. Two points determine a line.
2. Line segments can be extended.
3. A unique circle can be constructed given any point as its center and any line segment as its radius.
4. All right angles are equal to one another.
5. Given a line and a point not on the line, no more than one line can be drawn through the point and parallel to the given line.

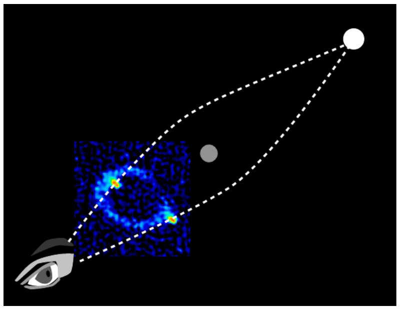


a / Noneuclidean effects, such as the discrepancy from 180° in the sum of the angles of a triangle, are expected to be proportional to area. Here, a noneuclidean equilateral triangle is cut up into four smaller equilateral triangles, each with 1/4 the area. As proved in problem 1, the discrepancy is quadrupled when the area is quadrupled.

[1] Einstein originally described the distinction between the two theories by saying that the special theory applied to nonaccelerating frames of reference, while the general one allowed any frame at all. The modern consensus is that Einstein was misinterpreting his own theory, and that special relativity actually handles accelerating frames just fine.

from 180° is expected to be proportional to the area of the triangle.

But we already know that light rays are bent by gravity. We expect it based on $E = mc^2$, which tells us that the energy of a light ray is equivalent to a certain amount of mass, and furthermore it has been verified experimentally by the deflection of starlight by the sun (example 8, p. 799). We therefore know that our universe is noneuclidean, and we gain the further insight that the level of deviation from Euclidean behavior depends on gravity.



b / An Einstein's ring. The distant object is a quasar, MG1131+0456, and the one in the middle is an unknown object, possibly a supermassive black hole. The intermediate object's gravity focuses the rays of light from the distant one. Because the entire arrangement lacks perfect axial symmetry, the ring is nonuniform; most of its brightness is concentrated in two lumps on opposite sides.

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

Einstein's ring *example 1*

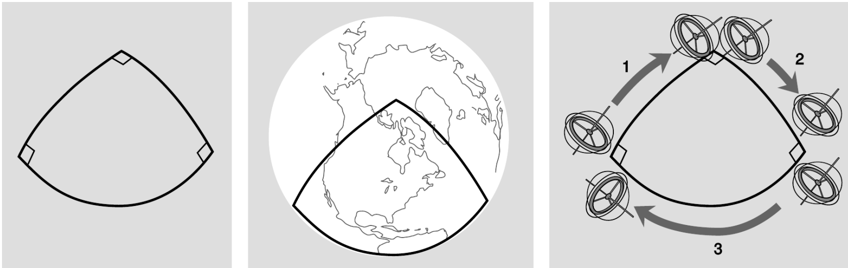
An Einstein's ring, figure b, is formed when there is a chance alignment of a distant source with a closer gravitating body. This type of gravitational lensing is direct evidence for the noneuclidean nature of space. The two light rays are lines, and they violate Euclid's first postulate, that two points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

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

Curvature

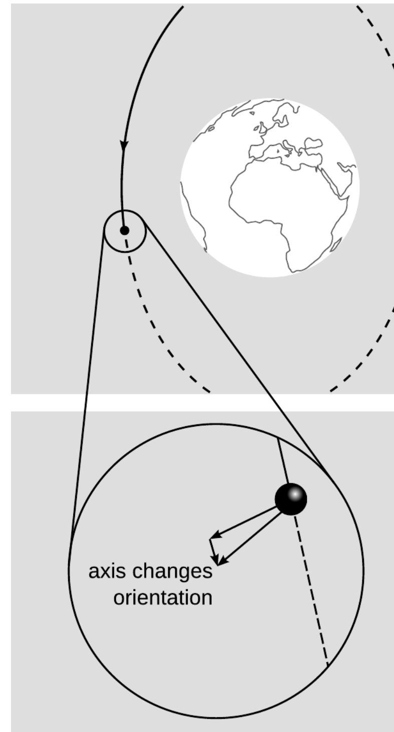
An example of such a non-optical measurement is the Gravity Probe B satellite, figure d, which was launched into a polar orbit in 2004 and operated until 2010. The probe carried four gyroscopes made of quartz, which were the most perfect spheres ever manufactured, varying from sphericity by no more than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

The triangle in figure c has angles that add up to more than 180°.



c / *Left*: A 90-90-90 triangle. Its angles add up to more than 180°. *Middle*: The triangle "pops" off the page visually. We intuitively want to visualize it as lying on a curved surface such as the earth's. *Right*: A gyroscope carried smoothly around its perimeter ends up having changed its orientation when it gets back to its starting point.

This type of curvature is referred to as positive. It is also possible to have negative curvature, as in figure e.

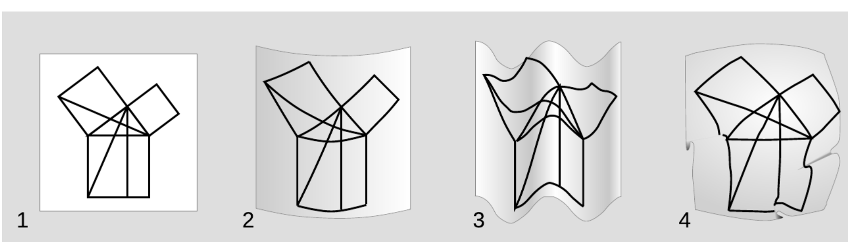


d / Gravity Probe B was in a polar orbit around the earth. As in the right panel of figure c, the orientation of the gyroscope changes when it is carried around a curve and back to its starting point. Because the effect was small, it was necessary to let it accumulate over the course of 5000 orbits in order to make it detectable.

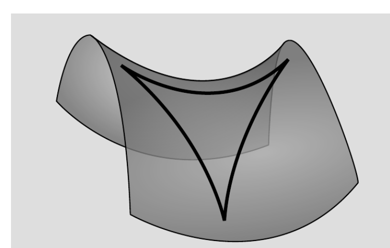
In general relativity, curvature isn't just something caused by gravity. Gravity *is* curvature, and the curvature involves both space and time, as may become clearer once you get to figure k. Thus the distinction between special and general relativity is that general relativity handles curved spacetime, while special relativity is restricted to the case where spacetime is flat.

Curvature doesn't require higher dimensions

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



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



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

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

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

27.2 The equivalence principle

Universality of free-fall

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

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

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

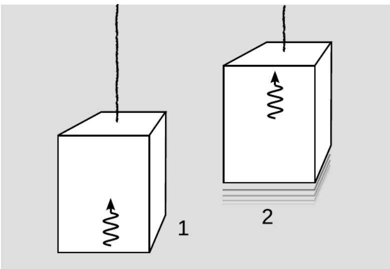
There is absolutely no way to tell which of these two scenarios is actually the case — unless you happen to have a chunk of FloatyStuff in your other pocket. If you release the FloatyStuff and it hovers above the deck, then you're on another planet and experiencing genuine gravity; your keychain responded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

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

The artificial horizon *example 2*

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

g / An artificial horizon.



h / 1. A ray of light is emitted upward from the floor of the elevator. The elevator accelerates upward. 2. By the time the light is detected at the ceiling, the elevator has changed its velocity, so the light is detected with a Doppler shift.

i / Pound and Rebka at the top and bottom of the tower.

output over the preceding several minutes. If the plane is flown in circles for several minutes, the artificial horizon will be fooled into indicating that the wrong direction is vertical.

Gravitational Doppler shifts and time dilation

An interesting application of the equivalence principle is the explanation of gravitational time dilation. As described on p. 672, experiments show that a clock at the top of a mountain runs faster than one down at its foot.

To calculate this effect, we make use of the fact that the gravitational field in the area around the mountain is equivalent to an acceleration. Suppose we're in an elevator accelerating upward with acceleration a, and we shoot a ray of light from the floor up toward the ceiling, at height h. The time Δt it takes the light ray to get to the ceiling is about h/c, and by the time the light ray reaches the ceiling, the elevator has sped up by $v = a\Delta t = ah/c$, so we'll see a red-shift in the ray's frequency. Since v is small compared to c, we don't need to use the fancy Doppler shift equation from section 24.7; we can just approximate the Doppler shift factor as $1 - v/c \approx 1 - ah/c^2$. By the equivalence principle, we should expect that if a ray of light starts out low down and then rises up through a gravitational field g, its frequency will be Doppler shifted by a factor of $1 - gh/c^2$. This effect was observed in a famous experiment carried out by Pound and Rebka in 1959. Gamma-rays were emitted at the bottom of a 22.5-meter tower at Harvard and detected at the top with the Doppler shift predicted by general relativity. (See problem 4.)

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

We have built up a picture of light rays interacting with gravity. To confirm that this make sense, recall that we have already observed on p. 715 and in problem 12 on p. 813 that light has momentum. The equivalence principle says that whatever has inertia must also participate in gravitational interactions. Therefore light waves must have weight, and must lose energy when they rise through a gravitational field(cf. p. 804).

Local flatness

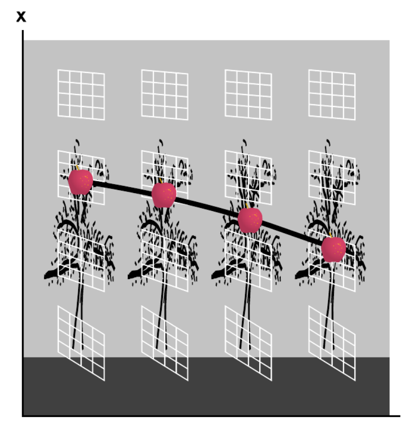
The noneuclidean nature of spacetime produces effects that grow in proportion to the area of the region being considered. Interpreting such effects as evidence of curvature, we see that this connects naturally to the idea that curvature is undetectable from close up.

For example, the curvature of the earth's surface is not normally noticeable to us in everyday life. Locally, the earth's surface is flat, and the same is true for spacetime.

Local flatness turns out to be another way of stating the equivalence principle. In a variation on the alien-abduction story, suppose that you regain consciousness aboard the flying saucer and find yourself weightless. If the equivalence principle holds, then you have no way of determining from local observations, inside the saucer, whether you are actually weightless in deep space, or simply free-falling in apparent weightlessness, like the astronauts aboard the International Space Station. That means that locally, we can always adopt a free-falling frame of reference in which there is no gravitational field at all. If there is no gravity, then special relativity is valid, and we can treat our local region of spacetime as being approximately flat.

j / The earth is flat — locally.

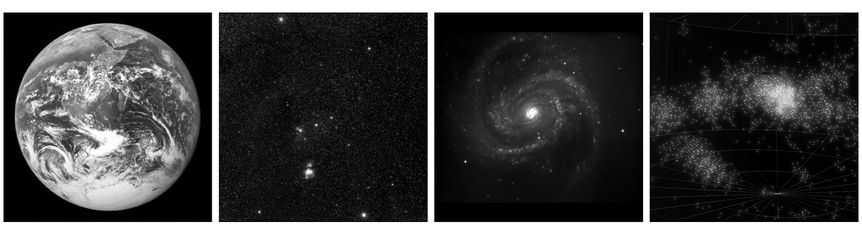
In figure k, an apple falls out of a tree. Its path is a "straight" line in spacetime, in the same sense that the equator is a "straight" line on the earth's surface.



k / Spacetime is locally flat.

Inertial frames

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



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

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

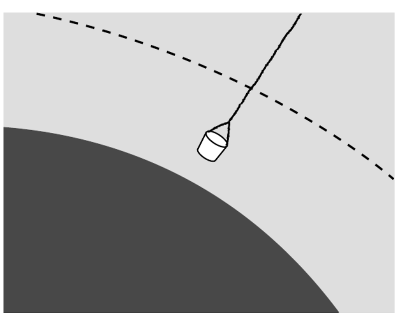
There is a different definition of an inertial frame that works better in relativity. A Newtonian inertial frame was defined by an object that isn't subject to any forces, gravitational or otherwise. In general relativity, we instead define an inertial frame using an object that that isn't influenced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

27.3 Black holes

The observations described so far showed only small effects from curvature. To get a big effect, we should look at regions of space in which there are strong gravitational fields. The prime example is a black hole. The best studied examples are two objects in our own galaxy: Cygnus X-1, which is believed to be a black hole with about ten times the mass of our sun, and Sagittarius A*, an object near the center of our galaxy with about four million solar masses. (See problem 14, p. 283 for how we know Sagittarius A*'s mass.)

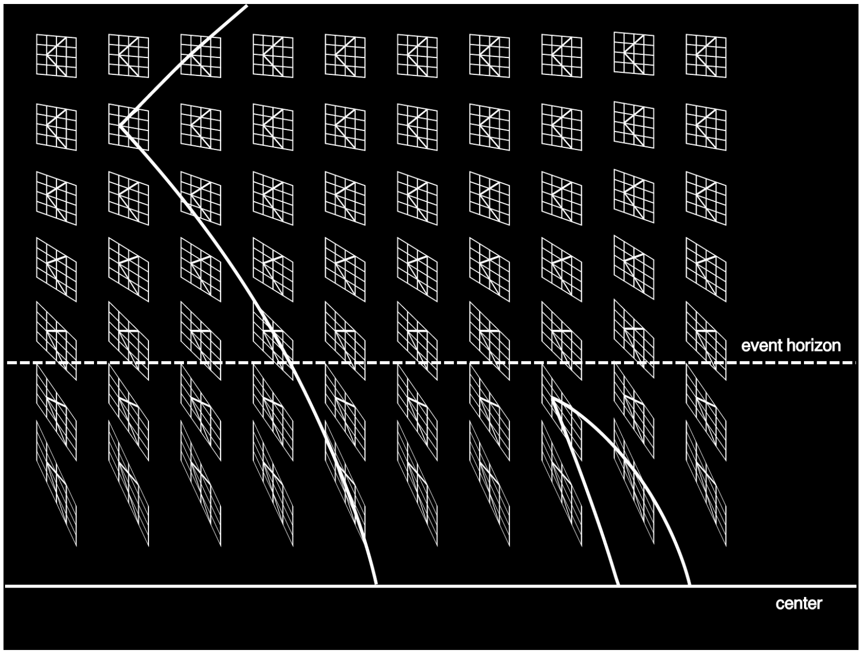
Although a black hole is a relativistic object, we can gain some insight into how it works by applying Newtonian physics. As shown in problem 21 on p. 358, a spherical body of mass M has an escape velocity $v = \sqrt{2GM/r}$, which is the minimum velocity that we would need to give to a projectile shot from a distance r so that it would never fall back down. If r is small enough, the escape velocity will be greater than c, so that even a ray of light can never escape.

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



m / Matter is lifted out of a Newtonian black hole with a bucket. The dashed line represents the point at which the escape velocity equals the speed of light. For a real, relativistic black hole, this is impossible.

In the Newtonian description of a black hole, matter could be lifted out of a black hole, m. Would this be possible with a real-world black hole, which is relativistic rather than Newtonian? No, because the bucket is causally separated from the outside universe. No rope would be strong enough for this job (problem 13, p. 813).



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

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

Although the light rays in figure n don't speed up or slow down, they do experience gravitational Doppler shifts. If a light ray is emitted from just above the event horizon, then it will escape to an infinite distance, but it will suffer an extreme Doppler shift toward low frequencies. A distant observer also has the option of interpreting this as a gravitational time dilation that greatly lowers the frequency of the oscillating electric charges that produced the ray. If the point of emission is made closer and closer to the horizon, the frequency and energy (see p. 804) measured by a distant observer approach zero, making the ray impossible to observe.

Information paradox

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

Newton's laws have built into them the implicit assumption that omniscience is possible, at least in principle. For example, Newton's definition of an inertial frame of reference leads to an infinite regress, as described on p. 825. For Newton this isn't a problem, because in principle an omnisicient observer can know the location of every mass in the universe. In this conception of the cosmos, there are no theoretical limits on human knowledge, only practical ones; if we could gather sufficiently precise data about the state of the universe at one time, and if we could carry out all the calculations to extrapolate into the future, then we could know everything that would ever happen. (See the famous quote by Laplace on p. 18.)

But the existence of event horizons surrounding black holes makes it impossible for any observer to be omniscient; only an observer inside a particular horizon can see what's going on inside that horizon.

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

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

This destruction of information is known as the black hole information paradox, and it's referred to as a paradox because quantum physics (ch. 33-36) has built into its DNA the requirement that information is never lost in this sense.

Formation

Around 1960, as black holes and their strange properties began to be better understood and more widely discussed, many physicists who found these issues distressing comforted themselves with the belief that black holes would never really form from realistic initial conditions, such as the collapse of a massive star. Their skepticism was not entirely unreasonable, since it is usually very hard in astronomy to hit a gravitating target, the reason being that conservation of angular momentum tends to make the projectile swing past. (See problem 13 on p. 426 for a quantitative analysis.) For example, if we wanted to drop a space probe into the sun, we would have to extremely precisely stop its sideways orbital motion so that it would drop almost exactly straight in. Once a star started to collapse, the theory went, and became relatively compact, it would be such a small target that further infalling material would be unlikely to hit it, and the process of collapse would halt. According to this point of view, theorists who had calculated the collapse of a star into a black hole had been oversimplifying by assuming a star that was initially perfectly spherical and nonrotating. Remove the unrealistically perfect symmetry of the initial conditions, and a black hole would never actually form.

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

27.4 Cosmology

The Big Bang

Section 19.5 presented the evidence, discovered by Hubble, that the universe is expanding in the aftermath of the Big Bang: when we observe the light from distant galaxies, it is always Doppler-shifted toward the red end of the spectrum, indicating that no matter what direction we look in the sky, everything is rushing away from us. This seems to go against the modern attitude, originated by Copernicus, that we and our planet do not occupy a special place in the universe. Why is everything rushing away from *our* planet in particular? But general relativity shows that this anti-Copernican conclusion is wrong. General relativity describes space not as a rigidly defined background but as something that can curve and stretch, like a sheet of rubber. We imagine all the galaxies as existing on the surface of such a sheet, which then expands uniformly. The space between the galaxies (but not the galaxies themselves) grows at a steady rate, so that any observer, inhabiting any galaxy, will see every other galaxy as receding. There is therefore no privileged or special location in the universe.

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

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

As with black holes, there was considerable skepticism about whether the existence of an initial singularity in these models was an arifact of the unrealistically perfect uniformity assumed in the models. Perhaps in the real universe, extrapolation of all the paths of the galaxies backward in time would show them missing each other by millions of light-years. But in 1972 Stephen Hawking proved a variant on the Penrose singularity theorem that applied to Big Bang singularities. By the Hawking singularity theorem, the level of uniformity we see in the present-day universe is more than sufficient to prove that a Big Bang singularity must have existed.

The cosmic censorship hypothesis

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

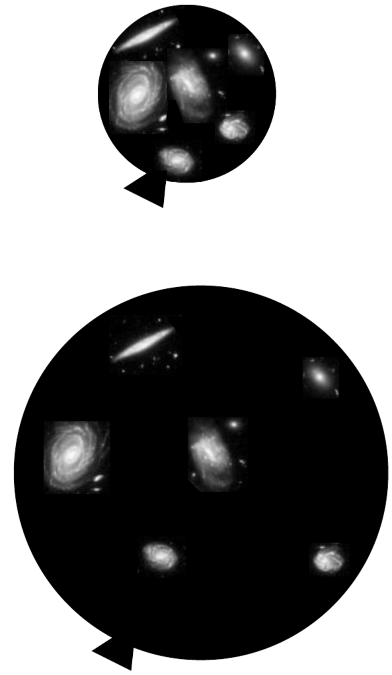
That would be such a devastating defeat for the enterprise of physics that in 1969 Penrose proposed an alternative, humorously named the "cosmic censorship hypothesis," which states that every singularity in our universe, other than the Big Bang, is hidden behind an event horizon. Therefore if green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

The advent of high-precision cosmology

We expect that if there is matter in the universe, it should have gravitational fields, and in the rubber-sheet analogy this should be represented as a curvature of the sheet. Instead of a flat sheet, we can have a spherical balloon, so that cosmological expansion is like inflating it with more and more air. It is also possible to have negative curvature, as in figure e on p. 821. All three of these are valid, possible cosmologies according to relativity. The positive-curvature type happens if the average density of matter in the universe is above a certain critical level, the negative-curvature one if the density is below that value.

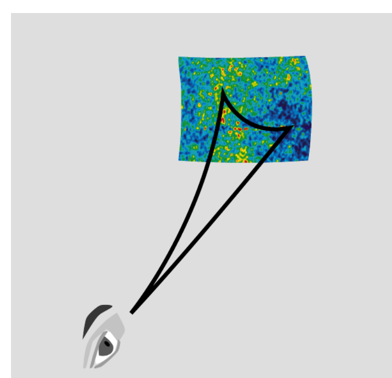
To find out which type of universe we inhabit, we could try to take a survey of the matter in the universe and determine its average density. Historically, it has been very difficult to do this, even to within an order of magnitude. Most of the matter in the universe probably doesn't emit light, making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

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



p / An expanding universe with positive spatial curvature can be imagined as a balloon being blown up. Every galaxy's distance from every other galaxy increases, but no galaxy is the center of the expansion.

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



q / The angular scale of fluctuations in the cosmic microwave background can be used to infer the curvature of the universe.

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

Dark energy and dark matter

Not everything works out so smoothly, however. One surprise, discussed in section 10.6, is that the universe's expansion is not currently slowing down, as had been expected due to the gravitational attraction of all the matter in it. Instead, it is currently speeding up. This is attributed to a variable in Einstein's equations, long assumed to be zero, which represents a universal gravitational repulsion of space itself, occurring even when there is no matter present. The current name for this is "dark energy," although the fancy name is just a label for our ignorance about what causes it.

Another surprise comes from attempts to model the formation of the elements during the era shortly after the Big Bang, before the formation of the first stars (section 26.4.10). The observed relative abundances of hydrogen, helium, and deuterium (^{2}H) cannot be reconciled with the density of low-velocity matter inferred from the observational data. If the inferred mass density were entirely due to normal matter (i.e., matter whose mass consisted mostly of protons and neutrons), then nuclear reactions in the dense early universe should have proceeded relatively efficiently, leading to a much higher ratio of helium to hydrogen, and a much lower abundance of deuterium. The conclusion is that most of the matter in the universe must be made of an unknown type of exotic matter, known as "dark matter." We are in the ironic position of knowing that precisely 96% of the universe is something other than atoms, but knowing nothing about what that something is. As of 2013, there have been several experiments that have been carried out to attempt the direct detection of dark matter particles. These are carried out at the bottom of mineshafts to eliminate background radiation. Early claims of success appear to have been statistical flukes, and the most sensitive experiments have not detected anything.[2]

[2] arxiv.org/abs/1310.8214

Problems

Key

√ A computerized answer check is available online.

∫ A problem that requires calculus.

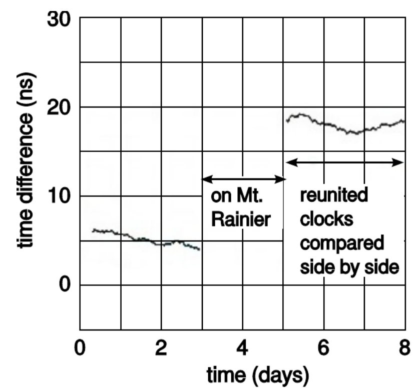
⋆ A difficult problem.

1 Prove, as claimed in the caption of figure a on p. 819, that $S - 180° = 4(s - 180°)$, where S is the sum of the angles of the large equilateral triangle and s is the corresponding sum for one of the four small ones. ▷ Solution, p. 1035

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

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

4 (a) Calculate the Doppler shift to be expected in the Pound-Rebka experiment described on p. 824. (b) In the 1978 Iijima mountain-valley experiment (p. 672), analysis was complicated by the clock's sensitivity to pressure, humidity, and temperature. A cleaner version of the experiment was done in 2005 by hobbyist Tom Van Baak. He put his kids and three of his atomic clocks in a minivan and drove from Bellevue, Washington to a lodge on Mount Rainier, 1340 meters higher in elevation. They spent the weekend there. Back at home, he compared the clocks to others that had stayed at his house. Verify that the effect shown in the graph is as predicted by general relativity.



Problem 4b. Redrawn from Van Baak, Physics Today 60 (2007) 16.

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

6 Section 27.3 presented a Newtonian estimate of how compact an object would have to be in order to be a black hole. Although this estimate is not really right, it turns out to give the right answer to within about a factor of 2. To roughly what size would the earth have to be compressed in order to become a black hole?

7 Clock A sits on a desk. Clock B is tossed up in the air from the same height as the desk and then comes back down. Compare the elapsed times. ▷ Hint, p. 1032 ▷ Solution, p. 1035

8 The angular defect d of a triangle (measured in radians) is defined as $s - \pi$, where s is the sum of the interior angles. The angular defect is proportional to the area A of the triangle. Consider the geometry measured by a two-dimensional being who lives on the surface of a sphere of radius R. First find some triangle on the sphere whose area and angular defect are easy to calculate. Then determine the general equation for d in terms of A and R. √

Exercise 27: Misconceptions about relativity

The following is a list of common misconceptions about relativity. The class will be split up into random groups, and each group will cooperate on developing an explanation of the misconception, and then the groups will present their explanations to the class. There may be multiple rounds, with students assigned to different randomly chosen groups in successive rounds.

1. How can light have momentum if it has zero mass?
2. What does the world look like in a frame of reference moving at c?
3. Alice observes Betty coming toward her from the left at $c/2$, and Carol from the right at $c/2$. Therefore Betty is moving at the speed of light relative to Carol.
4. Are relativistic effects such as length contraction and time dilation real, or do they just seem to be that way?
5. Special relativity only matters if you're moving close to the speed of light.
6. Special relativity says that everything is relative.
7. There is a common misconception that relativistic length contraction is what we would actually *see*. Refute this by drawing a spacetime diagram for an object approaching an observer, and tracing rays of light emitted from the object's front and back that both reach the observer's eye at the same time.
8. When you travel close to the speed of light, your time slows down.
9. Is a light wave's wavelength relativistically length contracted by a factor of gamma?
10. Accelerate a baseball to ultrarelativistic speeds. Does it become a black hole?
11. Where did the Big Bang happen?
12. The universe can't be infinite in size, because it's only had a finite amount of time to expand from the point where the Big Bang happened.

Optics

Chapter 28
The Ray Model of Light

Ads for one Macintosh computer bragged that it could do an arithmetic calculation in less time than it took for the light to get from the screen to your eye. We find this impressive because of the contrast between the speed of light and the speeds at which we interact with physical objects in our environment. Perhaps it shouldn't surprise us, then, that Newton succeeded so well in explaining the motion of objects, but was far less successful with the study of light.

The climax of our study of electricity and magnetism was discovery that light is an electromagnetic wave. Knowing this, however, is not the same as knowing everything about eyes and telescopes. In fact, the full description of light as a wave can be rather cumbersome. We will instead spend most of our treatment of optics making use of a simpler model of light, the ray model, which does a fine job in most practical situations. Not only that, but we will even backtrack a little and start with a discussion of basic ideas about light and vision that predated the discovery of electromagnetic waves.

28.1 The nature of light

The cause and effect relationship in vision

Despite its title, this chapter is far from your first look at light. That familiarity might seem like an advantage, but most people have never thought carefully about light and vision. Even smart people who have thought hard about vision have come up with incorrect ideas. The ancient Greeks, Arabs and Chinese had theories of light and vision, all of which were mostly wrong, and all of which were accepted for thousands of years.

One thing the ancients did get right is that there is a distinction between objects that emit light and objects that don't. When you see a leaf in the forest, it's because three different objects are doing their jobs: the leaf, the eye, and the sun. But luminous objects like the sun, a flame, or the filament of a light bulb can be seen by the eye without the presence of a third object. Emission of light is often, but not always, associated with heat. In modern times, we are familiar with a variety of objects that glow without being heated, including fluorescent lights and glow-in-the-dark toys.

How do we see luminous objects? The Greek philosophers Pythagoras (b. ca. 560 BC) and Empedocles of Acragas (b. ca. 492 BC), who unfortunately were very influential, claimed that when you looked at a candle flame, the flame and your eye were both sending out some kind of mysterious stuff, and when your eye's stuff collided with the candle's stuff, the candle would become evident to your sense of sight.

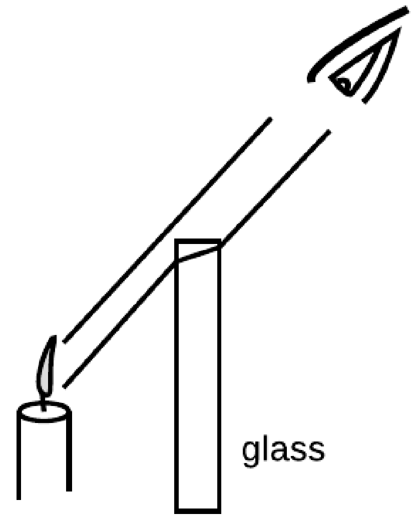
Bizarre as the Greek "collision of stuff theory" might seem, it had a couple of good features. It explained why both the candle and your eye had to be present for your sense of sight to function. The theory could also easily be expanded to explain how we see nonluminous objects. If a leaf, for instance, happened to be present at the site of the collision between your eye's stuff and the candle's stuff, then the leaf would be stimulated to express its green nature, allowing you to perceive it as green.

Modern people might feel uneasy about this theory, since it suggests that greenness exists only for our seeing convenience, implying a human precedence over natural phenomena. Nowadays, people would expect the cause and effect relationship in vision to be the other way around, with the leaf doing something to our eye rather than our eye doing something to the leaf. But how can you tell? The most common way of distinguishing cause from effect is to determine which happened first, but the process of seeing seems to occur too quickly to determine the order in which things happened. Certainly there is no obvious time lag between the moment when you move your head and the moment when your reflection in the mirror moves.

Today, photography provides the simplest experimental evidence that nothing has to be emitted from your eye and hit the leaf in order to make it "greenify." A camera can take a picture of a leaf even if there are no eyes anywhere nearby. Since the leaf appears green regardless of whether it is being sensed by a camera, your eye, or an insect's eye, it seems to make more sense to say that the leaf's greenness is the cause, and something happening in the camera or eye is the effect.

Light is a thing, and it travels from one point to another.

Another issue that few people have considered is whether a candle's flame simply affects your eye directly, or whether it sends out light which then gets into your eye. Again, the rapidity of the effect makes it difficult to tell what's happening. If someone throws a rock at you, you can see the rock on its way to your body, and you can tell that the person affected you by sending a material substance your way, rather than just harming you directly with an arm motion, which would be known as "action at a distance." It is not easy to do a similar observation to see whether there is some "stuff" that travels from the candle to your eye, or whether it is a case of action at a distance.



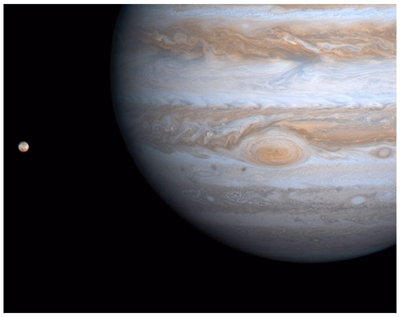
a / Light from a candle is bumped off course by a piece of glass. Inserting the glass causes the apparent location of the candle to shift. The same effect can be produced by taking off your eyeglasses and looking at which you see near the edge of the lens, but a flat piece of glass works just as well as a lens for this purpose.

Newtonian physics includes both action at a distance (e.g., the earth's gravitational force on a falling object) and contact forces such as the normal force, which only allow distant objects to exert forces on each other by shooting some substance across the space between them (e.g., a garden hose spraying out water that exerts a force on a bush).

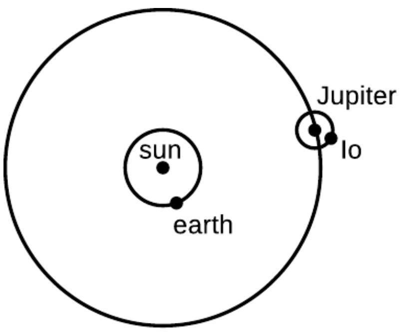
One piece of evidence that the candle sends out stuff that travels to your eye is that as in figure a, intervening transparent substances can make the candle appear to be in the wrong location, suggesting that light is a thing that can be bumped off course. Many people would dismiss this kind of observation as an optical illusion, however. (Some optical illusions are purely neurological or psychological effects, although some others, including this one, turn out to be caused by the behavior of light itself.)

A more convincing way to decide in which category light belongs is to find out if it takes time to get from the candle to your eye; in Newtonian physics, action at a distance is supposed to be instantaneous. The fact that we speak casually today of "the speed of light" implies that at some point in history, somebody succeeded in showing that light did not travel infinitely fast. Galileo tried, and failed, to detect a finite speed for light, by arranging with a person in a distant tower to signal back and forth with lanterns. Galileo uncovered his lantern, and when the other person saw the light, he uncovered his lantern. Galileo was unable to measure any time lag that was significant compared to the limitations of human reflexes.

The first person to prove that light's speed was finite, and to determine it numerically, was Ole Roemer, in a series of measurements around the year 1675. Roemer observed Io, one of Jupiter's moons, over a period of several years. Since Io presumably took the same amount of time to complete each orbit of Jupiter, it could be thought of as a very distant, very accurate clock. A practical and accurate pendulum clock had recently been invented, so Roemer could check whether the ratio of the two clocks' cycles, about 42.5 hours to 1 orbit, stayed exactly constant or changed a little. If the process of seeing the distant moon was instantaneous, there would be no reason for the two to get out of step. Even if the speed of light was finite, you might expect that the result would be only to offset one cycle relative to the other. The earth does not, however, stay at a constant distance from Jupiter and its moons. Since the distance is changing gradually due to the two planets' orbital motions, a finite speed of light would make the "Io clock" appear to run faster as the planets drew near each other, and more slowly as their separation increased. Roemer did find a variation in the apparent speed of Io's orbits, which caused Io's eclipses by Jupiter (the moments when Io passed in front of or behind Jupiter) to occur about 7 minutes early when the earth was closest to Jupiter, and 7 minutes late when it was farthest. Based on these measurements, Roemer estimated the speed of light to be approximately 2×10^8 m/s, which is in the right ballpark compared to modern measurements of 3×10^8 m/s. (I'm not sure whether the fairly large experimental error was mainly due to imprecise knowledge of the radius of the earth's orbit or limitations in the reliability of pendulum clocks.)



b / An image of Jupiter and its moon Io (left) from the Cassini probe.



c / The earth is moving toward Jupiter and Io. Since the distance is shrinking, it is taking less and less time for the light to get to us from Io, and Io appears to circle Jupiter more quickly than normal. Six months later, the earth will be on the opposite side of the sun, and receding from Jupiter and Io, so Io will appear to revolve around Jupiter more slowly.

Light can travel through a vacuum.

Many people are confused by the relationship between sound and light. Although we use different organs to sense them, there are some similarities. For instance, both light and sound are typically emitted in all directions by their sources. Musicians even use visual metaphors like "tone color," or "a bright timbre" to describe sound. One way to see that they are clearly different phenomena is to note their very different velocities. Sure, both are pretty fast compared to a flying arrow or a galloping horse, but as we have seen, the speed of light is so great as to appear instantaneous in most situations. The speed of sound, however, can easily be observed just by watching a group of schoolchildren a hundred feet away as they clap their hands to a song. There is an obvious delay between when you see their palms come together and when you hear the clap.

The fundamental distinction between sound and light is that sound is an oscillation in air pressure, so it requires air (or some other medium such as water) in which to travel. Today, we know that outer space is a vacuum, so the fact that we get light from the sun, moon and stars clearly shows that air is not necessary for the propagation of light.

Discussion questions

A If you observe thunder and lightning, you can tell how far away the storm is. Do you need to know the speed of sound, of light, or of both?

B When phenomena like X-rays and cosmic rays were first discovered, suggest a way one could have tested whether they were forms of light.

C Why did Roemer only need to know the radius of the earth's orbit, not Jupiter's, in order to find the speed of light?

28.2 Interaction of light with matter

Absorption of light

The reason why the sun feels warm on your skin is that the sunlight is being absorbed, and the light energy is being transformed into heat energy. The same happens with artificial light, so the net result of leaving a light turned on is to heat the room. It doesn't matter whether the source of the light is hot, like the sun, a flame, or an incandescent light bulb, or cool, like a fluorescent bulb. (If your house has electric heat, then there is absolutely no point in fastidiously turning off lights in the winter; the lights will help to heat the house at the same dollar rate as the electric heater.)

This process of heating by absorption is entirely different from heating by thermal conduction, as when an electric stove heats spaghetti sauce through a pan. Heat can only be conducted through matter, but there is vacuum between us and the sun, or between us and the filament of an incandescent bulb. Also, heat conduction can only transfer heat energy from a hotter object to a colder one, but a cool fluorescent bulb is perfectly capable of heating something that had already started out being warmer than the bulb itself.

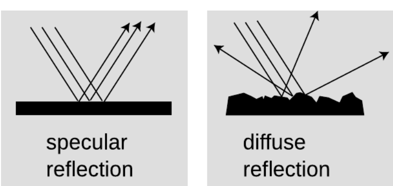
How we see nonluminous objects

Not all the light energy that hits an object is transformed into heat. Some is reflected, and this leads us to the question of how we see nonluminous objects. If you ask the average person how we see a light bulb, the most likely answer is "The light bulb makes light, which hits our eyes." But if you ask how we see a book, they are likely to say "The bulb lights up the room, and that lets me see the book." All mention of light actually entering our eyes has mysteriously disappeared.

Most people would disagree if you told them that light was reflected from the book to the eye, because they think of reflection as something that mirrors do, not something that a book does. They associate reflection with the formation of a reflected image, which does not seem to appear in a piece of paper.

Imagine that you are looking at your reflection in a nice smooth piece of aluminum foil, fresh off the roll. You perceive a face, not a piece of metal. Perhaps you also see the bright reflection of a lamp

d / Two self-portraits of the author, one taken in a mirror and one with a piece of aluminum foil.



e / Specular and diffuse reflection.

over your shoulder behind you. Now imagine that the foil is just a little bit less smooth. The different parts of the image are now a little bit out of alignment with each other. Your brain can still recognize a face and a lamp, but it's a little scrambled, like a Picasso painting. Now suppose you use a piece of aluminum foil that has been crumpled up and then flattened out again. The parts of the image are so scrambled that you cannot recognize an image. Instead, your brain tells you you're looking at a rough, silvery surface.

Mirror-like reflection at a specific angle is known as specular reflection, and random reflection in many directions is called diffuse reflection. Diffuse reflection is how we see nonluminous objects. Specular reflection only allows us to see images of objects other than the one doing the reflecting. In top part of figure d, imagine that the rays of light are coming from the sun. If you are looking down at the reflecting surface, there is no way for your eye-brain system to tell that the rays are not really coming from a sun down below you.

Figure f shows another example of how we can't avoid the conclusion that light bounces off of things other than mirrors. The lamp is one I have in my house. It has a bright bulb, housed in a completely opaque bowl-shaped metal shade. The only way light can get out of the lamp is by going up out of the top of the bowl. The fact that I can read a book in the position shown in the figure means that light must be bouncing off of the ceiling, then bouncing off of the book, then finally getting to my eye.

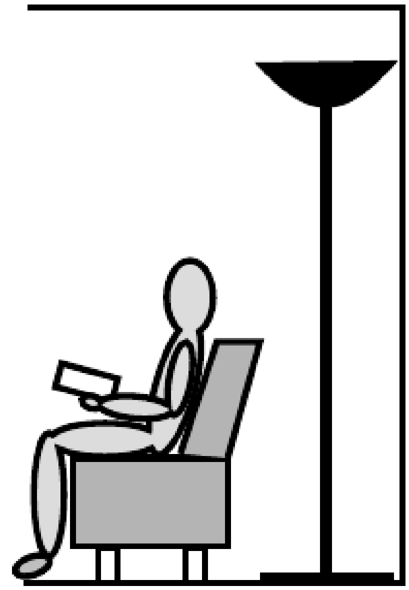
This is where the shortcomings of the Greek theory of vision become glaringly obvious. In the Greek theory, the light from the bulb and my mysterious "eye rays" are both supposed to go to the book, where they collide, allowing me to see the book. But we now have a total of four objects: lamp, eye, book, and ceiling. Where does the ceiling come in? Does it also send out its own mysterious "ceiling rays," contributing to a three-way collision at the book? That would just be too bizarre to believe!

The differences among white, black, and the various shades of gray in between is a matter of what percentage of the light they absorb and what percentage they reflect. That's why light-colored clothing is more comfortable in the summer, and light-colored upholstery in a car stays cooler that dark upholstery.

Numerical measurement of the brightness of light

We have already seen that the physiological sensation of loudness relates to the sound's intensity (power per unit area), but is not directly proportional to it. If sound A has an intensity of 1 $\mathrm{nW/m^2}$, sound B is 10 $\mathrm{nW/m^2}$, and sound C is 100 $\mathrm{nW/m^2}$, then the increase in loudness from B to C is perceived to be the same as the increase from A to B, not ten times greater. That is, the sensation of loudness is logarithmic.

The same is true for the brightness of light. Brightness is related to power per unit area, but the psychological relationship is a logarithmic one rather than a proportionality. For doing physics, it's the power per unit area that we're interested in. The relevant unit is $\mathrm{W/m^2}$. One way to determine the brightness of light is to measure the increase in temperature of a black object exposed to the light. The light energy is being converted to heat energy, and the amount of heat energy absorbed in a given amount of time can be related to the power absorbed, using the known heat capacity of the object. More practical devices for measuring light intensity, such as the light meters built into some cameras, are based on the conversion of light into electrical energy, but these meters have to be calibrated somehow against heat measurements.



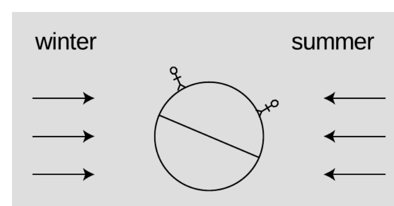
f / Light bounces off of the ceiling, then off of the book.

Discussion questions

A The curtains in a room are drawn, but a small gap lets light through, illuminating a spot on the floor. It may or may not also be possible to see the beam of sunshine crossing the room, depending on the conditions. What's going on?

B Laser beams are made of light. In science fiction movies, laser beams are often shown as bright lines shooting out of a laser gun on a spaceship. Why is this scientifically incorrect?

C A documentary film-maker went to Harvard's 1987 graduation ceremony and asked the graduates, on camera, to explain the cause of the seasons. Only two out of 23 were able to give a correct explanation, but you now have all the information needed to figure it out for yourself, assuming you didn't already know. The figure shows the earth in its winter and summer positions relative to the sun. Hint: Consider the units used to measure the brightness of light, and recall that the sun is lower in the sky in winter, so its rays are coming in at a shallower angle.

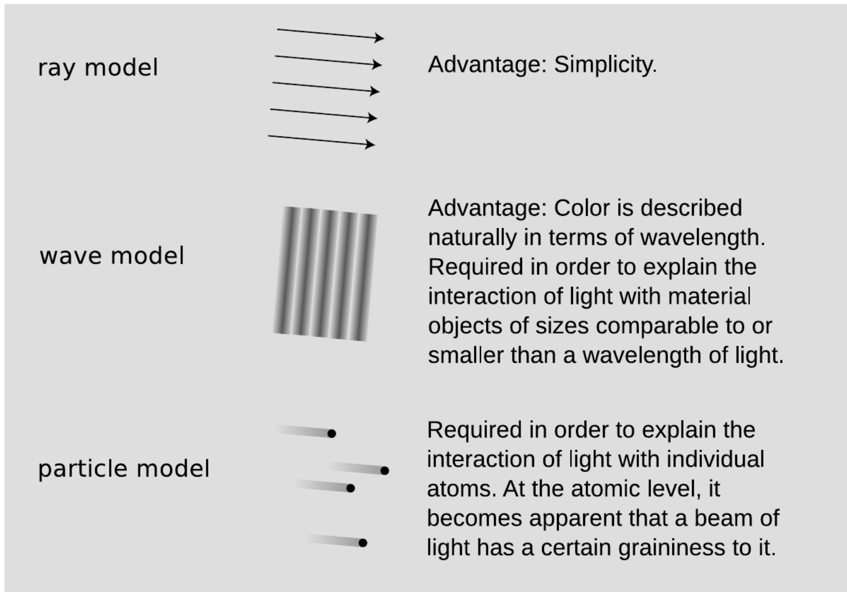


g / Discussion question C.

28.3 The ray model of light

Models of light

Note how I've been casually diagramming the motion of light with pictures showing light rays as lines on the page. More formally, this is known as the ray model of light. The ray model of light seems natural once we convince ourselves that light travels through space, and observe phenomena like sunbeams coming through holes in clouds. Having already been introduced to the concept of light as an electromagnetic wave, you know that the ray model is not the ultimate truth about light, but the ray model is simpler, and in any case science always deals with models of reality, not the ultimate nature of reality. The following table summarizes three models of light.

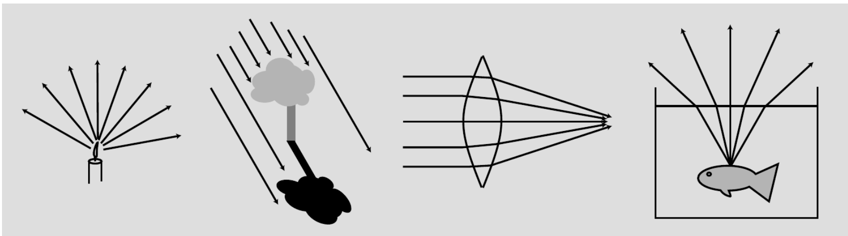


h / Three models of light.

The ray model is a generic one. By using it we can discuss the path taken by the light, without committing ourselves to any specific description of what it is that is moving along that path. We will use the nice simple ray model for most of our treatment of optics, and with it we can analyze a great many devices and phenomena. Not until chapter 32 will we concern ourselves specifically with wave optics, although in the intervening chapters I will sometimes analyze the same phenomenon using both the ray model and the wave model.

Note that the statements about the applicability of the various models are only rough guides. For instance, wave interference effects are often detectable, if small, when light passes around an obstacle that is quite a bit bigger than a wavelength. Also, the criterion for when we need the particle model really has more to do with energy scales than distance scales, although the two turn out to be related.

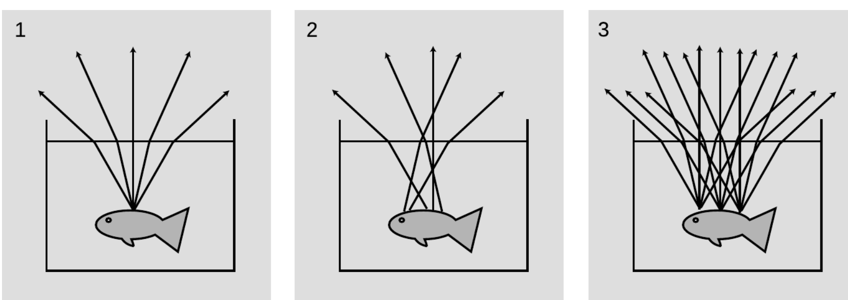
The alert reader may have noticed that the wave model is required at scales smaller than a wavelength of light (on the order of a micrometer for visible light), and the particle model is demanded on the atomic scale or lower (a typical atom being a nanometer or so in size). This implies that at the smallest scales we need *both* the wave model and the particle model. They appear incompatible, so how can we simultaneously use both? The answer is that they are not as incompatible as they seem. Light is both a wave and a particle, but a full understanding of this apparently nonsensical statement is a topic for chapter 34.



i / Examples of ray diagrams.

Ray diagrams

Without even knowing how to use the ray model to calculate anything numerically, we can learn a great deal by drawing ray diagrams. For instance, if you want to understand how eyeglasses help you to see in focus, a ray diagram is the right place to start. Many students under-utilize ray diagrams in optics and instead rely on rote memorization or plugging into formulas. The trouble with memorization and plug-ins is that they can obscure what's really going on, and it is easy to get them wrong. Often the best plan is to do a ray diagram first, then do a numerical calculation, then check that your numerical results are in reasonable agreement with what you expected from the ray diagram.



j / 1. Correct. 2. Incorrect: implies that diffuse reflection only gives one ray from each reflecting point. 3. Correct, but unnecessarily complicated

Figure j shows some guidelines for using ray diagrams effectively. The light rays bend when they pass out through the surface of the water (a phenomenon that we'll discuss in more detail later). The rays appear to have come from a point above the goldfish's actual location, an effect that is familiar to people who have tried spearfishing.

- A stream of light is not really confined to a finite number of narrow lines. We just draw it that way. In j/1, it has been necessary to choose a finite number of rays to draw (five), rather than the theoretically infinite number of rays that will diverge from that point.

- There is a tendency to conceptualize rays incorrectly as objects. In his Optics, Newton goes out of his way to caution the reader against this, saying that some people "consider ... the refraction of ... rays to be the bending or breaking of them in their passing out of one medium into another." But a ray is a record of the path traveled by light, not a physical thing that can be bent or broken.

- In theory, rays may continue infinitely far into the past and future, but we need to draw lines of finite length. In j/1, a judicious choice has been made as to where to begin and end the rays. There is no point in continuing the rays any farther than shown, because nothing new and exciting is going to happen to them. There is also no good reason to start them earlier, before being reflected by the fish, because the direction of the diffusely reflected rays is random anyway, and unrelated to the direction of the original, incoming ray.

- When representing diffuse reflection in a ray diagram, many students have a mental block against drawing many rays fanning out from the same point. Often, as in example j/2, the problem is the misconception that light can only be reflected in one direction from one point.

- Another difficulty associated with diffuse reflection, example j/3, is the tendency to think that in addition to drawing many rays coming out of one point, we should also be drawing many rays coming from many points. In j/1, drawing many rays coming out of one point gives useful information, telling us, for instance, that the fish can be seen from any angle. Drawing many sets of rays, as in j/3, does not give us any more useful information, and just clutters up the picture in this example. The only reason to draw sets of rays fanning out from more than one point would be if different things were happening to the different sets.

Discussion question

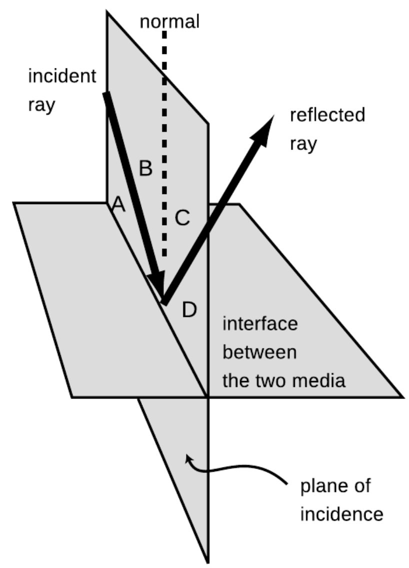
A Suppose an intelligent tool-using fish is spear-hunting for humans. Draw a ray diagram to show how the fish has to correct its aim. Note that although the rays are now passing from the air to the water, the same rules apply: the rays are closer to being perpendicular to the surface when they are in the water, and rays that hit the air-water interface at a shallow angle are bent the most.

28.4 Geometry of specular reflection

To change the motion of a material object, we use a force. Is there any way to exert a force on a beam of light? Experiments show that electric and magnetic fields do not deflect light beams, so apparently light has no electric charge. Light also has no mass, so until the twentieth century it was believed to be immune to gravity as well. Einstein predicted that light beams would be very slightly deflected by strong gravitational fields, and he was proved correct by observations of rays of starlight that came close to the sun, but obviously that's not what makes mirrors and lenses work!

If we investigate how light is reflected by a mirror, we will find that the process is horrifically complex, but the final result is surprisingly simple. What actually happens is that the light is made of electric and magnetic fields, and these fields accelerate the electrons in the mirror. Energy from the light beam is momentarily transformed into extra kinetic energy of the electrons, but because the electrons are accelerating they re-radiate more light, converting their kinetic energy back into light energy. We might expect this to result in a very chaotic situation, but amazingly enough, the electrons move together to produce a new, reflected beam of light, which obeys two simple rules:

- The angle of the reflected ray is the same as that of the incident ray.
- The reflected ray lies in the plane containing the incident ray and the normal (perpendicular) line. This plane is known as the plane of incidence.



k / The geometry of specular reflection.

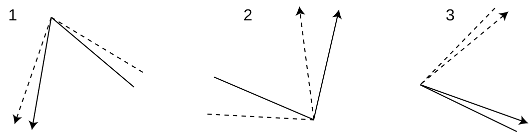
The two angles can be defined either with respect to the normal, like angles B and C in the figure, or with respect to the reflecting surface, like angles A and D. There is a convention of several hundred years' standing that one measures the angles with respect to the normal, but the rule about equal angles can logically be stated either as B=C or as A=D.

The phenomenon of reflection occurs only at the boundary between two media, just like the change in the speed of light that passes from one medium to another. As we have seen in chapter 20, this is the way all waves behave.

Most people are surprised by the fact that light can be reflected back from a less dense medium. For instance, if you are diving and you look up at the surface of the water, you will see a reflection of yourself.

self-check A

Each of these diagrams is supposed to show two different rays being reflected from the same point on the same mirror. Which are correct, and which are incorrect?



▷ Answer, p. 1045

Reversibility of light rays

The fact that specular reflection displays equal angles of incidence and reflection means that there is a symmetry: if the ray had come in from the right instead of the left in the figure above, the angles would have looked exactly the same. This is not just a pointless detail about specular reflection. It's a manifestation of a very deep and important fact about nature, which is that the laws of physics do not distinguish between past and future. Cannonballs and planets have trajectories that are equally natural in reverse, and so do light rays. This type of symmetry is called time-reversal symmetry.

Typically, time-reversal symmetry is a characteristic of any process that does not involve heat. For instance, the planets do not experience any friction as they travel through empty space, so there is no frictional heating. We should thus expect the time-reversed versions of their orbits to obey the laws of physics, which they do. In contrast, a book sliding across a table does generate heat from friction as it slows down, and it is therefore not surprising that this type of motion does not appear to obey time-reversal symmetry. A book lying still on a flat table is never observed to spontaneously start sliding, sucking up heat energy and transforming it into kinetic energy.

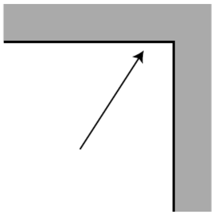
Similarly, the only situation we've observed so far where light does not obey time-reversal symmetry is absorption, which involves heat. Your skin absorbs visible light from the sun and heats up, but we never observe people's skin to glow, converting heat energy into visible light. People's skin does glow in infrared light, but that doesn't mean the situation is symmetric. Even if you absorb infrared, you don't emit visible light, because your skin isn't hot enough to glow in the visible spectrum.

These apparent heat-related asymmetries are not actual asymmetries in the laws of physics. The interested reader may wish to learn more about this from optional chapter 16 on thermodynamics.

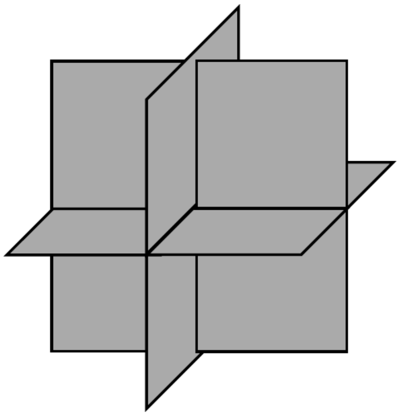
Ray tracing on a computer *example 1*

A number of techniques can be used for creating artificial visual scenes in computer graphics. Figure l shows such a scene, which was created by the brute-force technique of simply constructing a very detailed ray diagram on a computer. This technique requires a great deal of computation, and is therefore too slow to be used for video games and computer-animated movies. One trick for speeding up the computation is to exploit the reversibility of light rays. If one was to trace every ray emitted by every illuminated surface, only a tiny fraction of those would actually end up passing into the virtual "camera," and therefore almost all of the computational effort would be wasted. One can instead start a ray at the camera, trace it backward in time, and see where it would have come from. With this technique, there is no wasted effort.

l / This photorealistic image of a nonexistent countertop was produced completely on a computer, by computing a complicated ray diagram.



m / Discussion question B.



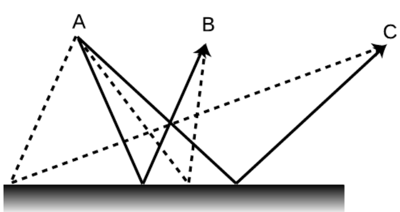
n / Discussion question C.

Discussion questions

A If a light ray has a velocity vector with components c_x and c_y, what will happen when it is reflected from a surface that lies along the y axis? Make sure your answer does not imply a change in the ray's speed.

B Generalizing your reasoning from discussion question A, what will happen to the velocity components of a light ray that hits a corner, as shown in the figure, and undergoes two reflections?

C Three pieces of sheet metal arranged perpendicularly as shown in the figure form what is known as a radar corner. Let's assume that the radar corner is large compared to the wavelength of the radar waves, so that the ray model makes sense. If the radar corner is bathed in radar rays, at least some of them will undergo three reflections. Making a further generalization of your reasoning from the two preceding discussion questions, what will happen to the three velocity components of such a ray? What would the radar corner be useful for?



o / The solid lines are physically possible paths for light rays traveling from A to B and from A to C. They obey the principle of least time. The dashed lines do not obey the principle of least time, and are not physically possible.

28.5 ⋆ The principle of least time for reflection

We had to choose between an unwieldy explanation of reflection at the atomic level and a simpler geometric description that was not as fundamental. There is a third approach to describing the interaction of light and matter which is very deep and beautiful. Emphasized by the twentieth-century physicist Richard Feynman, it is called the principle of least time, or Fermat's principle.

Let's start with the motion of light that is not interacting with matter at all. In a vacuum, a light ray moves in a straight line. This can be rephrased as follows: of all the conceivable paths light could follow from P to Q, the only one that is physically possible is the path that takes the least time.

What about reflection? If light is going to go from one point to another, being reflected on the way, the quickest path is indeed the one with equal angles of incidence and reflection. If the starting and ending points are equally far from the reflecting surface, o, it's not hard to convince yourself that this is true, just based on symmetry. There is also a tricky and simple proof, shown in figure p, for the more general case where the points are at different distances from the surface.