Gravity and geometry

Why is it that gravity can be described as a geometrical property of spacetime? Consider the trajectories of different particles in a gravitational field. Since the gravitational force is proportional to the inertial mass of an object (the Principle of Equivalence), all particles experience the same acceleration at the same position. Thus, if they begin at the same place and with the same initial velocity, Newton's Second Law says that they will follow the same trajectory. In other words, the trajectory of the particle does not depend on properties of the particle such as its mass or charge. In that sense, the trajectories are properties of the field and not properties of the particle. But then why think of the gravitational field as separate from the properties of spacetime itself. If *all* particles behave the same way, there is no way to separate the effects of gravity from the structure of spacetime. That tells us that a geometrical theory of gravity could work and that it is not futile to try it. It does not prove that it must work nor does it tell us how to go about formulating a theory. It was Einstein who told us how to do it. (In my view, the geometrization of gravity in his General Theory of Relativity is a much greater intellectual achievement than was Special Relativity. It was a greater leap from the conventional views of the age.)

The basic idea is that gravitational effects are replaced by the curvature of spacetime. There is no more gravitational force and no more gravitational field. But since Newton's theory of gravity works very well in most regions easily accessible to us, there must be some limit in which General Relativity reduces to Newtonian theory with a gravitational field. In a flat space, parallel lines stay parallel—they never cross. In the absence of forces, the trajectories of particles are straight lines, that never cross if they start out with parallel velocity vectors. However, when there is gravity (and, in fact, there is always gravity) particles that start out parallel, will approach each other as time increases. That effect becomes a consequence of the curvature of spacetime in General Relativity. A simple example is the surface of a sphere. Lines that begin parallel and heading north at the equator get closer together as they move farther north. The curvature of spacetime reproduces the effects that we have been attributing to a gravitational interaction.

So the motion of particles is controlled by the curvature of spacetime. But, where does the curvature of spacetime come from? It comes from the presence of matter! Matter curves spacetime, and then the motion of other matter is affected by that curvature. That's what gravity is.

So our problem is to set up the mathematical machinery that is needed to describe the curvature of spacetime. Unfortunately, this is a rather involved and elaborate business. It takes way more time than we can spend on it. 'Too bad. You can begin to see what the difficulties are by just thinking about the surface of a sphere. It is a two-dimensional surface with constant curvature and a great deal of symmetry. Except for the plane, you could not ask for something simpler. Nevertheless there is no way to put coordinates on the sphere that reflect its nice properties. In fact, there is no way to put global coordinates on the sphere without introducing some kind of a singularity. The usual way to describe position on a sphere is the polar angle θ and the azimuthal angle φ. They have bad behavior at the poles (φ is undetermined), and you will keep yourself busy for a long time trying to express a nice, simple great circle that does not go through the poles in these coordinates. Thus, we will have to be content with a very rough an qualitative discussion of how all this works and leave the real work for another course.

The basic idea is that the geometry of spacetime is encoded in the metric. We have seen the metric of flat Minkowski space appears in the expression for the spacetime interval

$$ds^2 = g_{\mu\nu}\, dx^\mu\, dx^\nu \ .$$

(Remember that we are using the Einstein summation convention here.) In this discussion, I will use a different symbol $\eta_{\mu\nu}$ for the special metric of flat Minkowski spacetime and use $g_{\mu\nu}$ for the general case. So the expression above is how the infinitesimal spacetime interval is written in the general case of a curved spacetime. The metric $g_{\mu\nu}$ might be a really complicated function of the position coordinates x^μ. In Minkowski spacetime, the metric becomes the simple version that we have already used $g_{\mu\nu} \to \eta_{\mu\nu}$ with

$$\eta_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \ .$$

There are two topics to deal with. One is the way that a particle moves in a given geometry with a given metric. Two is how the distribution of matter determines the metric. Let's talk about the first one first. We have already seen that in flat space, the spacetime interval between two events is obtained from the path that has the maximum proper time connecting the two events. We have also seen that it is the path with constant velocity. Thus it is also the path that a free particle follows. A particle traveling between two events follows the path that maximizes its proper time. The correct generalization to curved spacetime is that the previous statement is still true. But then the path is no longer one of constant velocity. Let's try to figure out what this means. Suppose that the metric is independent of the time coordinate t. If the particle is at rest at some position, we have $dx^i=0$ (I=1,2,3) and

$$d\tau = \sqrt{g_{00}dt^2} = \sqrt{g_{00}}\,dt \quad .$$

This says that proper time will run faster relative to coordinate time where g_{00} is larger. That will be preferred in the trajectory of the particle. On the other hand, we have already seen that adding a "unnecessary" velocity reduces the elapsed proper time relative to the path that uses lower velocity to get the job done. Thus there is a balance: the particle would like to get to a region with larger g_{00} to get its proper time moving faster, but to get there, it needs to add velocity which has the cost of slowing the proper time. For the static case, the metric can always be put in the form

$$g_{\mu\nu} = \begin{pmatrix} g_{00} & 0 & 0 & 0 \\ 0 & g_{11} & g_{12} & g_{13} \\ 0 & g_{21} & g_{22} & g_{23} \\ 0 & g_{31} & g_{32} & g_{33} \end{pmatrix} \quad .$$

The spatial parts g_{ij} i,j=1,2,3 are basically negative numbers as the form of the flat space Minkowski metric suggests they should be for the conventions that we have chosen. Then the form of the proper time is

$$d\tau = \sqrt{g_{00} + g_{ij}v^i v^j}\,dt$$

with the second term in the radical making a negative contribution. This decreases the rate of change of the proper time when the velocity is not zero. A balance must be struck. And that balance determines the actual trajectory. The mathematics that is used is called the calculus of variations. It allows one to convert the condition that the path must maximize the proper time into a differential equation that the path must satisfy. You have already seen the form of this differential equation that applies to the case of low speed motion with forces that are not too large. It is Newton's Second Law!

Let's reiterate the main point of this: The principle that the particle chooses the path with the largest proper time means that it must strike a balance between moving to regions with larger g_{oo} and increasing its velocity in order to get there.

The second issue is to consider how the geometry of spacetime, as reflected in the metric, is determined. Here we get at big hint from the gravitational red shift. It follows from the Principle of Equivalence and shows that when the rates of clocks at different gravitational potentials are compared, the clocks lower in the potential run slower. What does this tell us about the metric? The metric is exactly the thing that gives the relationship between clock rates and coordinate time. Consider the situation in which there is a spherically symmetric mass at rest at the origin and we are investigating the effect that it has on space outside if it. Lets restrict ourselves to dealing with just one of the spatial coordinates: distance from the origin r. Since the mass is not moving, we can hope for a geometry that is static and for a metric that does not depend on time. Now consider two light pulses that leave the lower clock separated by a small time dt. The trajectory that they follow is determined by the metric through the condition ds=0 along the path. Since the metric does not depend on time, the two paths are the same except that the second is shifted up the time axis by a uniform dt relative to the first. Thus, the separation of the times at which they arrive at a larger r is also dt.

Now consider something physical: proper time or clock rates. At fixed r, so that dr=0, $ds=\sqrt{[g_{00}(r)]}dt$. Comparing the rates of clocks at r_0 and r gives

$$\frac{ds(r)}{ds(r_0)} = \left[\frac{g_{00}(r)}{g_{00}(r_0)}\right]^{1/2}$$

On the other hand, the gravitational red shift tells us that this ratio is

$$\frac{ds(r)}{ds(r_0)} = 1 + \frac{1}{c^2}\, g(r - r_0)$$

(The g in the equation above is the local acceleration of gravity not the metric.) Note that the combination $g(r-r_0)$ is just the gravitational potential difference between r and r_0. With that we can rewrite it as

$$\frac{ds(r)}{ds(r_0)} = 1 + \frac{1}{c^2}\,[V(r) - V(r_0)]$$

For ordinary circumstance, this is very close to one. Thus, we expect that the metric g is close to the Minkowski form η, and we write $g=\eta+h$ with h small. In particular, $g_{00}=1+h_{00}$. When this is substituted above and the expression is expanded to first order in h, we find

$$\frac{ds(r)}{ds(r_0)} = 1 + \frac{1}{2}[h_{00}(r) - h_{00}(r_0)] \quad .$$

Comparing the last two expressions for the proper time ratio, we conclude that

$$h_{00} = \frac{2}{c^2}V(r) \quad \text{or} \quad g_{00} = 1 + \frac{2}{c^2}V(r) \quad .$$

Now we really have something. This gives us some idea about how the presence of the mass affects the curvature of spacetime and gives a metric that depends on position and is not the Minkowski metric.

So the space is curved, so what? What does that have to do with what we usually think of as the effect of the gravitational force on the motion of particles? We have already discussed this. A particle moves between two given events in such a way as to maximize the proper time for the trip. Suppose it is to go from r_0 at $t=t_1$ back to r_0 at $t=t_2$. Of course, we know from our previous experience what it really does. If there are no other masses around, the particle just sits there and waits out the time t_2-t_1. If it is here near the earth, to get it to return to your hand after a time t_2-t_1, you need to toss it straight up just the right amount. We can reproduce this result with the requirement that particle maximize the proper time. If there are no masses around, the metric is the flat Minkowski $\eta_{\mu\nu}$. That has already been discussed: the path with maximum proper time is the path with constant velocity. In this case, where it starts from and returns to the same point, that means zero velocity. In the case with a nearby mass and a different metric, there is the balance between the desire to get to regions with larger g_{00} and the desire to keep the velocity low. Thus it goes up a bit to get to larger V and g_{00} but not so high that the increase in velocity is too big. That's just another way of saying it goes up a ways and then comes down. It appears that we might be on the right track.

It's nice to reproduce things we already know, but it's nicer to predict something new. Einstein gave us the differential equations that determine the metric from a given distribution of mass. They are quite complicated and not yet fully understood. The reason they are complicated as that in a certain rough sense the gravitational field serves as a source for itself. So far, we have seen that mass changes the metric. Also we know that mass is just a form of energy, and that mass can be transformed to other forms of energy. Thus, it could not be consistent for it to really be mass that determines the metric, it must be energy. For slowly moving massive objects, most of the energy is in rest mass, and we easily miss the contribution from the kinetic energy. Speaking very roughly, the distortion of space that is caused by the presence of energy contains energy itself. Thus the curvature of spacetime becomes the cause of further curvature, and the equations become very nonlinear.

There are a few cases where relatively simple, exact solutions have been found. One of them is the problem at hand: a static, spherically mass. The result for radial motion outside a mass M is

$$ds^2 = \left(1 - \frac{2GM}{c^2 r}\right)dt^2 - \frac{1}{1 - \frac{2GM}{c^2 r}}\frac{dr^2}{c^2} \quad .$$

This is called the Schwarzschild solution. Notice that something odd is happening at

$$r = r_S \equiv \frac{2GM}{c^2}$$

which is called the Schwarzschild radius. In spite of the relative simplicity of this metric, it is still quite complicated to figure out what is really happening at r_S. For example, note that outside r_S , t is a timelike coordinate (dt^2 makes a positive contribution to ds^2) and r is a spacelike coordinate (dr^2 makes a negative contribution to ds^2) while inside r_S the roles are reversed. It turns out that what we have here is a black hole. Massive particles and light rays that cross from $r>r_S$ to $r<r_S$ can never return to the region outside r_S. The fall from $r>r_S$ to $r=0$ happens in finite proper time for the particle. However, for an observer at some fixed $r>r_S$, watching the process, the fall to r_S of the particle takes infinite time.

There is now quite a bit of observational evidence that black holes exist. They come in two basic forms. There are those that form at the end of the life of a massive star. Eventually it burns up all its nuclear fuel, and there is nothing left to resist the gravitational collapse al the way to a black hole. The other type is supermassive black holes at the centers of galaxies. If fact there is probably one at the center of our Milky Way. In either case, matter falling into the black hole is heated and it radiates and we can detect the radiation emitted before it crosses into the region $r<r_S$. Also the curvature of space around the black hole effects the orbits of nearby objects in ways that we can observe and deduce that they are probably caused by a black hole.