It is not hard to see why this problem can be cured with the term involving C, containing a product of Δx and Δy. Remember the "cosine rule" for the length ℓ of the side of a triangle opposite to an angle θ, formed from sides of length x and y:

$$\ell^2 = x^2 + y^2 - 2xy\cos\theta.$$

When the triangle's two sides are perpendicular, then $\theta = 90°$, so that the last term is zero, leaving the usual Pythagorean theorem. But in general one needs the cosine term to compensate for the fact that the distances x and y are measured in directions that are not perpendicular to each other. We see that the cosine formula is a special case of the formula Equation 18.5 with $C = -2\cos\theta$ and $A = B = 1$. The cosine formula can be thought of as the distance formula in flat space with straight coordinates that are skewed, so that there is an angle θ between the coordinate axes, an angle that is not necessarily a right angle.

Of course, the cosine rule is a formula that is correct only for a triangle in a flat two-dimensional plane. But every smooth curved space is locally flat (i.e. flat if we look at a small enough piece of it), so if the differences Δx and Δy are small enough, we can interpret the distance formula Equation 18.5 exactly as a version of the cosine rule. Therefore, C at any point just measures the angle θ between the directions of the coordinates at that point: $\cos\theta = -C/2AB$.

> This shows that an ordinary space (not spacetime), where squared distances must be positive, cannot have a distance formula with arbitrary coefficients: C^2 must be smaller than or equal to $4AB$ (in order that $\cos\theta$ should be less than or equal to one) and A and B themselves must be positive.

Figure 18.3 illustrates a coordinate system for a two-dimensional surface that is curved. It shows that, even when the coordinates are drawn in a very regular and smooth way, they stretch and turn to follow the surface. If we choose any grid line in the diagram and move along it, we see that the distances (measured along the surface, of course) between successive intersections with other grid lines do not remain a constant length: the grid is stretched and compressed. We also see that grid lines do not always intersect at right angles. The distance formula for this surface, in this coordinate system, will have non-zero functions A, B, and C.

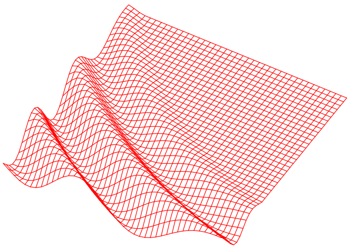

Figure 18.3. *This drawing shows a simple, nearly-rectangular coordinate system drawn on a curved (wavy) two-dimensional surface. It is impossible to keep the coordinates smooth without stretching them. Generally the coordinate lines also cannot be made to intersect at right angles.*

It is important to understand that the functions A, B, and C depend on the coordinates we have chosen, as well as on the curvature of the surface. There are many different ways I can draw coordinates on a surface, and the amount of stretching and squashing of the coordinates I need to do will depend on how I draw them, even though the surface will remain the same.

▷In fact, it is possible to have complicated functions even on the flat plane, just by choosing the coordinates differently from the usual Cartesian coordinates x and y. For example, the Pythagorean theorem for small distances using polar coordinates r and θ in the plane is

$$(\Delta\ell)^2 = (\Delta r)^2 + r^2(\Delta\theta)^2.$$

> Therefore, while the functions A, B, and C contain information about the curvature of the surface, they are not uniquely determined by the curvature: they depend on the coordinate system as well.

Newtonian gravity as the curvature of time

In this section: we learn how to understand the curvature of time.

How do we use distance measures in curved spaces to describe Newtonian gravity? We discussed curved two-dimensional surfaces because we could visualize them, but

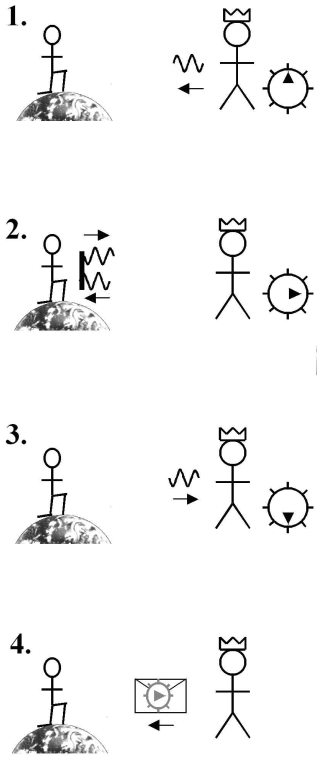

Figure 18.4. *How a time coordinate could be assigned to events on Earth by a distant experimenter using his own clock. The distant experimenter (the "master"), at rest, sends a photon to his "slave" on Earth (1), who reflects it back (2). The master notes the time he receives it (3). Since light takes the same time to travel to the Earth as to return, the master assigns the* average *of the times of (1) and (3) to the reflection event (2) as its* time-coordinate value. *He tells his slave this much later, in a letter (4), but that is okay: the event has been given a unique time. If the master does this again, then the slave will notice that the elapsed* coordinate *time is different (longer) than the time elapsed on a clock on the Earth. Being a slave, he is in no position to object to this! Nor should he: the master determines the coordinate time to be assigned to things, even if they are not the true (proper) times. That is what a time-coordinate is.*

it is harder to visualize a four-dimensional curved spacetime! The remarkable thing is that the mathematics that we have developed for measuring distances changes very little when we adapt it to describing gravity.

One big change is that in spacetime, the distance measure is the *spacetime-interval*, not the Euclidean distance, so we expect the coefficient of the time term to be negative.

The only other change is that in principle we need to use all three dimensions for space: gravity is a property of our real three-dimensional world. In the last chapter we saw how to write the spacetime-interval in just one space dimension, Equation 17.1 on page 217. To put in the other two dimensions is simple:

$$\Delta s^2 = -(c\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2. \tag{18.6}$$

The extra two space dimensions y and z have the same footing as x, and for purely spatial spacetime-intervals ($\Delta t = 0$) this is the standard Pythagorean rule in three dimensions. It follows from our discussion above that a general curved spacetime is described by modifying the spacetime-interval in Equation 18.6 to add in variable coefficients and "mixed" terms. Since this could get extremely messy, we'll only just note that it must be done to be fully general, but we won't need to do it for our discussions! Instead, we focus here on putting a variable coefficient in front of $(\Delta t)^2$.

The key to linking the notion of curvature in time with Newtonian gravity is the gravitational redshift. We have already seen in Chapter 2 that the gravitational redshift affects the rate at which clocks run. Imagine that we establish a time coordinate in the Solar System so that the time assigned to any event is the time that is recorded by a clock far from the Sun when the event occurs. It is worth thinking a little about how this time coordinate could be set up. Let's do it for the Earth, as a concrete example.

As we have remarked before, a clock on the surface of the Earth runs slower than one far away. How do we measure this? Let the clock on the Earth send out a radio signal each time it ticks. Then this signal will take a while to reach the distant clock, but the signals from both ticks take the same amount of time to travel, so the time between their arrivals at the distant clock will depend only on the time between their emissions: the time between ticks of the Earth-bound clock.

> We *define* our time coordinate t even at the Earth-bound clock to be the time elapsed on the distant clock between Earth-bound ticks. This will be longer than time on the Earth-bound clock, because of the gravitational redshift. But since clocks at different altitudes on the Earth will also run at different rates, there is nothing special about the Earth-bound clock. Our global time coordinate t has the advantage that it is possible to define it anywhere.

Of course, this time is not the time that the Earth-clock ticks. Our t is just a coordinate, a way of locating events in time. It is not meant to be directly a physical measurable. The proper time, given by the spacetime-interval, is the time on the local clock.

The gravitational redshift forces us to be careful about defining time. We saw in our discussion of the GPS in Chapter 2 that we now have to take into account the differences in redshifts of different clocks in our daily timekeeping on the Earth. The gravitational redshift causes local clock time (the proper time) to be different from our time-coordinate time t, so it is exactly the factor that converts from coordinate

Investigation 18.1. The gravitational redshift tells us how time curves

In Chapter 2 we saw that the effect of a Newtonian gravitational field was to change the rate at which clocks ticked. Now, the proper time given by the spacetime-interval is the time on clocks. The coordinate time is rather arbitrary, but it is helpful to take it to be the same as the proper time of clocks that are far from the gravity of the star or black hole that we are considering. These are "our" clocks, the clocks that we as astronomers far away from the system use to measure time.

Suppose that after a given amount of coordinate time Δt, a clock at rest in the gravitational field has ticked a proper time $\Delta\tau$. The relation between these depends on the clock's position in the gravitational field. For a clock at a distance r from a Newtonian body of mass M, a simple extension of the redshift calculation of Investigation 2.2 on page 16 shows that this relation is

$$\Delta\tau = \left(1 - \frac{GM}{c^2 r}\right)\Delta t. \tag{18.8}$$

Notice that proper time and coordinate time are equal when we are far away from the star or black hole ($r \rightarrow \infty$), which is how we defined the time coordinate t. This equation is only valid if the Newtonian field is weak, i.e. if $GM/c^2 r \ll 1$.

Now, along the world line of a clock that is at rest, the spatial coordinates don't change, so if we use the spacetime-interval to calculate the proper time, we can set $\Delta x = \Delta y = \Delta z = 0$. The negative of the spacetime-interval Δs^2 is the square of the proper time, $\Delta\tau^2$, times c^2, so we are led to

$$\Delta s^2 = -\left(1 - \frac{GM}{c^2 r}\right)^2 (c\Delta t)^2. \tag{18.9}$$

The term that is squared can be simplified by expanding the square:

$$\left(1 - \frac{GM}{c^2 r}\right)^2 = 1 - 2\frac{GM}{c^2 r} + \left(\frac{GM}{c^2 r}\right)^2.$$

The last term on the right-hand side is very small, since we are assuming the gravitational field is weak. For example, on the surface of the Earth we have $GM/c^2 r \approx 10^{-8}$, so the last term is about 10^{-8} as large as the second term. We will neglect it now. We must not throw away the second term, however, since that is where all the deviations from special relativity occur! We get

$$\Delta s^2 = -\left(1 - \frac{2GM}{c^2 r}\right)(c\Delta t)^2. \tag{18.10}$$

The time part of the spacetime-interval is therefore determined by the gravitational redshift effect. For the spatial coefficients, we make the simplest assumption: keep them the same as in special relativity. This gives the spacetime-interval Equation 18.7. We will see that this works perfectly for geodesics that represent particles going at non-relativistic speeds.

By the way, don't think that Equation 18.9 is more accurate than Equation 18.10, just because we have dropped the squared term to get the result. In fact, Equation 18.9 is itself not fully correct when gravitational fields are strong, so it can have errors of the same general size as the term we have neglected. Ironically, it turns out that Equation 18.10 is actually the form that is correct for strong fields as well as weak ones.

Exercise 18.1.1: ***Redshift near the Sun***

Derive Equation 18.8, starting from Investigation 2.2 on page 16. Calculate the redshift experienced by a photon with a wavelength of 0.5 μm as it travels from the surface of the Sun to a very distant observer. Calculate the redshift of the same photon if it is observed by a space observatory in the Earth's orbit but far from the Earth. Finally, calculate the redshift if the same photon is observed by an astronomer on the surface of the Earth.

---

time to proper time. The gravitational redshift is therefore precisely the "squashing" factor that we are looking for in the spacetime-interval formula for the Solar System. The details of how to put the redshift factor into the spacetime-interval are worked out in Investigation 18.1. The result is a spacetime-interval where only the time-coefficient is variable:

$$\Delta s^2 = -\left(1 - \frac{2GM}{c^2 r}\right)(c\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2, \tag{18.7}$$

where M is the mass of the Newtonian star (the Sun could be replaced in this argument by any star) whose gravity we represent by this curved spacetime, and where $r^2 = x^2 + y^2 + z^2$ is the distance from the star to the point $(x,\ y,\ z)$ in space where the clock is.

▷Note how different a time coordinate is from a time measurement. To *measure* the time passing on Earth, the clock must be on the Earth too. To assign an arbitrary *time-coordinate* one can use any clock, and it is particularly convenient to use one so far away that the Earth's gravity does not slow it down.

Do the planets follow the geodesics of this time-curvature?

The test of whether our expression Equation 18.7 for the spacetime-interval represents the real world is whether this spacetime has geodesics that are the Newtonian trajectories of particles orbiting the mass M. This may again sound like a hard thing to show, but it is not. In fact, we have essentially done all the work we need to show this. We just have to assemble the various components of the argument.

In this section: we show that the motion of any object acted on by a Newtonian gravitational force can be fully described instead as a free motion along a geodesic of a geometry with curved time.

The key is to realize that the locally flat coordinates in a spacetime are the coordinates of observers who fall freely with the acceleration of gravity. These observers can, by the equivalence principle, expect to do local experiments and have them come out exactly as in special relativity, as if there were no gravity. Now, one of the experiments they can do is to watch another nearby freely-falling particle. Since there is no gravity in their local (freely-falling) spacetime coordinate system, this particle will move on a straight line through their coordinates.

But this is the definition of a geodesic: a geodesic is a straight line in a locally flat coordinate system. Therefore, the geodesics of a spacetime in which the locally flat coordinates are those of experimenters falling freely in a gravitational field are the trajectories of freely-falling particles in the same gravitational field.

▷Fundamentally, we have turned our old derivation of the gravitational redshift completely around. In Chapter 2 we derived the redshift from the Newtonian gravitational force. In the present chapter, we have derived the Newtonian "force" (really, the equivalent spacetime geometry) from the gravitational redshift. From Einstein's point of view, the redshift is the more fundamental of the two, since it directly measures the geometry of spacetime. The motion of particles follows almost incidentally from that geometry.

> This proves that Equation 18.7 on the preceding page describes a spacetime geometry in which particles that follow geodesics will move on exactly the same trajectories as particles would do in a flat spacetime with the Newtonian gravitational force acting. We have therefore found a curved-spacetime picture of Newtonian gravity. The curvature here is *only* in the time-direction. Curvature in time is nothing more than the gravitational redshift: time advances at different rates in different places, so time is curved. We have found that the gravitational redshift fully determines the trajectories of particles in the gravitational field.

We have arrived at this goal with a minimum of calculation. We did not have to do any calculus or solve any differential equations. Yet we now know what it means physically when we say that time is curved: it means that the rate at which clocks run changes from place to place, even when the clocks are at rest with respect to one another. The curvature of time is in the gravitational redshift, and the gravitational redshift is enough to insure that freely-falling bodies follow their Newtonian trajectories.

> *All* of Newtonian gravitation is simply the curvature of time.

How to define the conserved energy of a particle

In this section: since the gravitational redshift changes the energy of a photon, it tells us how to compute energy changes for all particles as they move through curved time. We learn how to define an energy that is conserved along geodesics.

The frequency of photons changes because of the redshift, so their energy also changes. Nevertheless, it is possible to define a conserved energy in relativity, just as it was in Newtonian gravity. This should not be surprising, especially considering that energy depends not only on the particle but also on who is measuring it.

Let us identify the experimenters who measure the redshifted energy of a photon. They are local experimenters who are at rest with respect to the star. They each perform a local experiment to measure the frequency of the light, and they find that it decreases as the photon climbs away from the Earth. If they had been observing a freely-falling particle they would have found a similar result: the speed, and hence the kinetic energy, of a particle gets lower and lower as it gets further and further from the star.

An observer very far away, so far that GM/rc^2 is too small a correction to measure, is in a special position: this is where we are when we observe almost all astronomical systems outside the Solar System. In an ideal case, where we consider only the gravity of a single star, this distant observer lives for all practical purposes in the spacetime of special relativity. In special relativity, the energy of a particle or photon is constant as it moves. Although this is not true of the photon that moves near the star, it becomes true when that photon moves far from the star: as it leaves the gravitational influence of the star, its energy (frequency) becomes constant, regardless of where it is going.

▷Actually, although we are far from other astronomical bodies, we do sit deep in the Earth's gravitational field. Astronomers agree to use a universal time coordinate that matches proper time on the Earth, not in interstellar space.

▷The conservation of the energy we have defined is not actually trivial, as explained below.

This energy, as measured by a distant observer, is called the *conserved energy* of the photon. It is conserved in what might seem to the reader to be a trivial sense, in the sense that it is a number that is *defined* to be a constant equal to the energy when the photon is far away. If we associate this number with the photon even