432018 PHILOSOPHY OF PHYSICS (Spring 2002)

Lecture 15: Gravity and the curvature of space-time

Preliminary reading: Sklar, pp. 40-52.

We have seen that Einstein proposed the Special Theory of Relativity in order to account for the consequences of Maxwell's theory of electromagnetism. Specifically, in order to account for the fact that the speed of light should be the same for all observers at rest in an inertial frame. Indeed, the ramifications of Einstein's theory are far reaching, leading as they do to a revision of our understanding of space and time.

We now turn to Einstein's *General* Theory of Relativity. This theory is motivated by Einstein's desire to give an account of gravitational phenomena which is compatible with the revisions in our notions of space and time forced on us by his Special Theory of Relativity. And, as we shall see, this new theory introduces further revisions in our understanding of space and time. In particular, we shall start by considering a thought experiment that Einstein proposed to motivate his theory and then briefly discuss the nature of non-Euclidean geometries which provide the natural mathematical framework for his new theory. Then we shall summarise the most important aspects of his new theory.[1]

1 Gravity and relativity

To motivate Einstein's *General* Theory of Relativity, we start by examining the claims made by Newton's Gravitational Theory and the problems that this has in light of our discussion of the Special Theory of Relativity. This will lead us to consider a thought experiment which Einstein used to motivate the *equivalence principle* which is the basis of his new theory.

Newton's Gravitational Theory

In Newton's Gravitational Theory, gravity is a force which has three main properties:

- It acts on material bodies *at a distance*, i.e. gravity is a causal interaction that affects such bodies even when they are spatially separated.
- This causal interaction propagates between material bodies *instantaneously*, i.e. a change in one such body has an effect on the other bodies at the *same time* regardless of their separation. In particular, this means that the 'signals' that carry this interaction must travel between the bodies instantaneously (i.e. they take no time to traverse the distance separating them), and so such 'signals' must travel at an infinite speed.
- The *strength* of this interaction depends only on the [gravitational] masses of the bodies involved and the [inverse square of the] distance between them. In particular, the strength of the interaction is independent of the size and constitution of the bodies since it depends only on their [gravitational] mass.

An interesting consequence of Newton's Gravitational Theory in relation to his laws of motion is that there are, *prima facie*, two notions of mass at work here. To see this, let's consider a simple example:

> Suppose that I let go of a ball which I am holding a distance d above the surface of the Earth. The force that acts on this ball and which causes it to fall towards the ground is the gravitational force, due to the mass of the Earth, acting on the ball. Newton's theory tells us that:

[1] At least, as far as this course is concerned.

- this force acting on the ball is directed towards the centre of the Earth and has a magnitude given by

$$F = G\frac{Mm_g}{(R+d)^2},$$

where R and M denote the radius and mass of the Earth respectively and G is Newton's gravitational constant. Although, the important thing to note here is that m_g is the *gravitational* mass of the ball, i.e. the quantity that determines how strong the gravitational interaction between the Earth and *this* ball is.

However, Newton's second law tells us that given this information about the force, we can calculate the acceleration of the ball as it falls to the ground. That is:

- this force acting on the ball and directed towards the centre of the Earth with a magnitude given by F above causes the ball to accelerate towards the ground in accordance with the formula

$$F = m_i a,$$

where a is the acceleration of the ball and m_i is the *inertial* mass of the ball, i.e. the quantity that determines how much force is needed to give rise to a certain acceleration.[2]

Since the force involved in both of these equations is the same, this gives us

$$m_i a = G\frac{Mm_g}{(R+d)^2},$$

and due to the size of the Earth (i.e. $R \gg d$), we can approximate the quantity $GM/(R+d)^2$ by a number g which is effectively constant with respect to d, the distance of the ball from the ground.[3] Doing this, we get:

$$m_i a = m_g g,$$

Now, the quantity g is called the 'acceleration due to gravity [near the surface of the Earth]' and observations dictate that this is equal to the acceleration of the ball, i.e. a. But, this can only be true if m_i and m_g are *equal*.

Thus, for Newton's Mechanics and Gravitational Theory to match up so that they agree with experiment, these two masses which measure *prima facie* different quantities, i.e. the inertia of a body (m_i) and how strongly it interacts with other bodies *via* gravity (m_g), must be the same. This is easily accommodated within the theory by just saying that they are equal and the quantity in question is just the mass *simpliciter* of the object. But, although it can be *accommodated* in this way, within Newtonian physics there is no *explanation* of why these two different quantities should be the same. As such, within the Newtonian theory, this equality is just a coincidence, there is no deeper reason why it must hold.

The incompatibility of Newton's theory and STR

Obviously, from what we have seen above, the biggest problem for Newton's Gravitational Theory in the context of STR is the fact that it requires the gravitational interaction to propagate instantaneously. That is, Newton requires the gravitational causal 'signals' to propagate with an infinite speed which is manifestly incompatible with the fact that no 'signal' can travel faster than the speed of light in STR.

[2] More technically, m_i is a measure of the *inertia* of a body (i.e. its 'reluctance to move)' when a force is applied. For example, if you apply the same force to two bodies and find that it produces accelerations a and $2a$ in the first and second bodies respectively, then the first has twice as much inertia (inertial mass) than the second. That is, you get 'less motion' since it has 'more inertia'.

[3] As such, this means that the gravitational field is effectively *uniform* near to the surface of the Earth.

Einstein's motivation for GTR

However, this incompatibility was not the big problem for Einstein. By all accounts, he was more concerned by the fact that the equality of inertial and gravitational mass was just an accident in Newton's theory. In order to explain this equality, he postulates the following principle:

> **The Equivalence Principle:** In a small region of space-time (i.e. locally), it is not possible to experimentally distinguish between a frame 'at rest' in a uniform gravitational field and a frame being uniformly accelerated through empty space.[4]

This principle is motivated by a thought experiment concerning two observers 'trapped' in elevators, as illustrated in Figure 1. Now consider that:

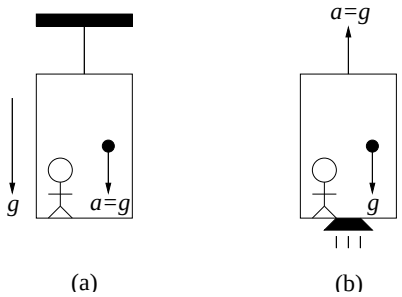

Figure 1: *Motivating the equivalence principle. (a) An elevator at rest in the uniform gravitational field of the Earth. (b) An elevator which is being uniformly accelerated through empty space.*

- In Figure 1 (a), an observer is at rest in an elevator that is itself at rest in the uniform gravitational field of the Earth. If he releases a ball, it falls to the floor with an acceleration given by g, the acceleration due to gravity mentioned above.
- In Figure 1 (b), an observer is at rest in an elevator which has an acceleration of g in the 'upwards' direction as it travels through empty space[5] If he too releases a ball, it will also fall to the floor with an acceleration given by g.

Thus, if both elevators were such that the observers couldn't see anything external to them, the observer couldn't use measurements of the ball's acceleration to decide which of these two situations he was in. That is, these two situations are empirically indistinguishable *vis-a-vis* measurements of the acceleration of a body.

The equivalence principle has two important consequences, firstly:

1. If all bodies experience one common acceleration g in (b), then m_i and m_g *must* be equal.

   This is because, if this was *not* the case, all bodies would accelerate with a common value g relative to the elevator in (b), but they would not do so in (a) since their accelerations a (given by $m_i a = m_g g$) would vary depending on the discrepancy between m_i and m_g. As such, these situations would be distinguishable and this contradicts the equivalence principle!

Thus, the equivalence principle *explains* why the gravitational and inertial masses of a body should be equal. Secondly, there is also a surprising prediction about the behaviour of light:

2. Light is affected by gravitational forces.

   To see why, consider the following extension of the above thought experiment, which is illustrated in Figure 2:

   - In Figure 2 (a), we have an elevator which has an acceleration of g in the 'upwards' direction as it travels through empty space. At some point a burst of light is emitted from a source located at a point A within the elevator and this travels towards the other side of the elevator. Now consider what is seen from the perspective of two observers O and O^*:

[4] That is, *empty* in the sense that there are no masses around to 'generate' gravitational forces.
[5] That is, there is no gravitational force acting in the region occupied by the accelerating elevator.

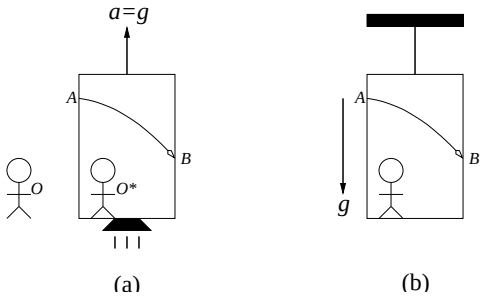

Figure 2: *A consequence of the equivalence principle — light is affected by gravitational fields. (a) The path of a light beam as seen in an elevator which is being accelerated through empty space. (a) The path of a light beam as seen in an elevator at rest in the uniform gravitational field of the Earth.*

  - An observer O who is 'at rest' in the empty space outside the elevator sees a burst of light being emitted from a source at point A within the elevator travelling towards the other side of the elevator. As this observer watches this burst of light, he will see it travel in a straight line through space, whilst at the same time, he will see the floor of the elevator move 'upwards' due to the acceleration of the elevator. Thus, as a consequence of this, he observes that the light hits the opposite wall of the elevator at a point B which is lower than A relative to the elevator.
  - As such, for an observer O^* who is inside the elevator and at rest relative to it, the burst of light appears to have 'dropped', i.e. followed a path which is *not* a straight line, since it hits the opposite wall at a point B which is lower than the point A relative to the elevator.

  This is a bit strange in itself, it appears that light isn't always observed to travel in straight lines!

- In Figure 2 (b), we have an observer who is at rest in an elevator that is itself at rest in the uniform gravitational field of the Earth. Now if he can only make observations concerning what is happening within the elevator, by the equivalence principle, he should not be able to decide whether the elevator is at rest in a uniform gravitational field or whether it is being uniformly accelerated relative to empty space. As such, if he was to observe a burst of light emitted from a source located at a point A within the elevator, he too *must* observe that, when this reaches the other side of the elevator, it has 'dropped' so that it arrives at the point B as before. That is, light moving in a uniform gravitational field must deviate from a straight line path!

In particular, the fact that light is being affected by gravitational fields in the same way as material bodies[6] is completely at odds with the 'classical' view. Indeed, for Newtonians, light has *no* mass and so there can be *no* gravitational force acting on it!

So light, which is classically treated as *massless*, now appears to act as if it has a mass! How can Einstein account for this? To see how this leads to GTR, we have to briefly consider the nature of non-Euclidean geometries.

2 Non-Euclidean geometry

Euclidean geometry, as put forward by Euclid, is an axiomatic system based on 'self-evident' postulates and axioms. The results that can be proved in this system are those that you are no doubt familiar with from the geometry you studied at school. For example, we have results like:

1. The shortest distance between two points is a straight line.
2. The sum of the internal angles of a triangle is equal to two right angles (i.e. 180°).

[6] Compare the shape we see with the parabolic path of a material body undergoing projectile motion!

However, there is one postulate of Euclidean geometry, 'the parallel postulate', that for many years was considered to be less 'evident' than the others. This postulate, in effect, states that:

> Given a straight line and a point not on that line, there is only one straight line that passes through the point parallel to the line.

where two lines are *parallel* if they fail to intersect, no matter how far they are extended. Firstly notice that this postulate is not as evident as the other postulates, for example the postulates which stipulate that:

- equal quantities added to equal quantities give equal quantities, and
- two points determine a straight line that joins them.

But, secondly, notice that we need this postulate to derive result (2.) above since, as shown in Figure 3, we need to be able to construct the parallel line CD in order to show that the angle $\alpha + \beta + \gamma$ is

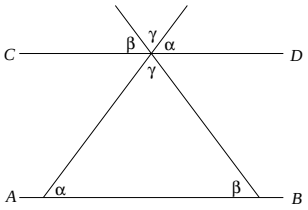

Figure 3: *In order to show that the sum of the internal angles of a triangle is equal to* 180° *we need to be able to construct a parallel line (i.e. CD) which is parallel to the line AB and goes through the apex of the triangle.*

the angle subtended by this straight line (i.e. 180°).

For a long time it was believed that Euclidean geometry was the *only* consistent geometry. That is, any other geometry which we may consider would ultimately be shown to be inconsistent and as such would be logically impossible. Thus, we had the view that geometry was *necessarily* Euclidean. Indeed, such ideas were 'backed up' by the view that geometry is the abstract study of the structure of space. And, since space appears to have the structure of Euclidean geometry, there is clearly no point in considering geometries that are not Euclidean.

Although, the fact that the parallel postulate was not as 'self-evident' as the other postulates still caused concern. As such, mathematicians tried to show that it was, in fact, not an independent postulate but a consequence of the other postulates. Clearly, if this could be shown, then the parallel postulate could be taken as a theorem arising from the 'self-evident' postulates, as opposed to being a postulate in its own right. However, how could they show that it is not independent of the other postulates? Well, they used the idea that, if you denied the parallel postulate and then found that this denial contradicted the results of the other postulates then, by *reductio ad absurdum*, the parallel postulate must be a consequence of these other postulates.[7] But, there are two ways of denying the parallel postulate, namely, you can insist that:

- there will be *no* such parallel lines.
- there will be more than one such parallel line.

and both of these avenues were pursued in order to establish the desired contradictions.

However, no contradiction (as such) was found and, in fact, it was discovered that the denial of the parallel postulate gave rise to *alternative* geometries. Geometries which, by denying the parallel postulate ceased to be Euclidean, but were consistent nonetheless. Thus, there were now two new *non-Euclidean* geometries available, Riemannian geometry (no parallel lines) and Lobachevskian

[7] That is, let Φ be a set of propositions containing these 'other' postulates and let p be the parallel postulate. If one can show that $\Phi \cup \{\neg p\} \vdash$ (i.e. the postulates in Φ and the parallel postulate lead to a contradiction), then $\Phi \vdash p$ (i.e. these postulates entail the parallel postulate) by *reductio*. (Although, technically, this inference uses the principle of indirect proof, not *reductio*!)

geometry (more than one parallel line). And, furthermore, in these geometries we get theorems which [unsurprisingly] tell us different things about geometric figures.

The most relevant way of thinking about the differences between these geometries is to think of it in terms of the work done by the mathematicians Gauss and Riemann. Here, the differences arise because these geometries can be thought of as the geometry that should be associated with two-dimensional figures in spaces that have different curvatures. For example, let's consider the case of Euclidean geometry and Riemannian geometry:

- Euclidean geometry is the geometry associated with a *flat* (i.e. zero curvature) space. Within this space, as we have seen,
  - the shortest distance between two points is given by a straight line.
  - the sum of the internal angles of a triangle is equal to 180°.

  and, incidentally, recall that in Minkowski space-time, Euclidean geometry describes the spatial structure of the planes of simultaneity of a given inertial observer.
- Riemannian geometry is the geometry associated with a space of positive curvature. For example, the surface of a sphere is a two-dimensional positively curved space (say the surface of the Earth as seen from the perspective of a two-dimensional creature that lives on that surface). Within this space, as we can see that
  - the shortest distance between two points is given by a *great circle* and this is the 'closest' thing we have to a straight line in such a space. (For example, lines of longitude or the equator.)
  - the sum of the internal angles of a triangle is *greater than* 180° where here, a triangle is a shape which is made out of three intersecting great circles.

    To see this, consider the triangular path you follow when, starting at the North Pole, you travel south down a line of longitude until you reach the equator, then turning East, you travel along the equator for a while until, turning North, you head back to the North Pole along another line of longitude. Clearly, you have traced out a triangle in the above sense since you have been travelling along great circles on the spherical surface of the Earth, but during your journey you have turned through 180°[8] and there is also the angle between the two [travelled] lines of longitude which meet at the North Pole. That is, on this trip you have followed a triangular path that has internal angles which sum up to *more than* 180°.

As such, we can see that there is a close connection between the curvature of a space and the geometry which should be used to describe it. In particular, we have the following definition,

> A *geodesic*, is a curve in space such that the distance between two points on this curve is the shortest possible distance between those two points within the space.

So, for example, straight lines are geodesics in Euclidean geometry (i.e. on a plane) and great circles are geodesics in a Riemannian geometry (e.g. on the surface of a sphere).[9]

3 The General Theory of Relativity

Given the equivalence principle, it is plausible to assume that gravity affects all objects *in the same way* regardless of their size or constitution.[10] But, we also know that material objects which travel

[8] That is, through 90° when you turned East on arriving at the equator and through another 90° when you turned North on leaving the equator.

[9] This raises an interesting question: If we were two-dimensional beings who lived in a two-dimensional space, could we determine its curvature? The answer is yes! We can measure the *intrinsic* curvature from within the space by using parallel transport. But, we shall not discuss this here.

[10] Since, otherwise, we could use objects of different sizes or constitutions to detect whether we are in a state of uniform accelerated motion or at rest in a uniform gravitational field.

with uniform velocities (i.e. at constant speeds in a fixed direction) in the absence of gravity or other forces, should follow different (i.e. curved) paths in the presence of a gravity. However now, due to the equivalence principle, the change of path can depend only on the initial *position* and *velocity* of the object in question since it does not depend on its size or constitution. Consequently, instead of thinking

- material objects follow a curved path in a Euclidean space due to the action of gravitational forces,

we can think

- material objects follow a curved path in a non-Euclidean space and there are no gravitational forces,

the idea being that we can take the curved path to be a manifestation of the curvature of space which is, in turn, *related* to something which we once thought of as a gravitational field. That is, we *geometrise* the gravitational field, replacing a 'geometry plus a force' picture with a new geometry that [somehow] incorporates the affects of the force.

So, basically, in the General Theory of Relativity, Einstein posits a *curved* space-time in which 'free' particles[11] and light rays follow the [timelike] geodesics of this space-time. And, in general, although these paths won't be straight lines as we understand them in the Euclidean sense, they will be the curves of 'shortest distance' or 'least curvature' in this space-time. As such, one should be able to determine the geometry of the space-time we inhabit by tracing out the paths of 'free' particles and light rays. That is, the particles and light rays which are only under the influence of gravity, now taken not as a force, but the curvature of space-time.

So, the only question we may have is 'How are the gravity-as-curvature and gravity-as-force pictures physically related?'. Well, recall that in the Newtonian Gravitational Theory:

- the gravitational force *acts* so as to accelerate all massive objects.
- massive particles are the *source* of the gravitational force.

whereas in Einstein's General Theory of Relativity we have:

- gravity is represented by the curvature of space-time, it *affects* 'free' particles and light rays in the sense that they travel along geodesic paths in this space-time.
- the curvature of space-time is *determined* by the mass[12] distribution of the world.

But, the equations of the General Theory of Relativity only tell us whether a given space-time is *compatible* with a given mass distribution. And, if this *is* the case, then this space-time/mass-distribution pair give us a 'possible world' of the theory. In particular, the fact that this is only a relation of 'compatibility' indicates that we should *not* think of the mass-distribution (i.e. matter) as *causing* the curvature.

James Ward (e-mail: `j.m.ward@lse.ac.uk`)

[11] That is, particles which are *not* acted on by [non-gravitational] forces.
[12] Strictly speaking, the non-gravitational mass-energy.