General Relativity for the Faint of Heart

I'm going to try to explain, as non-technically as I can, what it means to say that space-time is curved. If it doesn't make any sense to you, please understand that it's not your fault. It's just not something that is designed to make sense!

Also, in trying to explain this without too much math, I'll be talking about general relativity in a way that is quite out of date. Just the same, it is a fundamentally acceptable if mathematically inelegant way of looking at gravity in a general relativity context.

Those of you reading probably understand, intuitively, the way that gravity works in Newtonian physics. It's an attractive force that applies between any two objects with mass, such as the Earth and a tennis ball. Since the force is always attractive, if you throw the ball up in the air, it doesn't continue straight up forever (provided you have ordinary strength). Instead, it slows to a stop and then comes back down to your hand.

However, in general relativity, gravity is not a force, but an effect caused when objects move through curved space-time. We say that these objects follow "geodesics": paths that are as "straight" as can be on the curved surface of space-time. What is confusing is that this word "straight," like other words used in popular expositions of general relativity, doesn't mean what you're accustomed to. (Incidentally, the warped rubber sheet analogy commonly used to explain general relativity is actually a Newtonian construct; it's a curvature, all right, but not the right kind. The dimples created by masses represent potential energy wells.)

The root of the problem is that we're not accustomed to thinking about the term "straight" on a curved surface. The word "straight" presents no problem at all, as long as you're dealing with a flat piece of paper. If you're asked to draw a straight line between two points, you simply plop a ruler down and draw the line. We recognize that the ruler may not be perfectly straight, the paper has some slight roughness, and the pencil point has some finite thickness, but these are engineering issues. In *principle*, we can do it.

But what if the two points are on a curved surface, such as a sphere? We can't just plop the ruler down on the sphere, because the ruler is flat and the sphere isn't. The ruler can only touch the sphere at one point, which isn't going to help us draw a line. We can roll the ruler along the sphere, but there are many ways to do that, yielding many different lines. How can we tell which of those lines is as straight as possible?

It's not simply a matter of skill and care in laying a ruler down. The problem is defining what straight *means* on a curved surface such as the sphere. There are many ways to make this definition, but probably the most common is to recognize that on a flat piece of paper, a straight line is not only the path between two points that makes no turns, but also the shortest and most direct path between those two points. And this idea of the shortest path *can* be applied to a curved surface. We can, for instance, run a piece of string between the two endpoints. If we lay the string down on a flat surface, we can then use our ruler to measure the length of string we used.

On a sphere, the shortest path between two points lies on what is called a *great circle*. A great circle is any circle that cuts the sphere in half. In fact, "geodesic" comes from Greek words meaning "to cut the Earth." On the Earth, for instance, the equator is a great circle. So is any pair of diametrically opposed lines of longitude, such as 30 degrees west paired with 150 degrees east. You can see that if you're trying to get to the north pole from any spot on the Earth, the shortest path goes along the line of longitude passing through that spot, and follow it northward to the pole. Similarly, the shortest path between any two points on the equator follows the equator. These are so obvious they almost don't bear mentioning.

What's not so obvious, however, is the shortest path between, say, New York and Madrid, which both lie at a latitude of about 40 degrees north. You might think that the solution is to stay at 40 degrees north, all the way from New York to Madrid.

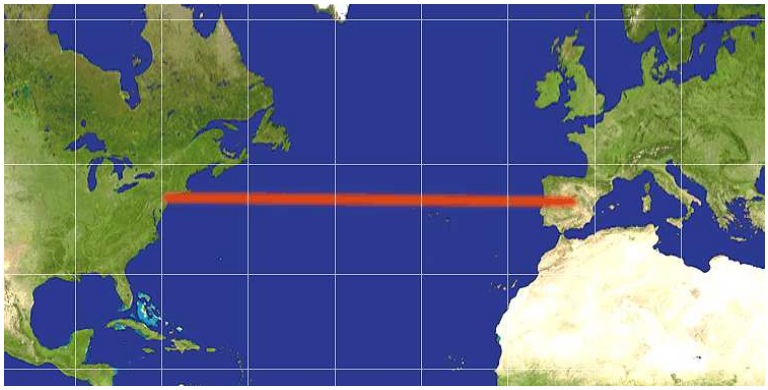

Figure 1: *Not the shortest path.*

But lines of constant latitude, unlike those of constant longitude, are not great circles, except at the equator. They don't cut the Earth exactly in half. So this can't be the shortest path. To find the great circle path between New York and Madrid, we must slice the Earth in half—figuratively speaking, of course!—passing through New York, Madrid, and the center of the Earth.

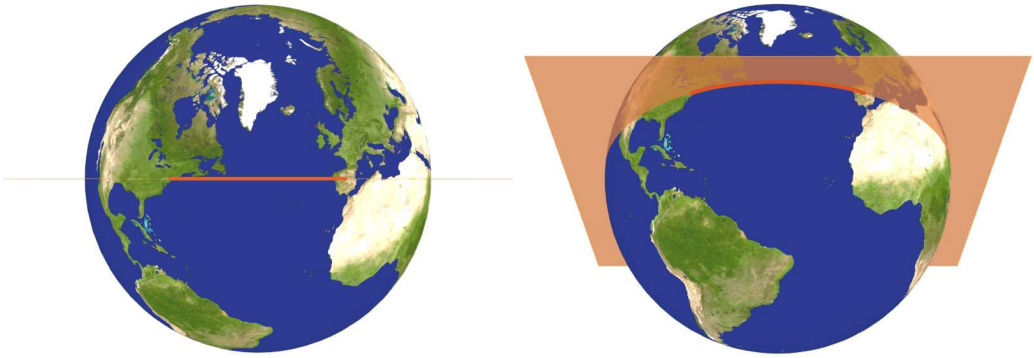

Figure 2: *Finding the great circle path.*

This slice cuts the Earth along a great circle. The path between the two cities is shorter, and thus straighter, than any other path between them. If you look at a globe, you can tell that this great circle is aligned in such a way that it is *north* of 40 degrees between New York and Madrid, and *south* of 40 degrees anywhere else.

Here's another way to think about it. Both cities are at a latitude of 40 degrees north, but Madrid has a longitude of about 4 degrees west, and New York one of 74 degrees west. To get from New York to Madrid, you must traverse those 70 intervening degrees of longitude. If you stay at a latitude of 40 degrees, those 70 degrees are each about 85 km wide. That path therefore covers a total distance of 70 times 85, or 5,950 km.

However, you can shave some distance off by taking advantage of the fact that degrees of longitude are narrower at higher latitudes. A degree of longitude at a latitude of 50 degrees is only 71 km wide, or about 5/6 as wide as that same degree at a latitude of 40 degrees. The total path length can be made shorter by arcing to higher latitudes and traversing those shorter degrees of longitude, and then returning to a latitude of 40 degrees. Indeed, degrees of longitude have zero width at the north pole. But to use those zero-width degrees, you would have to use up more distance getting to the north pole in the first place than you would save by not using up any distance traversing the 70 degrees of longitude.

The shortest path, then, is in between, going up gently in latitude to save distance in longitude. It turns out that the best path using this approach is the same as the great circle, which has a total length of about 5,770 km. Not tremendously shorter, but shorter just the same.

Now, consider what this path looks like on a map. A map, unlike a globe, is usually drawn on a flat piece of paper. Since a spherical globe doesn't unwrap nicely to a flat piece of paper, one has to make compromises when creating a map of the spherical Earth. You can make the areas of the various countries come out right, but not their shape; or you can get the shapes right, but not the areas; or you can get something in between. But in no case can you get everything exactly as it looks on the globe.[1]

Figure 3: *The great circle path on a Mercator map.*

[1] All Earth images in this article are courtesy Ernie Wright.

The process of creating a flat map from a spherical Earth is called *projection*, and there are many different projections, each with their own special properties. On a map with the Mercator projection, for instance, lines of longitude and latitude are drawn straight. Furthermore, those lines are spaced in such a way that paths with constant bearings (such as a path that always bears to the northwest) are drawn as straight lines. On such a map, it certainly looks as if the shortest path between New York and Madrid is to travel along the line of latitude at 40 degrees north. But we already know that this is not the shortest path. Instead, the shortest path follows a great circle.

On the Mercator map, that path looks curved, because it arcs up to a maximum latitude of about 46 degrees north, then levels out, and then comes back down to 40 degrees north again at the end. It's important to recognize that this apparent curvature is only the result of looking at a map with straight lines of longitude and latitude, and is not part and parcel of the path itself. On the curved Earth, the great circle path looks as straight as can be, because it *is* as straight as can be.

When we consider paths on the Earth, we must deal with the curvature of the Earth in three dimensions. But when we consider general relativity, we must deal with at least one additional dimension: the dimension of time. Any more or less complete treatment of the topic must therefore incorporate all four dimensions. However, since four dimensions are hard to visualize, and even three isn't easy, let's just concentrate on two dimensions: one time dimension, and one spatial dimension—up and down.

Go back to tossing the tennis ball up in the air. Suppose the height of your hand is 1 m, and we'll start our clock at a time of 12:00:00 noon. At this point, the space-time coordinate of the ball can be given as (12:00:00, 1), indicating that the ball is at a height of 1 m when the time is 12:00:00.

If you throw the ball up at a speed of 10 m/s, it returns to your hand in 2 seconds, having reached a peak of 6 m. At that peak, the space-time coordinate of the ball is (12:00:01, 6), and when it falls back down to your hand, it is (12:00:02, 1).

This all makes sense in Newtonian physics, because in that framework, gravity is a force. We can use Newton's law, $F = ma$, and figure out how the ball moves in response to the gravitational force. (In fact, that's how I arrived at the numbers above, using a gravitational acceleration of 10 m/s^2 instead of the more typical 9.8 m/s^2, just to make the figures come out easier.)

But in general relativity, it's different. We don't figure out what the force and resulting acceleration is on the ball at any point in time, because there is no gravitational force. Instead, we must ask ourselves, what is the straightest path between the initial and final coordinates of the ball, (12:00:00, 1) and (12:00:02, 1)?

If we plot these two points on a graph, with time along the horizontal axis, and the height of the ball along the vertical axis, the answer seems obvious: The straightest path is at a constant height of 1 m above the ground. But after considering the distortion caused by mapping the curved Earth on a flat piece of paper, I hope you'll see that this can't possibly be right, and not only because this apparently shortest path corresponds to the ball hovering in mid-air, 1 m above the ground, for 2 whole seconds!

The reason why this path isn't the straightest is that in plotting the path in the usual manner,

we are mapping curved space-time onto a flat piece of paper. Some distortion is inherent in this projection. We've made our lines of space and time straight, and just as in the Mercator map, the path that is actually shortest and straightest will look curved.

In what way are the lines of space and time curved? To answer that, we we have to look at two properties of general relativity—properties that are counter-intuitive, but which have been repeatedly confirmed in experiments. First of all, clocks run slower near a massive object such as the Earth than they do far from it. For example, a clock on the bottom floor of a tall building will run slower than one on the top floor, all other things remaining equal. There's nothing special about the Earth in this regard; any amount of mass warps time in this way. Even though the clocks are seemingly at rest with respect to one another, their respective heights above the Earth's mass puts them in two distinct frames of reference, which makes it possible for them to run at different speeds. But is there in fact any difference?

We can determine that by introducing a third clock, which starts out at the level of the lower clock, but is then thrown up in the air, so that at its peak, it's as high as the higher clock. This third clock is moving inertially, because it's not subject to any external forces; remember that in general relativity, gravity is not an external force. Therefore, we can use it as a reference point for comparing the speeds of the other two clocks; we'll call it our *reference clock*.

When the reference clock is first being thrown upward, it and the lower clock are moving rapidly past each another, so that even though they are in the same place, they don't share a common frame of reference: Each one measures the other as running slow. We know this from our understanding of special relativity. Now, we don't care what the lower clock measures, only what the reference clock does. From its point of view, the lower clock is running slow.

On the other hand, when the reference clock reaches the top of its arc, it's at the same height as the higher clock, and it's motionless like the higher clock. Therefore, they *do* share a common frame of reference and must therefore be running at the same speed. Well, if the reference clock measures the higher clock as running at the "right" speed, and the lower clock as running slow, then the higher clock must be running faster than the lower clock. And that turns out to be so.[2]

The second counter-intuitive property of relativity is that not everyone will agree on how long things take to happen, and how much distance they cover. For example, if I'm riding toward you on a bicycle as you toss the ball up, you'll measure the time elapsed as exactly 2 seconds, but I'll measure it as just barely more than 2 seconds.

What we *do* agree on, however, is what's called the *space-time interval* between your tossing the ball up and your catching the ball. The space-time interval between two events is like the regular distance between two points in space, with one important difference. In ordinary space, the distance between two points is given by the Pythagorean formula[3]:

[2]You might wonder why it isn't equally valid to measure the speed of the two stationary clocks in the opposite order, which would seem to lead to the opposite conclusion—that the lower clock was running faster. The reason is that we can't compare clocks that way unless they're in the same place with respect to the gravitational field—that is, when they're at the same height. And even though the reference clock is moving at all times, it actually measures both of the other clocks as running at a constant rate at all times. Any change in the altitude difference between two clocks is counteracted exactly by a corresponding change in the relative speed between the clocks.

[3]In what follows, I'm going to use the terms dx, dt, etc. to refer to difference in x-position, difference in time, and

$$ds^2 = dx^2 + dy^2$$

In other words, if you're standing 30 meters north and 40 meters east of me, you must be 50 meters away from me, because

$$30^2 + 40^2 = 900 + 1600 = 2500 = 50^2$$

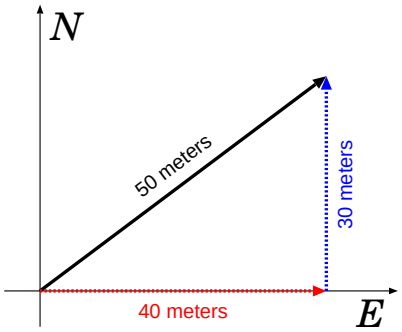

Figure 4: *Measuring the distance between two points.*

This distance is an *invariant* over rotations, meaning that it doesn't matter if our compasses are off from one another. Instead of 30 meters north and 40 meters east, you might measure it as 40 meters north and 30 meters east. Or, you might measure it as 10 meters north and about 49 meters east. Our dx and dy measurements might vary, but they don't vary freely: in every case, the separation between us must be exactly 50 meters.

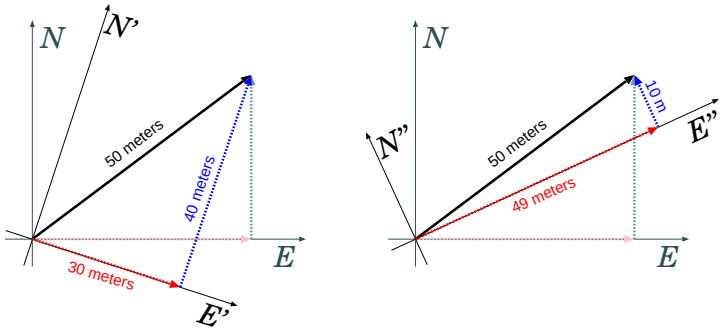

Figure 5: *The distance is not affected by how you align your north and east axes.*

so forth. When I write something like dy^2, that means the square of dy. Strictly speaking, because I'm going to be talking about small but measurable distances, I should be using Δx, Δt, and so forth, but since it matches standard notation and doesn't really affect the mathematics, I'll ask your indulgence in allowing this slight abuse of notation.

The same thing happens with space-time in general relativity, except that instead of *adding* the squares of the space and time intervals, you subtract the square of the time interval from the square of the space interval. That is,

$$ds^2 = dx^2 - dt^2$$

This is the important difference between distances and space-time intervals. It is this measure of the separation between two points in space-time that we use in determining the shortest, and therefore straightest, path. Because ds^2 is an invariant, everyone agrees on its value, and we can ask anyone what its value is. (See the endnote for more on this.)

In this case, we're going to ask an ant riding on the tennis ball as you toss it up and down. The advantage of asking the ant is that for the ant, the tennis ball doesn't move up and down—the rest of the world moves down and up around it. For the ant, therefore, dx and dx^2 are both zero. In that case, the ant just sees

$$ds^2 = -dt^2$$

and minimizing ds^2 amounts to maximizing dt^2. dt is the time interval covered by the ball, as measured by someone moving with the ball itself, so it is often called the *proper time*. To figure out how the ball moves, just ask the ant which path between (12:00:00, 1) and (12:00:02, 1) has the longest proper time! We don't *have* to ask the ant; anyone will figure out the same path for the ball. It's just that the ant sees $dx = 0$, so the question is easier for the ant.

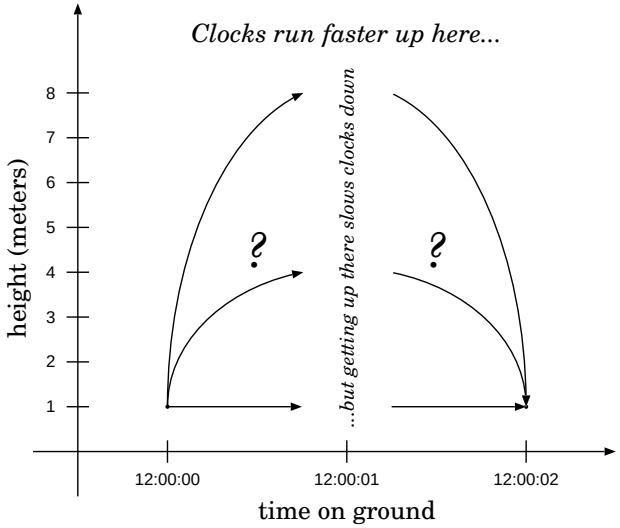

Figure 6: *Which path maximizes the time elapsed as measured by the ball? That is the path the ball actually travels.*

One path *is* to just hover at 1 m above the ground, in which case the proper time is 2 seconds. But that's not the best we can do, because high-altitude clocks run faster. So it makes sense for

the ball to move to a higher altitude, where more proper time will elapse. This is why the ball arcs upward.

So why not go really high up? Because the ball only has 2 seconds, as measured by you on the ground, to go up and come down. If it tries to go too high up, it will have to move very rapidly, and special relativity tells us that rapidly moving clocks run slow, ironically enough. You'll slow the clock down more than you gain by moving the clock to a higher altitude.

It's just like finding the shortest path between New York and Madrid, where you could go really far north, but you'd travel further just going that far north than you'd save by traversing narrower degrees of longitude. You have to take the in-between route that minimizes the distance covered. Well, in this case, the ball takes the in-between route that maximizes the proper time elapsed. It looks curved when you plot it on a grid with straight lines of space and time, but that's because lines of space and time aren't straight in curved space-time. If you could see space-time as it really is, you would see that the path taken by the ball is a geodesic—as straight as can be.

By the way, if you do plot it with seconds of time on one axis and meters of height on the other, you get a parabola that looks quite curved indeed—and this on the relatively weak gravity of the Earth. How could the measly Earth cause such a great apparent curvature of the tennis ball's path? But we're biased by using meters and seconds. In general relativity, units of length and time are tied to one another through the speed of light, about 300,000,000 m/s. It takes a lot of meters to make up just one light-second.

To see the true curvature caused by the Earth's gravity, we should really plot the path using seconds on one axis, and light-seconds on the other. Since the 5 m height gained by the ball amounts to just 17 billionths of a light-second, you can see that the curvature created by the Earth really isn't much at all. The plot would be a section of a parabola that is only about a hundred-millionth as tall as it is wide. To the unaided eye it would look perfectly straight. It takes very intense concentrations of mass—such as a black hole—to create curvature that you could see with the unaided eye in such a plot.

It turns out that the path of longest proper time is almost exactly the same as predicted by Newtonian physics. But if you get pretty much the same answer either way, why go through the more complex general relativity treatment, rather than the much simpler Newtonian approach?

The answer is, because the two are only *almost* the same. There is a slight difference, which is inconsequential so long as we restrict ourselves to tennis balls moving at ordinary speeds, but becomes significant when we consider faster objects such as subatomic particles or light itself. Roughly speaking, this difference arises because gravity warps not only time (the way that clocks run slower in a gravitational field), but space as well. Tennis balls travel through so little space in a given amount of time that the additional impact is small, but that's not the case for light, which travels at 1 light-second per second. Technologies such as GPS, which use signals travelling at the speed of light, are predicted to behave differently by the two theories. And in each case where the difference is measurable, experiment has always come down on the side of general relativity. In fact, we know of no measurable inaccuracy in general relativity. It is one of the crowning achievements of 20th-century physics.

Endnote on the Space-Time Interval

Why is the space-time interval defined as

$$ds^2 = dx^2 - dt^2$$

rather than as the sum $dx^2 + dt^2$? Because, as I said above, it's an invariant: Everyone agrees on the same value for ds^2, even if they disagree on dx^2 and dt^2. We can show this in the case of two reference frames that are both flat space-times, governed by special relativity. We can therefore use the Lorentz transforms to go from one frame to the other:

$$dx' = \gamma(dx - v\,dt)$$

$$dt' = \gamma(dt - v\,dx)$$

where $\gamma = 1/\sqrt{1 - v^2}$ is the Lorentz contraction coefficient. If we square both of these equations, we get

$$(dx')^2 = \gamma^2(dx^2 - 2v\,dx\,dt + v^2dt^2)$$

$$(dt')^2 = \gamma^2(dt^2 - 2v\,dt\,dx + v^2dx^2)$$

Now subtract the latter from the former to get

$$\begin{aligned}(dx')^2 - (dt')^2 &= \gamma^2(dx^2 + v^2dt^2 - dt^2 - v^2dx^2) \\ &= \gamma^2(dx^2 - v^2dx^2 - dt^2 + v^2dt^2) \\ &= \gamma^2(dx^2 - dt^2)(1 - v^2)\end{aligned}$$

But $\gamma^2 = 1/(1 - v^2)$, so

$$(dx')^2 - (dt')^2 = dx^2 - dt^2$$

so both reference frames agree on the value of ds^2. That wouldn't be the case if you defined ds^2 as $dx^2 + dt^2$, as you can easily confirm for yourself. (In four-dimensional space-time, the interval is defined analogously as $ds^2 = dx^2 + dy^2 + dz^2 - dt^2$.)

As an example, suppose I'm riding toward you on my bicycle at 3 m/s (about 7 mph), or one hundred-millionth of the speed of light. In that case, the same 2 seconds you measure between tossing the ball up and then catching it, I measure as just a tad longer than 2 seconds. The

difference is minuscule: about 100 attoseconds (1×10^{-16} seconds), which is the amount of time it takes light to cross about 300 atoms laid end to end.

As small as it is, however, that difference is not insignificant at a theoretical level; it can't simply be brushed under the rug. In those 2 seconds, I've travelled toward you about 6 m. Another way of expressing that distance is 20 light-nanoseconds—the distance covered by light in 20 billionths of a second. To you, you catch the ball in the same place you tossed it up: $dx = 0$. To me, $dx = 2 \times 10^{-8}$. Conversely, you measure the time as just 2 seconds: $dt = 2$. But to me, dt is just a bit larger, $2 + 1 \times 10^{-16}$.

Although we disagree about the values of dx and dt, we *don't* disagree on the value of ds^2. In both cases:

$$ds^2 = dx^2 - dt^2 = -4$$

It's just that to you, dx^2 is 0 and dt^2 is 4, while to me, dx^2 is 4×10^{-16} and dt^2 is $4 + 4 \times 10^{-16}$.