3

Newton versus Einstein

On cosmological scales (that is, on scales greater than 100 Mpc or so), the dominant force determining the evolution of the universe is gravity. The weak and strong nuclear forces are short-range forces; the weak force is effective only on scales of $\ell_w \sim 10^{-18}$ m or less, and the strong force on scales of $\ell_s \sim 10^{-15}$ m or less. Both gravity and electromagnetism are long-range forces. On small scales, gravity is negligibly small compared to electromagnetic forces; for instance, the electrostatic repulsion between a pair of protons is larger by a factor $\sim 10^{36}$ than the gravitational attraction between them. However, on large scales, the universe is electrically neutral, so there are no electrostatic forces on large scales. Moreover, intergalactic magnetic fields are sufficiently small that magnetic forces are also negligibly tiny on cosmological scales.

In referring to gravity as a force, we are implicitly adopting a Newtonian viewpoint. In physics, the two useful ways of looking at gravity are the Newtonian (classical) viewpoint and the Einsteinian (general relativistic) viewpoint. In Isaac Newton's view, as formulated by his laws of motion and law of gravity, gravity is a force that causes massive bodies to be accelerated. By contrast, in Einstein's view, gravity is a manifestation of the curvature of spacetime. Although Newton's view and Einstein's view are conceptually very different, in most contexts they yield the same predictions. The Newtonian predictions differ significantly from the predictions of general relativity only in the limit of deep potential minima (to use Newtonian language) or strong spatial curvature (to use general relativistic language). In these limits, general relativity yields the correct result.

In the limit of shallow potential minima and weak spatial curvature, it is permissible to switch back and forth between a Newtonian and a general relativistic viewpoint, adopting whichever one is more convenient. I will frequently adopt the Newtonian view of gravity in this book because, in many contexts, it is mathematically simpler and conceptually more familiar. The question of *why* it is possible to switch back and forth between the two very different viewpoints of Newton and Einstein is an intriguing one, and deserves closer investigation.

3.1 The Way of Newton

In Newton's view of the universe, space is unchanging and Euclidean. In Euclidean space, all the axioms and theorems of plane geometry, as codified by Euclid in the third century BC, hold true. (Euclidean space is also referred to as "flat" space. In this context, "flat" doesn't mean two-dimensional, like a piece of paper; you can have three-dimensional flat spaces as well as two-dimensional flat spaces.) In Euclidean space, the shortest distance between two points is a straight line, the angles at the vertices of a triangle sum to π radians, the circumference of a circle is 2π times its radius, and so on, through all the other axioms and theorems you learned in high school geometry. In Newton's view, moreover, an object with no net force acting on it moves in a straight line at constant speed. However, when we look at objects in the Solar System such as planets, moons, comets, and asteroids, we find that they move on curved lines, with constantly changing speed. Why is this? Newton would tell us, "Their velocities are changing because there is a force acting on them; the force called *gravity*."

Newton devised a formula for computing the gravitational force between two objects. Every object in the universe, said Newton, has a property that we may call the "gravitational mass." Let the gravitational masses of two objects be M_g and m_g, and let the distance between their centers be r. The gravitational force acting between the two objects (assuming they are both spherical) is

$$F = -\frac{GM_g m_g}{r^2}. \tag{3.1}$$

The negative sign in the above equation indicates that gravity, in the Newtonian view, is always an attractive force, tending to draw two bodies closer together.

What is the acceleration that results from this gravitational force? Newton had something to say about that as well. Every object in the universe, said Newton, has a property that we may call the "inertial mass." Let the inertial mass of an object be m_i. Newton's second law of motion says that force and acceleration are related by the equation

$$F = m_i a. \tag{3.2}$$

In Equations 3.1 and 3.2 we have distinguished, through the use of different subscripts, between the gravitational mass m_g and the inertial mass m_i. One of the fundamental principles of physics is that the gravitational mass and the inertial mass of an object are identical:

$$m_g = m_i. \tag{3.3}$$

When you stop to think about it, this equality is a remarkable fact. The property of an object that determines how strongly it is pulled on by the force of gravity is equal to the property that determines its resistance to acceleration by *any* force, not just the force of gravity. The equality of gravitational mass and inertial mass is called the *equivalence principle*.

If the equivalence principle did not hold, then the gravitational acceleration of an object toward a mass M_g would be (combining Equations 3.1 and 3.2)

$$a = -\frac{GM_g}{r^2}\left(\frac{m_g}{m_i}\right), \tag{3.4}$$

with the ratio m_g/m_i varying from object to object. However, when Galileo dropped objects from towers and slid objects down inclined planes, he found that the acceleration (barring the effects of air resistance and friction) was always the same, regardless of the mass and composition of the object. The magnitude of the gravitational acceleration close to the Earth's surface is $a = GM_{\text{Earth}}/r_{\text{Earth}}^2 = 9.8\,\text{m}\,\text{s}^{-2}$. Modern tests of the equivalence principle, which are basically more sensitive versions of Galileo's experiments, reveal that the inertial and gravitational masses are the same to within one part in 10^{13}. For the rest of this book, therefore, we'll just use the symbol m for mass, where $m = m_i = m_g$.

The equivalence principle implies that at every point $\vec{r}$ in the universe there is a unique gravitational acceleration $\vec{a}(\vec{r})$. It is useful to compute this acceleration in terms of a gravitational potential $\Phi(\vec{r})$. If the mass density of the universe is $\rho(\vec{r})$, then the gravitational potential is given by Poisson's equation:

$$\nabla^2\Phi = 4\pi G\rho. \tag{3.5}$$

If we start with a known density distribution ρ and want to find the associated potential Φ, it is more useful to use Poisson's equation in its integral form:

$$\Phi(\vec{r}) = -G\int \frac{\rho(\vec{x})}{|\vec{x}-\vec{r}|}d^3x. \tag{3.6}$$

The gravitational acceleration is then $\vec{a} = -\vec{\nabla}\Phi$.

3.2 The Special Way of Einstein

After the publication of Newton's *Principia Mathematica* in 1687, the immense power of Newtonian physics became apparent to Newton's contemporaries. As Alexander Pope wrote shortly after Newton's death:

Nature and Nature's law lay hid in night.
God said *Let Newton be!* and all was light.

Two centuries later, however, the poet John Collings Squire was able to write:

It did not last: the Devil howling *Ho!*
Let Einstein be! restored the status quo.

In popular culture, Newton's laws were regarded as rational and comprehensible; Einstein's theories were regarded as esoteric and incomprehensible. In fact, the

theory of special relativity (as first published by Einstein in 1905) is mathematically rather simple. It's only when we turn to general relativity (as published by Einstein in 1915) that the mathematics becomes more complicated. Let's start, as a warmup exercise, by considering special relativity.

Special relativity deals with the *special* case in which gravity is not present. In the absence of gravity, space is Euclidean, just as in Newtonian theory. Suppose we place a particle of mass m in three-dimensional Euclidean space. It is straightforward to measure the particle's coordinates (x, y, z) relative to a set of cartesian coordinate axes, which provide a *reference frame* for measuring positions, velocities, and accelerations. The reference frame is *inertial* if the motion of a particle, with speed $v \ll c$ relative to the reference frame, obeys Newton's second law of motion,

$$\frac{d^2\vec{r}}{dt^2} = \frac{1}{m}\vec{F}, \tag{3.7}$$

when the acceleration is measured relative to the reference frame. A rotating reference frame, for example, is *not* an inertial frame, since the equation of motion in a rotating frame contains a Coriolis term and a centrifugal term. Whether or not a reference frame is inertial can be determined empirically. Take a particle, apply a known force to it, and measure whether its acceleration is equal to that predicted by Newton's second law. (The necessary caution is that your test is limited by the precision and accuracy with which you can measure accelerations. Newton, after all, devised his second law after performing experiments in which accelerations were measured relative to a frame of reference attached to the rotating Earth. The resulting Coriolis and centrifugal terms, however, were too small for Newton to measure.)

Suppose you've taken out your accelerometer and have satisfied yourself that your cartesian reference frame is inertial. Now consider a second reference frame, moving relative to the first at a constant speed v in the $+x$ direction, as shown in Figure 3.1. If the first reference frame (let's call it the "unprimed" frame) is

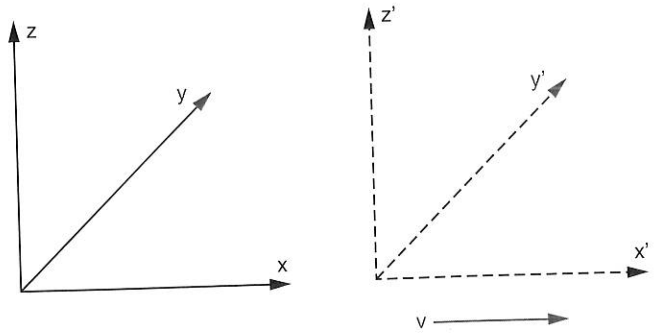

Figure 3.1 A pair of inertial reference frames (unprimed and primed), moving at a constant relative velocity $\vec{v}$.

inertial, the second reference frame (the "primed" frame) is inertial as well, as long as the relative velocity $\vec{v}$ of the two frames is constant.

In Newtonian physics, time is independent of the reference frame in which it is measured. As Newton himself put it, "Absolute, true, and mathematical time, of itself, and from its own nature, flows equably without relation to anything external." If the origins of the unprimed and primed reference frames coincide at some time $t = t' = 0$, then at some other time $t = t' \neq 0$, the coordinates in the two frames are related, in Newtonian physics, by the *Galilean transformation*:

$$\begin{aligned} x' &= x - vt \\ y' &= y \\ z' &= z \\ t' &= t. \end{aligned} \tag{3.8}$$

The Galilean transformation implies that a particle that has a velocity $\vec{u}$ measured relative to the unprimed frame has a velocity $\vec{u}' = \vec{u} - v\hat{e}_x$ relative to the primed frame.

Newtonian physics and the Galilean transformation were seriously questioned by Einstein at the beginning of the 20th century. Einstein's *first postulate of special relativity* is:

1st: The equations describing the basic laws of physics are the same in all inertial frames of reference.

Einstein's first postulate, on its surface, doesn't seem very radical. It's just an extension of what Galileo said in the 17th century, even before the birth of Newton. Galileo pointed out that if you were below decks in a sailing ship with no portholes, there would be no experiment you could conduct that would enable you to tell whether you were anchored on a placid sea or sailing along at a constant velocity. Einstein's key realization, though, was that Maxwell's equations, as well as Newton's laws of motion, are unchanged in a switch between inertial reference frames. Maxwell's equations, which describe the behavior of electric and magnetic fields, imply the existence of electromagnetic waves traveling through a vacuum at speed c. If Maxwell's equations are identical in all inertial frames of reference, as Einstein assumed, then electromagnetic waves must travel with the identical speed c in all inertial frames. This realization led Einstein to what is sometimes called the *second postulate of special relativity*:

2nd: The speed of light in a vacuum has the same value c in all inertial frames of reference.

The constancy of the speed of light had been demonstrated by Michelson and Morley as early as 1887 (although it is unclear whether Einstein was aware of their results when he published the theory of special relativity in 1905).

Let's return to the unprimed and primed frames of reference shown in Figure 3.1. At the instant when the origins of the two frames coincide, we synchronize the clocks associated with the frames, so that $t = t' = 0$. We celebrate the synchronization by having a lamp located at the joint origin emit a brief flash of light. If space is empty, then a spherical shell of light expands outward with speed c, regardless of the frame in which it is observed. At a later time $t > 0$, the equation giving the size of the shell in the unprimed frame is

$$c^2t^2 = x^2 + y^2 + z^2. \tag{3.9}$$

At the corresponding time t' in the primed frame,

$$c^2(t')^2 = (x')^2 + (y')^2 + (z')^2. \tag{3.10}$$

Equations 3.9 and 3.10 are incompatible with the Galilean transformation, as you can verify by substitution from Equation 3.8.

Equations 3.9 and 3.10 are, however, compatible with the *Lorentz transformation*:[1]

$$\begin{aligned} x' &= \gamma(x - vt) \\ y' &= y \\ z' &= z \\ t' &= \gamma(t - vx/c^2), \end{aligned} \tag{3.11}$$

where γ is the *Lorentz factor*,

$$\gamma \equiv \frac{1}{\sqrt{1 - v^2/c^2}}. \tag{3.12}$$

In special relativity, the Lorentz transformation is the correct way to convert between coordinates in two inertial frames of reference.

To see how the Lorentz transformation disrupts Newtonian ideas about space and time, consider two events. In the unprimed frame, event 1 occurs at time t_1 at location (x_1, y_1, z_1); event 2 occurs at time t_2 at location (x_2, y_2, z_2). Since space is Euclidean in special relativity, we can easily compute the spatial distance between the two events in the unprimed frame,

$$(\Delta\ell)^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2. \tag{3.13}$$

The time elapsed between the two events in the unprimed frame is

$$\Delta t = t_1 - t_2. \tag{3.14}$$

We can use the Lorentz transformation to compute the spatial distance between the two events measured in the primed frame,

[1] The Lorentz transformation was first published by Joseph Larmor in 1897; Hendrik Lorentz didn't independently find the Lorentz transformation until 1899. (The law of misonomy strikes again.)

$$(\Delta\ell')^2 = (x_1' - x_2')^2 + (y_1' - y_2')^2 + (z_1' - z_2')^2 \qquad (3.15)$$
$$= \gamma^2 [x_1 - x_2 - v(t_1 - t_2)]^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2.$$

The time elapsed between the two events in the primed frame is

$$\Delta t' = t_1' - t_2' = \gamma \left[t_1 - t_2 - \frac{v}{c^2}(x_1 - x_2) \right]. \qquad (3.16)$$

Observers in the primed and unprimed frames will measure different spatial distances between the two events. They will also measure different time intervals between the two events; under some circumstances, they will even disagree on which event occurred first. Contrary to Newton's thinking, special relativity tells us that there is no "absolute time." Observers in different reference frames will measure time differently.

Although observers in different inertial reference frames will disagree on the spatial distance between two events, and also on the time interval between the events, there is something that they will agree on: the *spacetime* separation between the events. In the unprimed frame, the spacetime separation between event 1 and event 2 is

$$(\Delta s)^2 = -c^2(t_1 - t_2)^2 + (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \qquad (3.17)$$

or

$$(\Delta s)^2 = -c^2(\Delta t)^2 + (\Delta \ell)^2. \qquad (3.18)$$

Notice the choice of signs in this relation: two events have a spacetime separation $\Delta s = 0$ if the light travel time between their spatial locations, $\Delta\ell/c$, is equal to the time that elapses between the events, $|\Delta t|$.

The spacetime separation in the primed frame is

$$(\Delta s')^2 = -c^2(\Delta t')^2 + (\Delta \ell')^2, \qquad (3.19)$$

where $\Delta\ell'$ is given by Equation 3.15 and $\Delta t'$ is given by Equation 3.16. Making the substitutions into Equation 3.19, we find

$$(\Delta s')^2 = -\gamma^2 \left[c(t_1 - t_2)^2 - \frac{v}{c}(x_1 - x_2)^2 \right]^2$$
$$+ \gamma^2 [x_1 - x_2 - v(t_1 - t_2)]^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2. \qquad (3.20)$$

A little algebraic simplification reveals that

$$(\Delta s')^2 = -c^2(t_1 - t_2)^2 + (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2, \qquad (3.21)$$

and therefore, comparing Equations 3.17 and 3.21, that $(\Delta s)^2 = (\Delta s')^2$.

Using the Galilean transformation, the separation in *time* between two events is the same in all inertial frames of reference. Using the Lorentz transformation, the separation in *spacetime* is the same in all inertial frames. In Newtonian physics, it makes sense to think about space and time as two separate entities;

however, in special relativity, it is more useful to think about a four-dimensional spacetime, with the four-dimensional separation Δs between two events being given by Equation 3.17.

3.3 The General Way of Einstein

The theory of special relativity has limited usefulness, since it deals only with the case in which gravity is non-existent. It took Einstein a decade, from 1905 to 1915, to generalize his theory. To see how Einstein was inspired by the equivalence principle to devise his theory of general relativity, let's begin with a thought experiment. Suppose you wake up one morning to find that you have been sealed up (bed and all) within an opaque, soundproof, hermetically sealed box. "Oh no!" you say. "This is what I've always feared would happen. I've been abducted by space aliens who are taking me away to their home planet." Startled by this realization, you drop your teddy bear. Observing the bear, you find that it falls toward the floor of the box with an acceleration $a = 9.8\,\mathrm{m\,s^{-2}}$. "Whew!" you say, with some relief. "At least I am still on the Earth's surface; they haven't taken me away in their spaceship yet." At that moment, a window in the side of the box opens to reveal (much to your horror) that you are inside an alien spaceship that is being accelerated at $a = 9.8\,\mathrm{m\,s^{-2}}$ by a rocket engine.

When you drop a teddy bear, or any other object, within a sealed box (Figure 3.2), the equivalence principle permits two possible interpretations, with no way of distinguishing between them:

(1) The box is static, or moving with a constant velocity, and the bear is being accelerated downward by a gravitational force.

(2) The bear is static, or moving at a constant velocity, and the box is being accelerated upward by a non-gravitational force.

The behavior of the bear in each case is identical. In each case, a big bear falls at the same rate as a little bear; in each case, a bear stuffed with cotton falls at the same rate as a bear stuffed with lead; and in each case, a sentient anglophone bear would say, "Oh, bother. I'm weightless," during the interval before it collides with the floor of the box.

Einstein's insight, starting from the equivalence principle, led him to the theory of general relativity. To understand Einstein's thought processes, imagine yourself back in the sealed box, being accelerated through interplanetary space at $9.8\,\mathrm{m\,s^{-2}}$. You grab the flashlight that you keep on the bedside table and shine a beam of light perpendicular to the acceleration vector (Figure 3.3). Since the box is accelerating upward, the path of the light beam will appear to you to be bent downward, as the floor of the box rushes up to meet the photons. However, thanks to the equivalence principle, we can replace the accelerated box with a

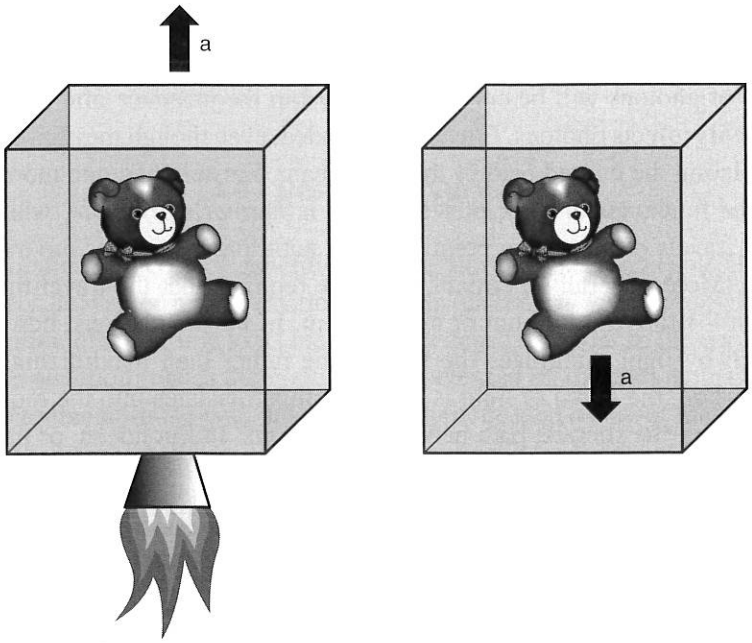

Figure 3.2 Equivalence principle (teddy bear version). The behavior of a bear in an accelerated box (left) is identical to that of a bear being accelerated by gravity (right).

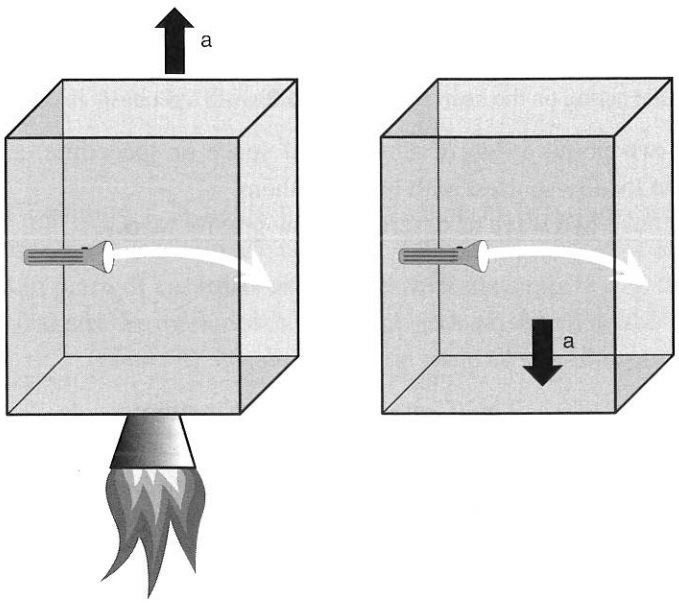

Figure 3.3 Equivalence principle (photon version). The path followed by a light beam in an accelerated box (left) is identical to the path followed by a light beam being accelerated by gravity (right). [The deflection shown is greatly exaggerated for the sake of visualization. The actual deflection will be $\sim 2 \times 10^{-14}$ m if the box is 2 meters across.]

stationary box experiencing a constant gravitational acceleration. Since there's no way to distinguish between these two cases, we are led to the conclusion that the paths of photons will be curved downward in the presence of a gravitational field. Gravity affects photons, Einstein concluded, even though they have no mass. Contemplating the curved path of the light beam, Einstein had one more insight. One of the fundamental principles of optics is *Fermat's principle*, which states that light travels between two points along a path that minimizes the travel time required. (More generally, Fermat's principle requires that the travel time be an extremum – either a minimum or a maximum. In most situations, however, the path taken by light minimizes the travel time rather than maximizing it.) In a vacuum, where the speed of light is constant, this translates into the requirement that light takes the shortest path between two points. In Euclidean, or flat, space, the shortest path between two points is a straight line. However, in the presence of gravity, the path taken by light is not a straight line. Thus, Einstein concluded, space is *not* Euclidean.

The presence of mass, in Einstein's view, causes space to be curved. In fact, in the fully developed theory of general relativity, mass and energy (which Newton thought of as two separate entities) are interchangeable, via the famous equation $E = mc^2$. Moreover, space and time (which Newton thought of as two separate entities) form a four-dimensional spacetime. A more accurate summary of Einstein's viewpoint, therefore, is that the presence of mass-energy causes spacetime to be curved. We now have a third way of thinking about the motion of the teddy bear in the box:

(3) No forces are acting on the bear; it is simply following a *geodesic* in curved spacetime.

If you take two points in an N-dimensional space or spacetime, a geodesic is defined as the locally shortest path between them.

We now have two ways of describing how gravity works.

The Way of Newton:

Mass tells gravity how to exert a force ($F = -GMm/r^2$),
Force tells mass how to accelerate ($F = ma$).

The (General) Way of Einstein:

Mass-energy tells spacetime how to curve,
Curved spacetime tells mass-energy how to move.[2]

Einstein's description of gravity gives a natural explanation for the equivalence principle. In the Newtonian description of gravity, the equality of the gravitational mass and the inertial mass is a remarkable coincidence. However, in Einstein's theory of general relativity, curvature is a property of spacetime itself.

[2] This pocket summary of general relativity was coined by the physicist John Wheeler, who also popularized the term "black hole."

It then follows automatically that the gravitational acceleration of an object should be independent of mass and composition – it's just following a geodesic, which is dictated by the geometry of spacetime.

3.4 Describing Curvature

In developing his theory of general relativity, Einstein faced multiple challenges. Ultimately, he wanted a mathematical formula (called a *field equation*) that relates the curvature of spacetime to its mass-energy density, similar to the way in which Poisson's equation relates the gravitational potential of space to its mass density. En route to this ultimate goal, however, Einstein needed a way of mathematically describing curvature. Since picturing the curvature of a four-dimensional spacetime is difficult, let's start by considering ways of describing the curvature of two-dimensional spaces, and then extend what we have learned to higher dimensions.

The simplest of two-dimensional spaces is a plane, as illustrated in Figure 3.4, for which Euclidean geometry holds true. On a plane, a geodesic is a straight line. If a triangle is constructed on a plane by connecting three points with geodesics, the angles at its vertices (α, β, and γ in Figure 3.4) obey the relation

$$\alpha + \beta + \gamma = \pi, \tag{3.22}$$

where angles are measured in radians. On a plane, we can set up a cartesian coordinate system, and assign to every point a coordinate (x, y). On a plane, the Pythagorean theorem holds, so the distance $d\ell$ between points (x, y) and $(x + dx, y + dy)$ is given by the relation[3]

$$d\ell^2 = dx^2 + dy^2. \tag{3.23}$$

Stating that Equation 3.23 holds true everywhere in two-dimensional space is equivalent to saying that the space is a plane. Of course, other coordinate systems can be used in place of cartesian coordinates. For instance, in a polar coordinate system, the distance between points (r, θ) and $(r + dr, \theta + d\theta)$ is

$$d\ell^2 = dr^2 + r^2 d\theta^2. \tag{3.24}$$

Although Equations 3.23 and 3.24 are different in appearance, they both represent the same flat geometry, as you can verify by making the simple coordinate substitution $x = r\cos\theta$, $y = r\sin\theta$.

Now consider another simple two-dimensional space, the surface of a sphere (Figure 3.5). On the surface of a sphere, a geodesic is a portion of a great circle; that is, a circle whose center corresponds to the center of the sphere. If a triangle is

[3] Starting with this equation, I adopt the convention, commonly used among relativists, that $d\ell^2 = (d\ell)^2$, and not $d(\ell^2)$. Omitting the parentheses makes the equations less cluttered.

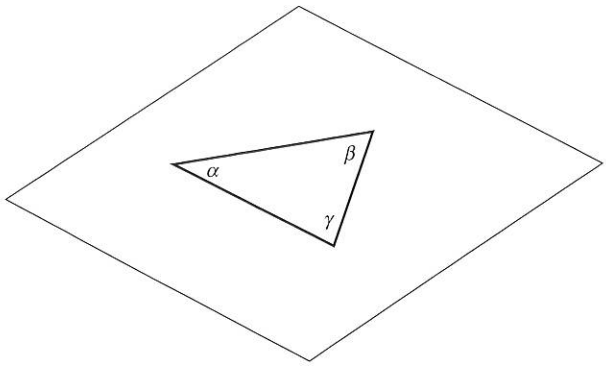

Figure 3.4 A Euclidean, or flat, two-dimensional space.

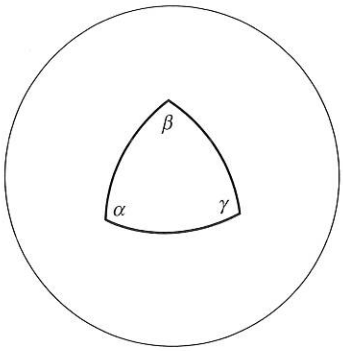

Figure 3.5 A positively curved two-dimensional space.

constructed on the surface of a sphere by connecting three points with geodesics, the angles at its vertices (α, β, and γ) obey the relation

$$\alpha + \beta + \gamma = \pi + A/R^2, \tag{3.25}$$

where A is the area of the triangle, and R is the radius of the sphere. All spaces in which $\alpha + \beta + \gamma > \pi$ are called positively curved spaces. The surface of a sphere is a special variety of positively curved space; it has curvature that is both homogeneous and isotropic. That is, no matter where you draw a triangle on the surface of a sphere, or how you orient it, it must always satisfy Equation 3.25, with the radius R being the same everywhere and in all directions. For brevity, we can describe a space where the curvature is homogeneous and isotropic as having "uniform curvature." Thus, the surface of a sphere can be described as a two-dimensional space with uniform positive curvature.

On the surface of a sphere, we can set up a polar coordinate system by picking a pair of antipodal points to be the "north pole" and "south pole" and by picking a geodesic from the north to the south pole to be the "prime meridian." If r is the distance from the north pole, and θ is the azimuthal angle measured relative to the

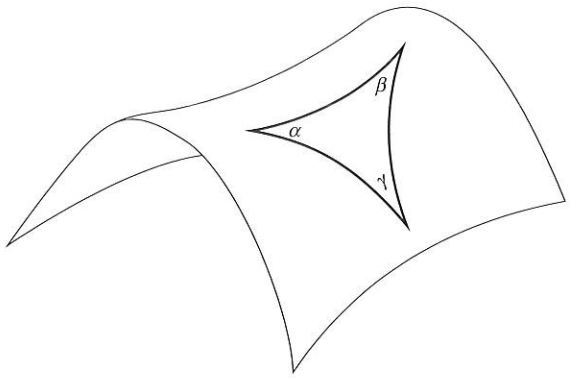

Figure 3.6 A negatively curved two-dimensional space.

prime meridian, then the distance $d\ell$ between a point (r, θ) and another nearby point $(r + dr, \theta + d\theta)$ is given by the relation

$$d\ell^2 = dr^2 + R^2 \sin^2(r/R) d\theta^2. \tag{3.26}$$

Note that the surface of a sphere has a finite area, equal to $4\pi R^2$, and a maximum possible distance between points. (In a non-Euclidean space, the distance between two points is defined as the length of the geodesic connecting them.) The distance between antipodal points, at the maximum possible separation, is $\ell_{\max} = \pi R$. By contrast, a plane has infinite area, and has no upper limit on the possible distance between points.[4]

In addition to flat spaces and positively curved spaces, there exist negatively curved spaces. An example of a negatively curved two-dimensional space is the hyperboloid, or saddle shape, shown in Figure 3.6. For illustrative purposes, it would be useful to show you a surface of uniform negative curvature, just as the surface of a sphere has uniform positive curvature. Unfortunately, the mathematician David Hilbert proved that a two-dimensional surface of uniform negative curvature cannot be constructed in a three-dimensional Euclidean space. The saddle shape illustrated in Figure 3.6 has uniform curvature only in the central region, near the "seat" of the saddle.

Despite the difficulties in visualizing a surface of uniform negative curvature, its properties can be written down easily. Consider a two-dimensional surface of uniform negative curvature, with radius of curvature R. If a triangle is constructed on this surface by connecting three points with geodesics, the angles at its vertices (α, β, and γ) obey the relation

$$\alpha + \beta + \gamma = \pi - A/R^2, \tag{3.27}$$

where A is the area of the triangle.

[4] Since the Syndicate of Cambridge University Press objected to producing a book of infinite size, Figure 3.4 actually shows only a portion of a plane.

On a surface of uniform negative curvature, we can set up a polar coordinate system by choosing some point as the pole, and some geodesic leading away from the pole as the prime meridian. If r is the distance from the pole, and θ is the azimuthal angle measured relative to the prime meridian, then the distance $d\ell$ between a point (r, θ) and a nearby point $(r + dr, \theta + d\theta)$ is given by

$$d\ell^2 = dr^2 + R^2 \sinh^2(r/R) d\theta^2. \tag{3.28}$$

A surface of uniform negative curvature has infinite area, and has no upper limit on the possible distance between points.

Relations like those presented in Equations 3.24, 3.26, and 3.28, which give the distance $d\ell$ between two nearby points in space, are known as *metrics*. In general, curvature is a local property. A rubber tablecloth can be badly rumpled at one end of the table and smooth at the other end; a bagel (or other toroidal object) is negatively curved on part of its surface and positively curved on other portions.[5] However, if you want a two-dimensional space to be homogeneous and isotropic, only three possibilities can fit the bill: the space can be uniformly *flat*; it can have uniform *positive* curvature; or it can have uniform *negative* curvature. Thus, if a two-dimensional space has curvature that is homogeneous and isotropic, its geometry can be specified by two quantities, κ, and R. The number κ, called the *curvature constant*, is $\kappa = 0$ for a flat space, $\kappa = +1$ for a positively curved space, and $\kappa = -1$ for a negatively curved space. If the space is curved, then the quantity R, which has dimensions of length, is the radius of curvature.

The results for two-dimensional space can be extended straightforwardly to three dimensions. A three-dimensional space, if its curvature is homogeneous and isotropic, must be flat, or have uniform positive curvature, or have uniform negative curvature. If a three-dimensional space is flat ($\kappa = 0$), it has the metric

$$d\ell^2 = dx^2 + dy^2 + dz^2, \tag{3.29}$$

expressed in cartesian coordinates, or

$$d\ell^2 = dr^2 + r^2[d\theta^2 + \sin^2\theta d\phi^2], \tag{3.30}$$

expressed in spherical coordinates.

If a three-dimensional space has uniform positive curvature ($\kappa = +1$), its metric is

$$d\ell^2 = dr^2 + R^2 \sin^2(r/R)[d\theta^2 + \sin^2\theta d\phi^2]. \tag{3.31}$$

A positively curved three-dimensional space has finite volume, just as a positively curved two-dimensional space has finite area. The point at $r = \pi R$ is the antipodal point to the origin, just as the south pole is the antipodal point to the north pole

[5] You can test this assertion, if you like, by drawing triangles on a bagel.

on the surface of a sphere. By traveling a distance $C = 2\pi R$, it is possible to "circumnavigate" a space of uniform positive curvature.

Finally, if a three-dimensional space has uniform negative curvature ($\kappa = -1$), its metric is

$$d\ell^2 = dr^2 + R^2 \sinh^2(r/R)[d\theta^2 + \sin^2\theta d\phi^2]. \tag{3.32}$$

Like flat space, negatively curved space has infinite volume.

The three possible metrics for a homogeneous, isotropic, three-dimensional space can be written more compactly in the form

$$d\ell^2 = dr^2 + S_\kappa(r)^2 d\Omega^2, \tag{3.33}$$

where

$$d\Omega^2 \equiv d\theta^2 + \sin^2\theta d\phi^2 \tag{3.34}$$

and

$$S_\kappa(r) = \begin{cases} R\sin(r/R) & (\kappa = +1) \\ r & (\kappa = 0) \\ R\sinh(r/R) & (\kappa = -1). \end{cases} \tag{3.35}$$

In the limit $r \ll R$, $S_\kappa \approx r$, regardless of the value of κ. When space is flat, or negatively curved, S_κ increases monotonically with r, with $S_\kappa \to \infty$ as $r \to \infty$. By contrast, when space is positively curved, S_κ increases to a maximum of $S_{\max} = R$ at $r/R = \pi/2$, then decreases again to 0 at $r/R = \pi$, the antipodal point to the origin.

The coordinate system (r, θ, ϕ) is not the only possible system. For instance, if we switch the radial coordinate from r to $x \equiv S_\kappa(r)$, the metric for a homogeneous, isotropic, three-dimensional space can be written in the form

$$d\ell^2 = \frac{dx^2}{1 - \kappa x^2/R^2} + x^2 d\Omega^2. \tag{3.36}$$

Although the metrics written in Equations 3.33 and 3.36 appear different on the page, they represent the same homogeneous, isotropic spaces. They merely have a different functional form because of the different choice of radial coordinates.

3.5 The Robertson–Walker Metric

So far, we've considered the metrics for simple two-dimensional and three-dimensional spaces. However, relativity teaches us that space and time together constitute a four-dimensional spacetime. Just as we can compute the distance between two points in space using the appropriate metric for that space, so we can compute the four-dimensional separation between two events in spacetime.

Consider two events, one occurring at the spacetime location (t, r, θ, ϕ), and another occurring at the spacetime location $(t + dt, r + dr, \theta + d\theta, \phi + d\phi)$. According to the laws of special relativity, the spacetime separation between these two events is

$$ds^2 = -c^2dt^2 + dr^2 + r^2d\Omega^2. \tag{3.37}$$

The metric given in Equation 3.37 is called the *Minkowski metric*, and the spacetime that it describes is called Minkowski spacetime. Note, from comparison with Equation 3.33, that the spatial component of Minkowski spacetime is Euclidean, or flat.

A photon's path through spacetime is a four-dimensional geodesic – and not just any geodesic, mind you, but a special variety called a *null geodesic*. A null geodesic is one for which, along every infinitesimal segment of the photon's path, $ds = 0$. In Minkowski spacetime, then, a photon's trajectory obeys the relation

$$ds^2 = 0 = -c^2dt^2 + dr^2 + r^2d\Omega^2. \tag{3.38}$$

If the photon is moving along a radial path, toward or away from the origin, this means, since θ and ϕ are constant,

$$c^2dt^2 = dr^2, \tag{3.39}$$

or

$$\frac{dr}{dt} = \pm c. \tag{3.40}$$

The Minkowski metric of Equation 3.37 applies only within the context of special relativity. With no gravity present, Minkowski spacetime is flat and static. When gravity is added, however, the permissible spacetimes are more interesting. In the 1930s, the physicists Howard Robertson and Arthur Walker asked, "What form can the metric of spacetime assume if the universe is spatially homogeneous and isotropic at all time, and if distances are allowed to expand or contract as a function of time?" The metric they derived (independently of each other) is called the *Robertson–Walker metric*.[6] It can be written in the form

$$ds^2 = -c^2dt^2 + a(t)^2\left[dr^2 + S_\kappa(r)^2d\Omega^2\right], \tag{3.41}$$

where the function $S_\kappa(r)$ is given by Equation 3.35, with $R = R_0$. The spatial component of the Robertson–Walker metric consists of the spatial metric for a uniformly curved space of radius R_0 (compare Equation 3.33), scaled by the square of the scale factor $a(t)$. The scale factor, first introduced in Section 2.3,

[6] The Robertson–Walker metric is also called the Friedmann–Robertson–Walker (FRW) metric or the Friedmann–Lemaître–Robertson–Walker (FLRW) metric, depending on which subset of pioneering cosmologists you want to acknowledge.

describes how distances in a homogeneous, isotropic universe expand or contract with time.

The time variable t in the Robertson–Walker metric is the cosmological proper time, called the *cosmic time* for short, and is the time measured by an observer who sees the universe expanding uniformly around him or her. The spatial variables (r, θ, ϕ) are called the *comoving coordinates* of a point in space; if the expansion of the universe is perfectly homogeneous and isotropic, then the comoving coordinates of any point remain constant with time.

The assumption of homogeneity and isotropy is an extremely powerful one. If the universe is perfectly homogeneous and isotropic, then everything we need to know about its geometry is contained within the scale factor $a(t)$, the curvature constant κ (which can be $\kappa = +1$, 0, or -1), and, if $\kappa \neq 0$, the present-day radius of curvature R_0. Much of modern cosmology is devoted in one way or another to finding the values of $a(t)$, κ, and R_0. The assumption of spatial homogeneity and isotropy is so powerful that it was adopted by cosmologists such as Einstein, Friedmann, Lemaître, Robertson, and Walker long before the available observational evidence gave support for such an assumption.[7]

The Robertson–Walker metric is an approximation that holds good only on large scales; on smaller scales, the universe is lumpy, and hence does not expand uniformly. Small, dense lumps, such as humans, teddy bears, and interstellar dust grains, are held together by electromagnetic forces, and hence do not expand. Larger lumps, as long as they are sufficiently dense, are held together by their own gravity, and hence do not expand. Examples of such gravitationally bound systems are galaxies (such as the Milky Way Galaxy in which we live) and clusters of galaxies (such as the Local Group in which we live). It's only on scales larger than $\sim$ 100 Mpc that the expansion of the universe can be treated as the ideal, homogeneous, isotropic expansion described by the single scale factor $a(t)$.

3.6 Proper Distance

Consider a galaxy far away from us – sufficiently far away that we may ignore the small scale perturbations of spacetime and adopt the Robertson–Walker metric. One question we may ask is, "Exactly how far away is this galaxy?" In an expanding universe, the distance between two objects is increasing with time. Thus, if we want to assign a spatial distance d between two objects, we must specify the time t at which the distance is the correct one. Suppose that you are at the origin, and that the galaxy that you are observing is at a comoving coordinate position (r, θ, ϕ), as illustrated in Figure 3.7. The proper distance $d_p(t)$ between two points

[7] If homogeneity and isotropy did not exist, as Voltaire might have said, it would be necessary to invent them – at least if your desire is to have a simple, analytically tractable form for the metric of spacetime.

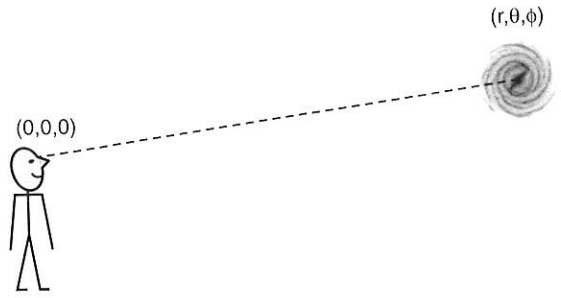

Figure 3.7 An observer at the origin observes a galaxy at coordinate position (r, θ, ϕ). A photon emitted by the galaxy at cosmic time t_e reaches the observer at cosmic time t_0.

is equal to the length of the spatial geodesic between them when the scale factor is fixed at the value $a(t)$. The proper distance between the observer and galaxy in Figure 3.7 can be found using the Robertson–Walker metric at a fixed time t:

$$ds^2 = a(t)^2[dr^2 + S_\kappa(r)^2 d\Omega^2]. \tag{3.42}$$

Along the spatial geodesic between the observer and galaxy, the angle (θ, ϕ) is constant, and thus

$$ds = a(t)dr. \tag{3.43}$$

The proper distance d_p is found by integrating over the radial comoving coordinate r:

$$d_p(t) = a(t) \int_0^r dr = a(t)r. \tag{3.44}$$

Because the proper distance has the form $d_p(t) = a(t)r$, with the comoving coordinate r constant with time, the rate of change for the proper distance between us and a distant galaxy is

$$\dot{d}_p = \dot{a}r = \frac{\dot{a}}{a} d_p. \tag{3.45}$$

Thus, at the current time ($t = t_0$), there is a linear relation between the proper distance to a galaxy and its recession speed:

$$v_p(t_0) = H_0 d_p(t_0), \tag{3.46}$$

where

$$v_p(t_0) \equiv \dot{d}_p(t_0) \tag{3.47}$$

and

$$H_0 = \left(\frac{\dot{a}}{a}\right)_{t=t_0}. \tag{3.48}$$

In a sense, this is just a repetition of what was demonstrated in Section 2.3; if the distance between points is proportional to $a(t)$, there will be a linear relation between the relative velocity of two points and the distance between them. Now, however, we are interpreting the change in distance between widely separated galaxies as being associated with the expansion of space. As the distance between galaxies increases, the radius of curvature of the universe, $R(t) = a(t)R_0$, increases at the same rate.

Some cosmology books contain a statement like "As space expands, it drags galaxies away from each other." Statements of this sort are somewhat misleading because they make galaxies appear to be entirely passive. On the other hand, a statement like "As galaxies move apart, they drag space along with them" would be equally misleading because it makes space appear to be entirely passive. As the theory of general relativity points out, spacetime and mass-energy are intimately linked. Yes, the curvature of spacetime does tell mass-energy how to move, but then it's mass-energy which tells spacetime how to curve.

The linear velocity–distance relation given in Equation 3.46 implies that points separated by a proper distance greater than the Hubble distance,

$$d_H(t_0) \equiv c/H_0, \tag{3.49}$$

will have

$$v_p = \dot{d}_p > c. \tag{3.50}$$

Using the observationally determined value of $H_0 = 68 \pm 2\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, the current value of the Hubble distance in our universe is

$$d_H(t_0) = c/H_0 = 4380 \pm 130\,\mathrm{Mpc}. \tag{3.51}$$

Thus, galaxies farther than $\sim$ 4400 megaparsecs from us are currently moving away from us at speeds greater than that of light. Cosmological innocents sometimes exclaim, "Gosh! Doesn't this violate the law that massive objects can't travel faster than the speed of light?" Actually, it doesn't. The speed limit that states that massive objects must travel with $v < c$ relative to each other is one of the results of special relativity, and refers to the relative motion of objects within a static space. In the context of general relativity, there is no objection to having two points moving away from each other at superluminal speed due to the expansion of space.

When we observe a distant galaxy, we know its angular position very well, but not its distance. That is, we can point in its direction, but we don't know its current proper distance $d_p(t_0)$. We can, however, measure the redshift z of the light we receive from the galaxy. Although the redshift doesn't tell us the proper distance to the galaxy, it does tell us what the scale factor a was at the time the light from that galaxy was emitted. To see the link between a and z, consider the galaxy illustrated in Figure 3.7. Light that was emitted by the galaxy at a time t_e

is observed by us at a time t_0. During its travel from the distant galaxy to us, the light traveled along a null geodesic, with $ds = 0$. The null geodesic has θ and ϕ constant.[8] Thus, along the light's null geodesic,

$$c^2 dt^2 = a(t)^2 dr^2. \tag{3.52}$$

Rearranging this relation, we find

$$c\frac{dt}{a(t)} = dr. \tag{3.53}$$

In Equation 3.53, the left-hand side is a function only of t, and the right-hand side is independent of t. Suppose the distant galaxy emits light with a wavelength λ_e, as measured by an observer in the emitting galaxy. Fix your attention on a single wave crest of the emitted light. The wave crest is emitted at a time t_e and observed at a time t_0, such that

$$c\int_{t_e}^{t_0} \frac{dt}{a(t)} = \int_0^r dr = r. \tag{3.54}$$

The next wave crest of light is emitted at a time $t_e + \lambda_e/c$, and is observed at a time $t_0 + \lambda_0/c$, where, in general, $\lambda_0 \neq \lambda_e$. For the second wave crest,

$$c\int_{t_e+\lambda_e/c}^{t_0+\lambda_0/c} \frac{dt}{a(t)} = \int_0^r dr = r. \tag{3.55}$$

Comparing Equations 3.54 and 3.55, we find that

$$\int_{t_e}^{t_0} \frac{dt}{a(t)} = \int_{t_e+\lambda_e/c}^{t_0+\lambda_0/c} \frac{dt}{a(t)}. \tag{3.56}$$

That is, the integral of $dt/a(t)$ between the time of emission and the time of observation is the same for every wave crest in the emitted light. If we subtract the integral

$$\int_{t_e+\lambda_e/c}^{t_0} \frac{dt}{a(t)} \tag{3.57}$$

from each side of Equation 3.56, we find the relation

$$\int_{t_e}^{t_e+\lambda_e/c} \frac{dt}{a(t)} = \int_{t_0}^{t_0+\lambda_0/c} \frac{dt}{a(t)}. \tag{3.58}$$

That is, the integral of $dt/a(t)$ between the emission of successive wave crests is equal to the integral of $dt/a(t)$ between the observation of the same two wave crests. This relation becomes still simpler when we realize that during the time between the emission or observation of two wave crests, the universe doesn't have time to expand by a significant amount. The time scale for expansion of the

[8] In a homogeneous, isotropic universe there's no reason for the light to swerve to one side or the other.

universe is the Hubble time, $H_0^{-1} \approx 14\,\text{Gyr}$. The time between wave crests, for visible light, is $\lambda/c \approx 2 \times 10^{-15}\,\text{s} \approx 10^{-32} H_0^{-1}$. Thus, $a(t)$ is effectively constant in the integrals of Equation 3.58. We may then write

$$\frac{1}{a(t_e)} \int_{t_e}^{t_e+\lambda_e/c} dt = \frac{1}{a(t_0)} \int_{t_0}^{t_0+\lambda_0/c} dt, \tag{3.59}$$

or

$$\frac{\lambda_e}{a(t_e)} = \frac{\lambda_0}{a(t_0)}. \tag{3.60}$$

Using the definition of redshift, $z = (\lambda_0 - \lambda_e)/\lambda_e$, we find that the redshift of light from a distant object is related to the expansion factor at the time it was emitted via the equation

$$1 + z = \frac{a(t_0)}{a(t_e)} = \frac{1}{a(t_e)}. \tag{3.61}$$

Here, we have used the usual convention that $a(t_0) = 1$.

Thus, if we observe a galaxy with a redshift $z = 2$, we are observing it as it was when the universe had a scale factor $a(t_e) = 1/3$. The redshift we observe for a distant object depends only on the relative scale factors at the time of emission and the time of observation. It doesn't depend on how the transition between $a(t_e)$ and $a(t_0)$ was made. It doesn't matter if the expansion was gradual or abrupt; it doesn't matter if the transition was monotonic or oscillatory. All that matters is the scale factors at the time of emission and the time of observation.

Exercises

3.1 What evidence can you provide to support the assertion that the universe is electrically neutral on large scales?

3.2 Suppose you are a two-dimensional being, living on the surface of a sphere with radius R. An object of width $d\ell \ll R$ is at a distance r from you (remember, all distances are measured on the surface of the sphere). What angular width $d\theta$ will you measure for the object? Explain the behavior of $d\theta$ as $r \to \pi R$.

3.3 Suppose you are *still* a two-dimensional being, living on the same sphere of radius R. Show that if you draw a circle of radius r, the circle's circumference will be

$$C = 2\pi R \sin(r/R). \tag{3.62}$$

Idealize the Earth as a perfect sphere of radius $R = 6371\,\text{km}$. If you could measure distances with an error of ± 1 meter, how large a circle would you have to draw on the Earth's surface to convince yourself that the Earth is spherical rather than flat?

3.4 Consider an equilateral triangle, with sides of length L, drawn on a two-dimensional surface of uniform curvature. Can you draw an equilateral triangle of arbitrarily large area A on a surface with $\kappa = +1$ and radius of curvature R? If not, what is the maximum possible value of A? Can you draw an equilateral triangle of arbitrarily large area A on a surface with $\kappa = 0$? If not, what is the maximum possible value of A? Can you draw an equilateral triangle of arbitrarily large area A on a surface with $\kappa = -1$ and radius of curvature R? If not, what is the maximum possible value of A?

3.5 By making the substitutions $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, and $z = r\cos\theta$, demonstrate that Equations 3.29 and 3.30 represent the same metric.

4

Cosmic Dynamics

The idea that the universe could be curved, or non-Euclidean, long predates Einstein's theory of general relativity. As early as 1829, half a century before Einstein's birth, Nikolai Ivanovich Lobachevski, one of the founders of non-Euclidean geometry, proposed observational tests to demonstrate whether the universe was curved. In principle, measuring the curvature of the universe is simple; in practice, it is much more difficult. In principle, we could determine the curvature by drawing a really, really big triangle, and measuring the angles α, β, and γ at the vertices. Equations 3.22, 3.25, and 3.27 generalize to the equation

$$\alpha + \beta + \gamma = \pi + \frac{\kappa A}{R_0^2}, \tag{4.1}$$

where A is the area of the triangle. Therefore, if $\alpha + \beta + \gamma > \pi$ radians, the universe is positively curved, and if $\alpha + \beta + \gamma < \pi$ radians, the universe is negatively curved. If, in addition, we measure the area of the triangle, we can determine the radius of curvature R_0. Unfortunately for this elegant geometric plan, the area of the biggest triangle we can draw is much smaller than R_0^2, and the deviation of $\alpha + \beta + \gamma$ from π radians would be too small to measure.

We can conclude from geometric arguments that if the universe is curved, it can't have a radius of curvature R_0 that is significantly smaller than the current Hubble distance, $c/H_0 \approx 4380$ Mpc. To see why, consider a galaxy of diameter D that is at a distance r from the Earth. In a flat universe, in the limit $D \ll r$, we can use the small angle formula to compute the observed angular size α of the galaxy:

$$\alpha = \frac{D}{r}. \tag{4.2}$$

In a positively curved universe, the angular size is

$$\alpha_+ = \frac{D}{R_0 \sin(r/R_0)}. \tag{4.3}$$