General Relativity

Joe Keir

Please send comments/corrections to Joseph.Keir@maths.ox.ac.uk

Chapter 1

A brief history of spacetime

> The observed facts about earth are not only that it remains at the centre, but also that it moves to the centre. The place to which any fragment of earth moves must necessarily be the place to which the whole moves; and in the place to which a thing naturally moves, it will naturally rest.
>
> Aristotle, *De Caelo*, translated by J. L. Stocks.

Before the development of General Relativity, 'space', 'time', and later (after special relativity) 'spacetime', were simply the background on which physics took place. They were fairly simple and uninteresting, although some of their properties did play an important role in the foundations of physics. Nevertheless, aside from foundational considerations, these structures were not central objects of study in physics: they formed a part of the necessary substrata on which a theory of physics could be built, but were not particularly interesting in themselves.

All this changed dramatically with General Relativity. As we will see in this course, in GR spacetime is no longer flat and featureless, but curved in interesting and sometimes extreme ways. This curvature turns out to be what we experience as gravity. Moreover, spacetime is no longer static but *dynamical*; no longer playing just a background role but instead taking a full part in the dynamical evolution of physical systems.

Before we get to GR, though, let us take a brief look at some of the different historical notions of space and time, and how this underlying structure informed the physical theories which were built on top of it. Thinking precisely about these structures and what they mean – rather than letting them fade into the background – will be helpful when it comes time to formulate GR[1].

1.1 "Aristotelian" spacetime

This is the most 'common sense' view of space and time. Something of this sort was held by some ancient philosophers[2] (see the quote at the start of this chapter), and is still common among children, though in both cases without the level of mathematical sophistication we will give!

[1] The presentation of the 'history' below is extremely anachronistic: I will present historical ideas in modern notation and language. Some of the terms used are also rather idiosyncratic – the term "relativity" is usually reserved for the Galilean and subsequent views, even though there is some kind of 'relativity' present in even earlier theories.

[2] Actually, Aristotle's view differed from the point of view put forward here in two important respects: first, he did not believe in a finite past, and second, he believed that the universe is finite in spatial extent (erroneously believing that one could not make sense of an "origin" or centre in an infinite space), so that spacetime is really more like $\mathbb{E} \times B$, with B a ball of some large radius, and $\mathbb{E}$ the one-dimensional Euclidean space (see section 1.2). Perhaps the spacetimes described in this section should instead be called "Thomist", after Thomas Aquinas, who modified Aristotle's arguments to fit better with Christianity, in particular inserting a finite past.

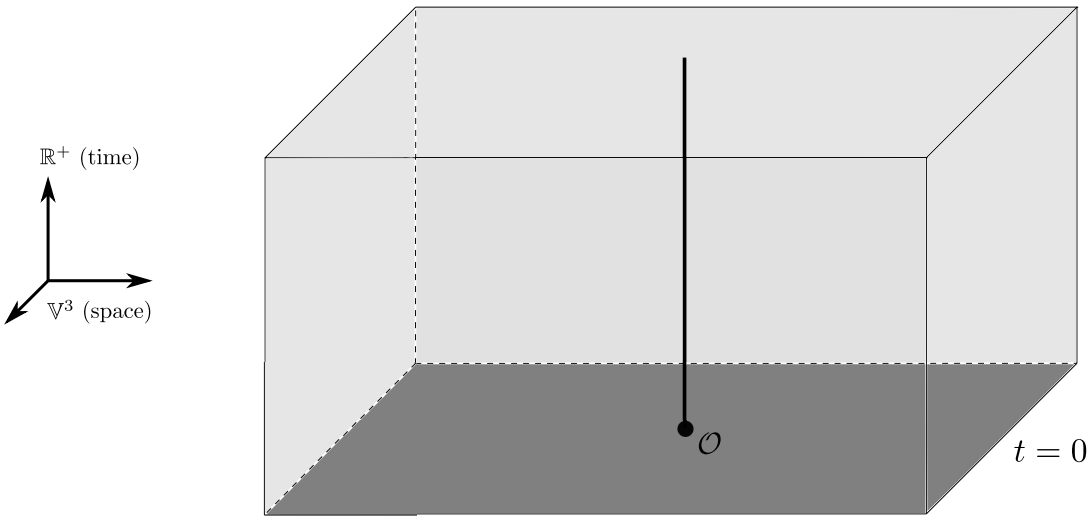

Figure 1.1: A sketch of "Aristotelian" spacetime, suppressing one spatial direction. Time goes up the page, and "space at a given time" is represented by a horizontal slice through spacetime. Space extends infinitely in all directions, while time extends to the future of (or "above", in the sketch) the plane $t = 0$. $\mathcal{O}$ is the origin of space and time, and the the origin of $\mathbb{V}^3$ at various times, i.e. the curve $(t, 0)$, is shown as a solid black straight line going up the page, emanating from $\mathcal{O}$.

In this point of view, spacetime is simply $\mathbb{R}^+ \times \mathbb{V}^3$, with times t taking values in $\mathbb{R}^+$ and points p in space taking values in a 3-dimensional vector space $\mathbb{V}^3$. $\mathbb{V}^3$ is to be equipped with a positive definite inner product (the 'dot product', $\boldsymbol{X} \cdot \boldsymbol{Y}$), which can be used to measure distances and angles. The origin of $\mathbb{V}^3$ specifies some special place – presumably the centre of the earth – and the origin of $\mathbb{R}^+$ specifies some special time – perhaps the 'moment of creation'. If we choose a basis for $\mathbb{V}^3$ that diagonalizes the dot product, then $\mathbb{V}^3$ can be identified with $\mathbb{R}^3$, and the dot product is the usual vector dot product. The choice of such a basis is unique up to rotations.

"Aristotelian" relativity

Physical laws should "respect" this background structure, meaning that they should be invariant under changes that leave this structure intact. In this case, these changes are simply rotations of $\mathbb{V}^3$, which leave the dot product invariant. In other words, rotations about the centre of the Earth (which is obviously spherical – we're not *that* primitive!) should not change the laws of physics.

Problems with Aristotelian spacetime

With this view of spacetime, physical laws should be the same at different places on the surface of the earth, which is good. However, there is no reason why physics should be the same if we perform other transformations, such as rotating about some other point. Physical laws could (and in general *should*) depend on the distance away from the centre of the earth, and the time that has passed since $t = 0$ – these are invariant under the allowed transformations. At first, this might seem like a good thing: Aristotle used this idea to argue that the physical laws cause the element 'earth' to fall towards the centre of the universe. However, from this point of view there's no real reason why the physical laws at one time or altitude should be anything like the physical laws at some other time or altitude, so it would be very hard to build a physical theory with any predictive powers on these foundations. Of course, such a theory becomes untenable once it is appreciated that the Earth (or even the sun) is not the centre of the universe!

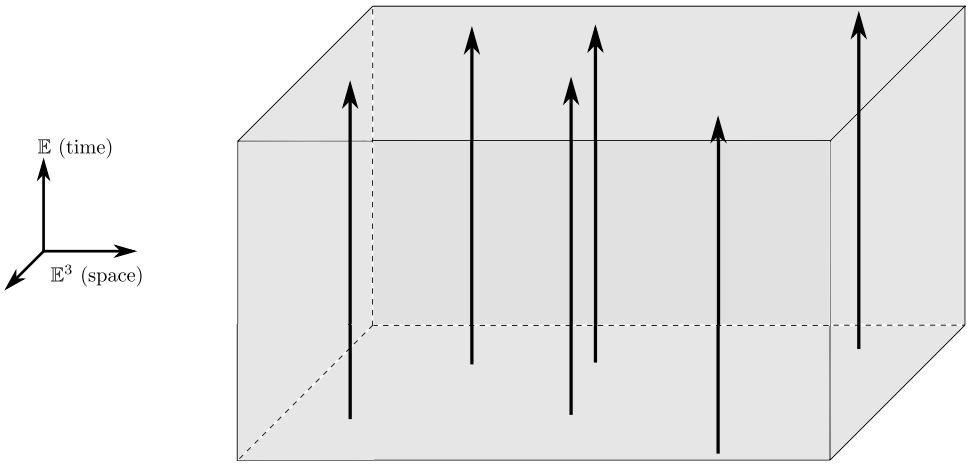

Figure 1.2: A sketch of "Atomist" spacetime. Unlike "Aristotelian" spacetime, there is no special origin of time or space, both of which extend infinitely in all directions. The paths of several particles at rest are shown as solid black arrows.

1.2 "Atomist" spacetime

This is a slightly more sophisticated version of the 'common sense' view of space and time. The ancient atomist philosophers, for example, believed in an infinite universe with no central point[3]. This formulation harmonizes with this viewpoint, while also making use of the more advanced Euclidean geometry.

Atomist spacetime is modelled on $\mathbb{E} \times \mathbb{E}^3$, where $\mathbb{E}^n$ is the n-dimensional *Euclidean* space. By this we mean an affine space[4] such that the associated (real) vector space comes equipped with a positive definite inner product (the 'dot product', $X \cdot Y$). In other words, we have the same structures as in the Aristotelian spacetime, but we *forget about the origin.*

"Atomist relativity"

Again, physical laws should respect the background structure. This time, the transformations leaving the structure invariant are rotations (about *any* point), together with translations in space or in time. In other words, physical laws should be the same no matter where you are in time or space, and no matter in which direction you are facing.

Problems with Atomist spacetime

This is a big improvement over the Aristotelian spacetime considered above: we no longer have a preferred centre of space or start of time, which is a much better foundation on which to build a predictive theory. Despite these nice properties, there is still a notion of *absolute space*, and a restrictive notion of objects being *at rest*. As Galileo would later show, these notions are incompatible with observations.

How do the relatively flexible structures of affine space give rise to these rigid structures? First, we can define *absolute space* $\mathbb{E}^3$ using the projection

$$\begin{aligned} \mathbb{E} \times \mathbb{E}^3 &\to \mathbb{E}^3 \\ (t, x) &\mapsto x \end{aligned}$$

so, although there is no distinguished origin to space, there is a unique point in "absolute space" (thought

[3] This viewpoint is also similar to the one put forward by the pantheist occultist Giordano Bruno in support of Copernicus.

[4] An affine space can be thought of as just a vector space where we forget the special role played by the origin. More concretely, for each pair of points in the affine space there is a unique vector - the "difference" between the pair of points - in an associated vector space.

of as Euclidean space $\mathbb{E}^3$) associated with every point in spacetime.

Next, consider a path through spacetime:

$$\begin{aligned}\gamma : \mathbb{R} &\to \mathbb{E} \times \mathbb{E}^3 \\ (s) &\mapsto (s, p(s))\end{aligned}$$

where we have chosen some (arbitrary up to translations) identification of $\mathbb{E}$ with $\mathbb{R}$. Such a path through spacetime is called a *worldline*: it could represent, for example, the path taken by a point particle, which is at position $p(s)$ at time $t(s)$.

Now, this worldline is said to be the worldline of a particle *at rest* if, for all s, $\tilde{s} \in [0, 1]$, we have $p(s) - p(\tilde{s}) = 0$. Recall that, for two points in the affine space $\mathbb{E}^3$, their difference defines a vector in $\mathbb{R}^3$. Moreover, the transformations allowed by atomist relativity – translations and rotations – induce corresponding transformations on $\mathbb{R}^3$, which in this case are just the rotations. So, properties of these vectors in $\mathbb{R}^3$ are allowed to have a physical meaning *if they are invariant under rotations* – in other words, the *length* of such a vector can have a physical meaning. In particular, the zero vector is invariant under rotations. Hence, the equation $p(s) - p(\tilde{s}) = 0$ defines a physically distinguished class of worldlines on Atomist spacetime (see figure 1.1). Given one worldline of a particle at rest, it is fairly easy to see that all the other such worldlines are given by translations of this one reference worldline.

Theories of physics built on these foundations were fairly successful: for example, the earth can be considered "at rest", and then it is natural to think that other objects will tend to follow other worldlines which are at rest, at least in the absence of external factors (such as forces). This matches observations fairly well, however, there are several difficulties with such a proposal: for example, it is very difficult to explain projectile motion in this framework. Projectiles appear not to be being affected by any external factors[5] and yet they are clearly not moving along world lines which are at rest!

1.3 Galilean spacetime

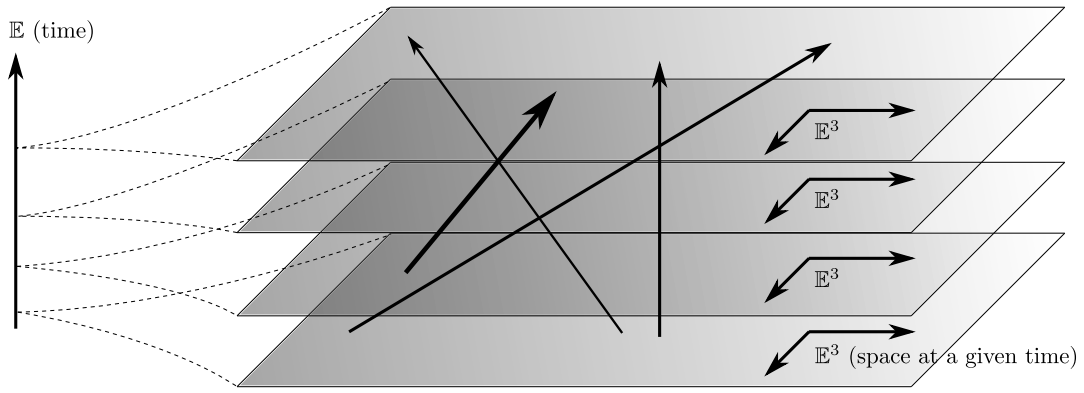

Figure 1.3: A sketch of Galilean spacetime. To each point in time we associate an $\mathbb{E}^3$, which corresponds to space at that time. Unlike in the previous examples, there is no preferred identification of the different $\mathbb{E}^3$'s, leading to the absence of *absolute space*. However, there is a special family of curves through spacetime, which are the *worldlines of inertial observers*, some of which are shown as arrows through spacetime in the figure

After rolling some balls around on inclined planes, Galileo recognised the problems with models built on absolute space. In a modern formulation the alternative model of spacetime proposed by Galileo takes the form of a *fibre bundle*, with base space $\mathbb{E}$ and fibres $\mathbb{E}^3$. If you're not familiar with fibre bundles, think of the line $\mathbb{E}$, and, above every point on this line, place an $\mathbb{E}^3$ (see figure 1.3). This differs from

[5]Here we are ignoring air resistance, but this can be justified by taking projectiles with very large mass so that the air resistance is small.

$\mathbb{E}^1 \times \mathbb{E}^3$ in that there is no preferred identification of the different $\mathbb{E}^3$'s that are situated 'above' different times – in other words, there is no *absolute space*.

We need a bit more structure than this: clearly space at one time and space at another time are not completely unrelated to one another! Mathematically, the extra structure we need is a *connection* on this fibre bundle, which allows us to identify a special class of "straight" worldlines on the fibre bundle. In Galilean spacetime, these worldlines are all related to one another by *Galilean transformations*, discussed below.

Galilean relativity

In a Galilean spacetime, the distinguished class of worldlines are said to be the worldlines of *inertial observers*. Suppose that we have one such worldline: we can use this to identify the different $\mathbb{E}^3$'s by using coordinates on each $\mathbb{E}^3$ so that this worldline passes through the origin[6]. In terms of these coordinates, other worldlines of inertial observers can be found by means of *Galilean transformations*:

$$\begin{aligned} t &\mapsto t + t_0 \\ x^i &\mapsto R^i{}_j x^j + v^i t + (x_0)^i. \end{aligned}$$

Here $R^i{}_j$ is a special orthogonal matrix, t_0 is a constant scalar and v^i and $(x_0)^i$ are some constant vectors. Note that *repeated indices are summed over*.

As usual, physical theories should respect this background structure, which means that the physical laws should be invariant under these Galilean transformations. Famously, this *principle of Galilean relativity* is obeyed by Newton's equations $F = ma$: the acceleration is not quite invariant (under a Galilean transformation $a^i \mapsto R^i{}_j a^j$), but it does transform like a vector, and so it is able to play a physical role, as long as we understand that the force F should also transform like a vector.

Problems with Galilean spacetime

This conception of spacetime was tremendously successful and underpins Newtonian mechanics. It is still the view held by most people who have never learned relativity. Nevertheless, it has a few problems, the main one being the constancy of the speed of light.

Under a Galilean transformation, a velocity V will transform as

$$V^i \mapsto R^i{}_j V^j + v^i.$$

In particular, the magnitude of a velocity is *not invariant* under Galilean transformations, and in fact, given a velocity V there is always some transformation setting V to zero. The speed of light, therefore, should be different when measured by different inertial observers. On the other hand, Maxwell's equations predict just one speed of light, independent of the observer, and this was eventually confirmed by experiment.

Note that the magnitude of the difference between two velocities (or *relative velocity*) *is* invariant under Galilean transformations. A popular solution to the problem of the constancy of the speed of light was therefore to suppose the existence of "the aether" through which light propagates. The velocity c, derived from Maxwell's equations, was then supposed to be just the relative velocity of light waves through the aether. However, all attempts to detect the aether failed, most famously in the *Michelson-Morley experiment*, which showed that the speed to light is the same in all directions.

There is also the issue of *absolute time*. With the Galilean viewpoint, we have successfully done away with the notion of absolute space, but there is still a notion of absolute time. This can be defined by using the canonical projection from the fibre bundle to $\mathbb{E}$, which can be viewed as "collapsing" each of the $\mathbb{E}^3$'s onto the point along the line $\mathbb{E}$ that it sits above (see figure 1.3). Using this, each "event" (that is, each point in spacetime) can be assigned a unique "time" in $\mathbb{E}$. In particular, this gives a unique ordering

[6]Technically, we also need to ensure that the $\mathbb{E}^3$'s don't rotate relative to one another. This can be done by choosing a basis for one of the $\mathbb{E}^3$'s, and then using the connection to *parallel transport* this basis onto the other $\mathbb{E}^3$'s– see chapter 4

on events: certain events happen before others, and all inertial observers will agree on what precedes what. However, as Einstein showed (in thought experiments involving lightning bolts hitting trains), there can be instances where different inertial observers will *disagree* about the ordering of events: this is the famous *relativity of simultaneity*.

1.4 Minkowski spacetime

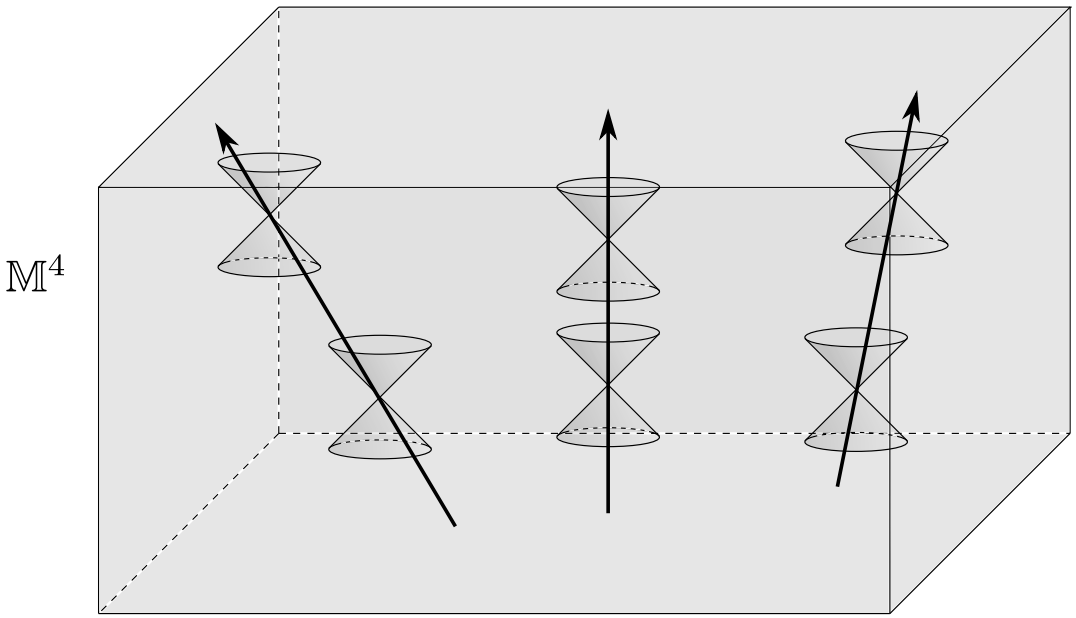

Figure 1.4: A sketch of Minkowski spacetime $\mathbb{M}^4$. Space and time are no longer separate entities, but are merged into the single object spacetime. There is no longer a (unique) notion of space at a given time: horizontal slices are space at a given time *according to an inertial observer moving vertically up the page*, but other observers will split space and time differently. The *light cones* (or *null cones*) are invariant under Lorentz transformations, so all observers will agree on them - some of these are shown in the sketch. Worldlines of inertial observers are straight lines and (if the observers have nonzero mass) pass through the interior of the light cones.

Up until this point, 'space' and 'time' have actually been quite distinct, playing distinct and separate roles in the formulations above. The first true 'spacetime', in which space and time are combined and unified in a natural way, was put forward by Einstein in his special theory of relativity, and subsequently geometrised by Minkowksi.

According to this view, spacetime is to be identified with $\mathbb{M}^4$, which is a four dimensional affine space. Unlike the Euclidean spaces $\mathbb{E}^n$ of the previous examples, this doesn't come equipped with a positive definite inner product, but instead it comes equipped with an *indefinite* (but *non-degenerate*) quadratic form m with Lorentzian signature $(-,+,+,+)$.

Special relativity

This spacetime structure constructed above is invariant under *Poincaré transformations*. If we choose coordinates x^a, $a = 0,1,2,3$ for the affine space $\mathbb{M}^4$ (i.e. we make an arbitrary choice of origin) then these transformations are given by

$$x^a \mapsto \Lambda_b{}^a x^b + y^a$$

where Λ is a matrix in $SO(1,3)$, whose action is called a *Lorentz transformation*, and y^a is a fixed spacetime vector, corresponding to space and time translations.

The set of timelike[7]. straight lines in Minkowski space defines inertial observers in this view of spacetime. Note that a Poincaré transformation takes straight lines to straight lines, therefore the property of *being an inertial observer* can have physical content.

Physical theories should be invariant under Poincaré transformations: this is the *principle of relativity*. In practice, this means that physical theories can be of the form: "For an inertial observer, the physical laws are…". We will review special relativity in more detail in chapter 3.

Problems with Minkowski spacetime

Unlike the previous views of spacetime, which were maintained for hundreds or even thousands of years, Minkowski spacetime and special relativity held the crown as the state-of-the-art view of spacetime for a mere ten years. This is not because of their lack of success, but rather because of the rapid progress made by Einstein, who was able to quickly supersede his own theory!

The main problem (in fact, the *only* real problem) with this view of spacetime is that it fails to incorporate gravity. As we shall see in the next chapter, Newtonian gravity is incompatible with special relativity, so Einstein set about trying to remedy this. In the process, not only did he produce a new theory of gravity - which remains the current best theory of gravity, after more than 100 years - but he revolutionised our conception of spacetime for a second time.

[7]See chapter 3

Chapter 2

Newtonian gravity

> The history of the apple is too absurd. Whether the apple fell or not, how can any one believe that such a discovery could in that way be accelerated or retarded? Undoubtedly, the occurrence was something of this sort. There comes to Newton a stupid, importunate man, who asks him how he hit upon his great discovery. When Newton had convinced himself what a noodle he had to do with, and wanted to get rid of the man, he told him that an apple fell on his nose; and this made the matter quite clear to the man, and he went away satisfied.
>
> Carl Friedrich Gauss, as quoted by Robert Chambers, *The Book of Days* (1832).

In Newtonian gravity, the gravitational potential Φ is a function on spacetime satisfying Poisson's equation:

$$\Delta\Phi = 4\pi G\rho.$$

Here Δ is the Laplacian in 3 dimensions: in an inertial frame (that is, in the natural rectangular coordinate associated with an inertial observer), it is given by

$$\Delta = \sum_{i=1}^{3} \partial_i^2,$$

G is Newton's constant, and ρ, mapping points in spacetime to $\mathbb{R}$, is the matter density.

Particles in a gravitational field experience a gravitational force in the direction of the gradient of the gravitational potential. If a particle has position $x = x(t)$, then

$$m\ddot{x} = -m\nabla\Phi.$$

2.1 Problems with Newtonian gravity

From the Galilean viewpoint, Newtonian gravity is completely acceptable. However, from the point of view of special relativity, there are several problems.

From a theoretical point of view, the major problem is that the equations of Newtonian gravity are not invariant under Lorentz transformations. Two different inertial observers will generally construct two different gravitational potentials, and predict two different and incompatible motions for particles.

In the Newtonian theory, gravity propagates instantaneously: Poisson's equation is solved separately at each instant of time. So if, for example, the Sun were to suddenly disappear, then it would take about 8 minutes for the light to go out, but we would notice the gravitational effect instantly. But what does

"instantly" mean? Remember that, in special relativity, different inertial observers will define different time coordinates, and will label different times as "now".

There are also various observational problems with Newtonian gravity, which were beginning to cause problems for physicists around the time of Einstein. The orbit of the planet Mercury did not quite match with astronomers' predictions, leading to the prediction of an extra planet (dubbed 'Vulcan') to account for the observed deviations. Then, there is the bending of light: if light is considered as a wave (i.e. treating Maxwell's equations classically) then there should be no bending of light, since the gravitational field does not appear in Maxwell's equations. On the other hand, considering light as a particle, then there should be some bending[1].

Finally, there is the famous *equivalence principle*, which leads to a philosophical argument against Newtonian gravity.

2.2 The equivalence principle

Consider the following two situations: situation (A), in which a closed box (or elevator) is in free-fall towards the Earth[2], and situation (B), in which a similar closed box is freely floating in space (figure 2.1). Then there is no *local* way (i.e. on a short length and timescale) that an experimenter inside the box could tell whether they are in situation (A) or situation (B). This is despite the fact that, in Newtonian gravity, these two situations are described in completely different manners: in situation (A) we would say that the box sits in a gravitational field and the experimenter experiences a gravitational force, while in situation (B) no such force is present.

In Newtonian gravity, the reason why it is difficult to tell these situations apart is that the gravitational force is proportional to the mass. This should remind you of *fictitious forces* such as the centrifugal force[3].

There is an amusing variant of this thought experiment, which goes as follows. Imagine sitting in a sealed room on the Earth with no windows. Then imagine that the entire Earth suddenly vanishes, except for the room you are sitting in, but at the same moment the room starts to accelerate upwards at 9.8ms^{-2}, perhaps pushed upwards by some rockets beneath the room (figure 2.2). Then there is no way that, from inside the room, you will be able to tell that this has happened!

These ideas are formalised in the *equivalence principle*, which comes in three versions:

The weak equivalence principle

The trajectories of all test particles moving in gravitational fields depend only on their initial positions and velocities.

Here a *test particle* is a point mass which does not self-interact gravitationally. The weak equivalence principle implies that the mass of such a test particle has no effect on its motion through a gravitational field.

The Einstein equivalence principle

All local, non-gravitational experiments performed in freely falling laboratories will obtain the same outcomes, regardless of the position and velocity of the lab.

Here a *freely falling lab* is one in which no local acceleration of the lab can be measured. Such a lab could either be floating in space or falling to Earth, as discussed earlier.

[1] This might seem odd, given that the gravitational force on a massless particle is zero. However, the approach to massless particles in Newtonian theory was to write out the equations with the mass m as a parameter, and then to take the limit $m \to 0$. When working out the trajectory of a massless particle in a gravitational field, m appears on both the right and left hand sides of the equation of motion, and can therefore be cancelled off both sides.

[2] As usual, we picture the ideal situation in which there is no air resistance!

[3] It is a universal feature of "fictitious forces" that they are proportional to the mass of the object on which they are acting, just like the force of gravity.

This upgrades the weak equivalence principle to apply to experiments which involve things other than the motion of test particles, e.g. experiments involving other forces such as electromagnetism. The experiment must still be *local*, meaning that it can't involve large time or length scales, otherwise *tidal forces* can be measured. The Einstein equivalence principle also rules out experiments involving significant gravitational interaction between the measured objects.

The strong equivalence principle

The motion of all sufficiently small bodies moving in gravitational fields depends only on their initial positions and velocities. Also all local experiments performed in freely falling laboratories will obtain the same outcomes, regardless of the position and velocity of the lab.

The first part upgrades the weak equivalence principle to apply to sufficiently small extended bodies, while the second part upgrades the Einstein equivalence principle to allow for gravitational interaction between the objects of study.

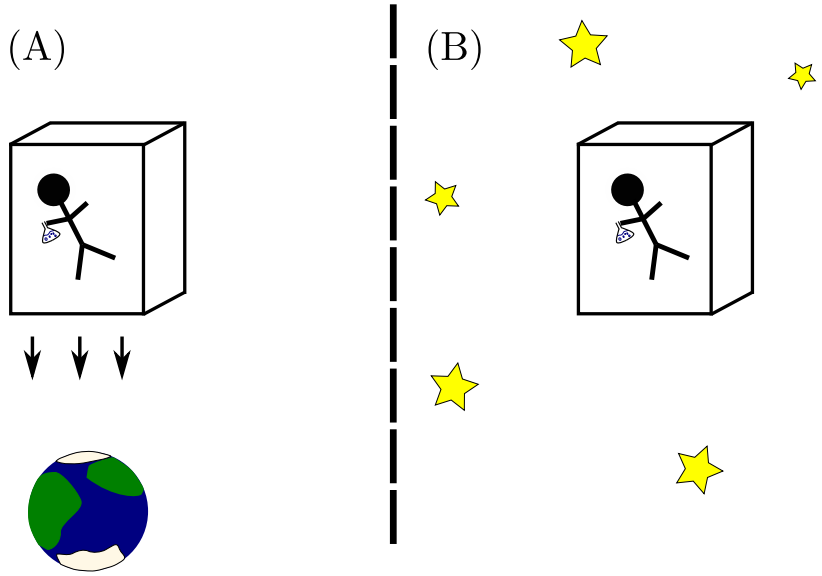

Figure 2.1: Einstein's thought experiment: (A) a scientist in an elevator falling to Earth, while in (B) they are floating in space. No local experiment can distinguish between (A) and (B).

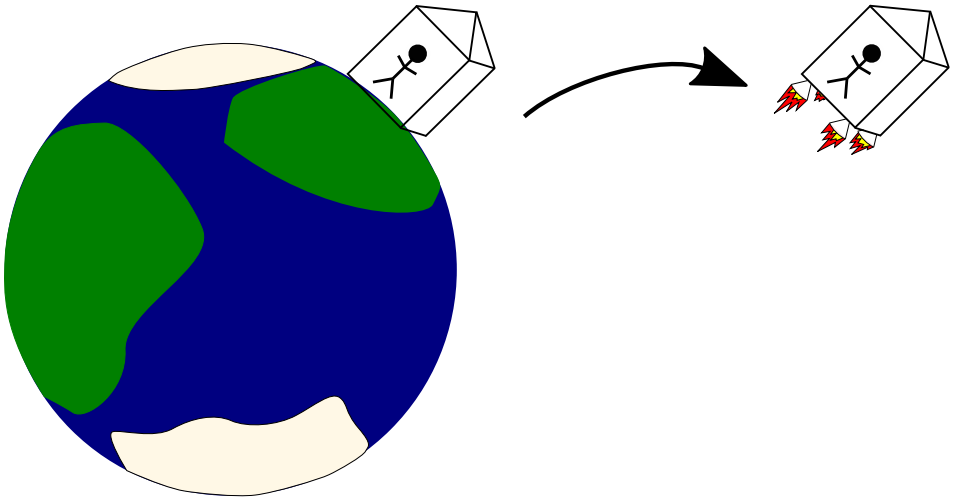

Figure 2.2: Did this just happen?

The equivalence principle strongly suggests that spacetime is curved. The intuition runs as follows: in previous versions of spacetime, special curves have been distinguished, which we have called the worldlines of inertial observers. Physically, these are supposed to represent paths taken by observers who experience no external forces, and we have always assumed that such paths are straight lines. On the other hand, the equivalence principle suggests that *freely falling observers* experience no external forces. However, freely falling objects don't move along paths which we would normally consider "straight lines" – they can orbit the Earth, for example. We can still consider these to be a kind of straight line, but only if *spacetime itself is curved*!

Chapter 3

Review of special relativity

> Since the mathematicians have invaded the theory of relativity, I do not understand it myself anymore.
>
> Einstein, quoted in *Zum Siebzigsten Geburtstag Albert Einsteins* (To Albert Einstein's Seventieth Birthday), translated by Paul A. Schilpp.

In this chapter we'll do a lightning survey of special relativity, putting special emphasis on the aspects of the theory which will be central in the transition to GR, and viewing things from a "geometric" perspective.

3.1 Conventions

First, let's fix some of the conventions we'll use thoughout the course.

The *signature* of the metric is $(-,+,+,+)$. In other words, a $-$ sign is associated with times, not lengths, and the Minkowski metric is $\mathrm{diag}(-1,1,1,1)$. The alternative signature, $(+,-,-,-)$, is frequently used, particularly in high energy physics, but we'll take the viewpoint that time is weird and lengths are normal.

We'll *always* use the Einstein summation convention, so whenever repeated indices appear we will sum over them. For example, if v^μ and w_μ are a spacetime vector and covector respectively, then

$$v^\mu w_\mu := \sum_{\mu=0}^{3} v^\mu w_\mu.$$

On the very rare occasions where we do not wish to sum over a pair of repeated indices, then we will write this explicitly.

Because of this notation, repeated indices are sometimes called *dummy indices*. It doesn't matter which letters are used for these indices: for example, $v^\mu w_\mu = v^a w_a = v^\nu w_\nu$.

We will use Greek indices $\mu, \nu, \rho \ldots$ to refer to *abstract spacetime indices*: if we write out some equation involving this kind of index, then we have not picked any particular set of coordinates. Consequently, any such equation should be independent of the coordinates we choose!

Latin indices from the start of the alphabet, $a, b, c \ldots$ will be reserved for *concrete indices*, that is, they will always refer to a particular set of coordinates (or, occasionally, some special class of coordinates). Thus, equations written with these indices might not be true in an alternative coordinate system. In general, when such an index takes the value 0 then it refers to 'time', and when it takes a value $1, 2$ or 3 then it refers to 'space'.

Latin indices from the middle of the alphabet, $i, j, k \ldots$ will be used to refer to *spatial* indices, excluding time.

We will use "geometrized units": the speed of light $c = 1$, and Newton's constant $G = 1$, except on the few occasions where we display them explicitly for clarity.

3.2 The metric and causal structure

In *inertial coordinates* $x^a = (x^0, x^1, x^2, x^3)$, the Minkowski metric is m_{ab} is given by

$$m := \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

So, for example, $m_{00} = -1$. Similarly, in these coordinates its inverse is

$$m^{-1} := \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Given a (nonzero) spacetime vector v, we say that the vector is *spacelike* if $m(v, v) > 0$, *timelike* if $m(v, v) < 0$ and *null* if $m(v, v) = 0$. Similarly, given two points p and q in Minkowski spacetime, we say that these points are *spacelike separated*, *timelike separated*, or *null separated* (or *lightlike separates*) depending on whether the vector $v = p - q$ is timelike, spacelike or null respectively. A point in spacetime is sometimes called an *event*.

The set of points that are null separated from a point p are said to lie on the *light cone* of the point p (see figure 3.1).

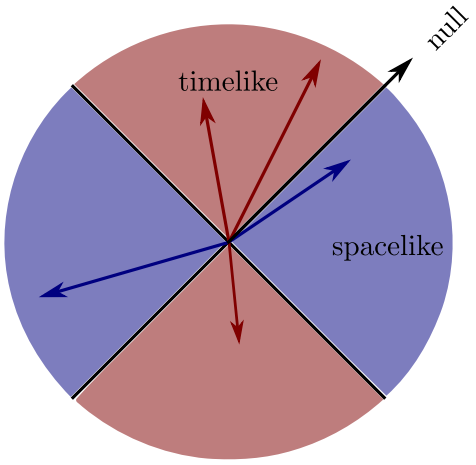

Figure 3.1: Timelike (red), spacelike (blue) and null (black) vectors, in a sketch where two spatial dimensions have been suppressed. The set of timelike vectors has two connected components, which we call *past directed* and *future directed* vectors. Despite its appearance in this sketch, the set of spacelike vectors has only a single connected component: one spatial dimension can be restored by revolving this diagram around the vertical axis.

3.3 Lorentz transformations

A Lorentz transformation is a transformation from one set of inertial coordinates to another, fixing the origin. These are given by linear transformations

$$y^{a'} := \Lambda_a{}^{a'} x^a,$$

where the matrix $\Lambda_a{}^{a'}$ has unit determinant and preserves the form of the metric, that is,

$$m_{ab} = \Lambda_a{}^{a'} \Lambda_b{}^{b'} m_{a'b'}.$$

These can be split up into boosts, which mix the time and space coordinates - e.g. a boost with relative velocity v in the x direction is

$$\Lambda_a{}^{a'} = \begin{pmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$
$$\gamma = \frac{1}{\sqrt{1-v^2}}$$

and rotations, which have the form

$$\Lambda_a{}^{a'} = \begin{pmatrix} 1 & 0 \\ 0 & R_i{}^j \end{pmatrix} \tag{3.1}$$

with $R_i{}^j \in SO(3)$.

3.4 Curves, tangent vectors, proper time and proper length

A *curve* is a map $\gamma : [0,1]$ (or $\mathbb{R}$) $\to \mathbb{M}^4$. In inertial coordinates, we can write $\gamma(\lambda) = x^a(\lambda)$.

The *tangent vector* to the curve γ at the point p with coordinate $x^a(\lambda_0)$ is

$$v^a\big|_p := \frac{\mathrm{d}}{\mathrm{d}\lambda} x^a(\lambda)\big|_{\lambda=\lambda_0}.$$

We can use tangent vectors to measure the rate of change of a function along a curve. Given a function $f : \mathbb{M}^4 \to \mathbb{R}$, using the chain rule we have

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(f \circ \gamma(\lambda)\right) = \frac{\mathrm{d}\gamma^a}{\mathrm{d}\lambda} \frac{\partial f}{\partial x^a},$$

where $\gamma^a(\lambda) = x^a(\lambda)$ are the coordinates of the curve in the coordinates (x^a). Hence

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(f \circ \gamma(\lambda)\right) = v^a \frac{\partial f}{\partial x^a}.$$

The tangent vector to a curve can be either timelike, spacelike or null. Note the character of the tangent vector (i.e. whether it is timelike, spacelike or null) at a point p along the curve does not depend on the parametrisation of the curve[1] (**exercise**).

A curve is said to be timelike, spacelike or null if its tangent vector is everywhere timelike, spacelike or null. Massive particles move along timelike curves, and massless particles move along null curves[2].

[1] If $\sigma : \mathbb{R} \to \mathbb{R}$ is monotonic increasing, then an alternative parametrisation of the curve is given by $\gamma \circ \sigma$.

[2] These curves are also called *worldlines*, in both the timelike and null cases.

If γ is a timelike curve, then the *proper time* along γ, τ, is defined to be the parameter along the curve so that the tangent vector to the curve satisfies

$$m(v,v) = m_{ab}\left(\frac{\mathrm{d}}{\mathrm{d}\tau}x^a(\tau)\right)\left(\frac{\mathrm{d}}{\mathrm{d}\tau}x^b(\tau)\right) = -1.$$

Such a parameter can always be found along a timelike curve (**exercise** - see the previous exercise for the transformation of tangent vectors under reparametrisation). Such a parameter is unique up to the choice of origin, that is, the point along the curve at which $\tau = 0$. It has a physical meaning in special relativity, given by:

Postulate (The clock postulate)**.** *An accurate clock moving along a timelike worldline measures the proper time along the worldline.*

Similarly, if γ is a spacelike curve, then the *proper length* s of γ is the parameter defined so that

$$m_{ab}\frac{\mathrm{d}x^a}{\mathrm{d}s}\frac{\mathrm{d}x^b}{\mathrm{d}s} = 1.$$

There is no analogue to proper time or distance along a generic null curve. However, there are special null curves, which are *generated* by a null vector as follows: let $p \in \mathbb{M}^4$ be some fixed position, and let v be a null vector. Then consider the curve "generated" by the vector v through the point p, defined by the equation

$$\gamma(\lambda) - p = \lambda v.$$

This defines a null curve with tangent vector v. The parameter λ along such a curve is called an *affine parameter*. We can find another affine parametrisation of such a curve by choosing a different point p along the curve, and choosing a different null vector which is proportional to v. Under such a reparametrisation, the affine parameter λ transforms as $\lambda \mapsto a\lambda + b$, for constants a and b.

3.5 Vectors, covectors, tensors and their transformations

Given a point $p \in \mathbb{M}^4$, the *tangent space* at the point p is the vector space consisting of all the vectors from the point p, that is, we define

$$T_p(\mathbb{M}^4) := \{(q-p) \in \mathbb{R}^4 \,\big|\, q \in \mathbb{M}^4\}.$$

Equivalently, we can define $T_p(\mathbb{M}^4)$ as the set of all tangent vectors to curves through the point p.

Note that there is a natural way to identify 'different' tangent spaces, i.e. the tangent spaces $T_p(\mathbb{M}^4)$ and $T_q(\mathbb{M}^4)$ with $q \neq p$. Suppose X is a vector in $T_p(\mathbb{M}^4)$; then there is some point $r \in \mathbb{M}^4$ such that $X + (q-p) = (r-p)$. Then we can define the vector $X' \in T_q(\mathbb{M}^4)$, which corresponds to the vector $X \in T_p(\mathbb{M}^4)$, as

$$X' := (r-q).$$

See the parallelogram in figure 3.2. Although this identification seems trivial, the fact that, when we come to consider curved spacetimes, *there is no such natural identification* leads to a lot of important consequences for GR.

Consider inertial coordinates x^a, chosen so that the point p is at the origin $x^a = 0$. Then the point q has coordinates $x^a = q^a$. Under a Lorentz transformation $y^{a'} = \Lambda_a{}^{a'}x^a$, the coordinates of the point q transform as

$$\begin{aligned} q^a &= x^a \\ &= (\Lambda^{-1})_{a'}{}^a y^{a'} \\ \Rightarrow y^{a'} &= \Lambda_a{}^{a'} q^a. \end{aligned}$$

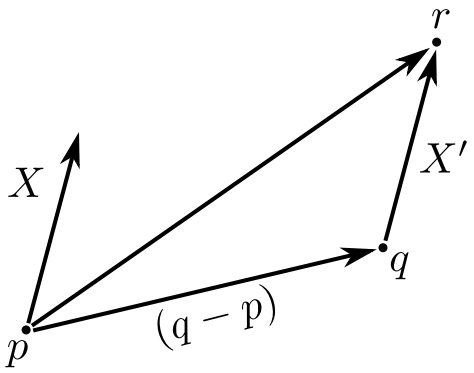

Figure 3.2: The identification of a vector X at p and the corresponding vector X' at q.

In other words, in terms of the new coordinates $y^{a'}$, the coordinates of the point q are

$$q'^{a'} = \Lambda_a{}^{a'} q^a.$$

Hence, the components of the vector $X = q - p$ transform under a Lorentz transformation as

$$X'^{a'} = \Lambda_a{}^{a'} X^a. \tag{3.2}$$

This is the *transformation law for the components of a vector*. Equation (3.2) is sometimes used to *define* vectors, i.e. a vector is any quantity which transforms under this rule.

The *cotangent space* at the point p, $T_p^*(\mathbb{M}^4)$, is the dual space to the tangent space $T_p(\mathbb{M}^4)$, i.e. it is the space of linear maps from $T_p(\mathbb{M}^4)$ to $\mathbb{R}$. Elements of the cotangent space are called *covectors* (or sometimes *1-forms*).

The transformation law for vector components can be used to deduce the corresponding transformation law for covectors: let $\eta \in T_p^*(\mathbb{M}^4)$ be a covector. The components of η with respect to some inertial coordinates (chosen so that p is at the origin) are the four numbers η_a where, for any vector $X \in T_p(\mathbb{M}^4)$, we have

$$\eta(X) = \eta_a X^a,$$

where X^a are the components of the vector X in the inertial coordinates. Note that $\eta(X)$ is just a real number (a *scalar*) – it clearly doesn't transform at all under Lorentz transformations! So, if $\eta'_{a'}$ are the components of η with respect to the inertial coordinates $y^{a'} = \Lambda_a{}^{a'} x^a$, we must have

$$\begin{aligned} \eta_a X^a &= \eta'_{a'} X'^{a'} \\ &= \eta'_{a'} \Lambda_a{}^{a'} X^a, \end{aligned}$$

and so the *transformation law for covectors* is

$$\eta'_{a'} = (\Lambda^{-1})_{a'}{}^a \eta_a. \tag{3.3}$$

Vectors are sometimes said to transform *contravariantly*, while covectors transform *covariantly*.

A *tensor* at the point p is an element of $\left(T_p(\mathbb{M}^4)\right)^n \times \left(T_p^*(\mathbb{M}^4)\right)^m$ for some n, $m \geq 0$. Such a tensor is said to be of *rank* (n, m), or to have *valency* (n, m). For example, given a vector X and a covector η we can form the tensor $X\eta$, which has components (in *any* coordinate system)

$$(X\eta)^\mu{}_\nu := X^\mu \eta_\nu.$$

It is often useful to view a (n, m) tensor as a map from $\left(T_p(\mathbb{M}^4)\right)^m \times \left(T_p^*(\mathbb{M}^4)\right)^n \to \mathbb{R}$, which is always possible due to the finite dimensionality of the tangent and cotangent spaces.

Tensors transform under Lorentz transformations in the obvious way: for a rank (n, m) tensor T, its components transform as

$$(T')^{a'_1 a'_2 \ldots a'_n}{}_{b'_1 b'_2 \ldots b'_m} = \Lambda_{a_1}{}^{a'_1} \Lambda_{a_2}{}^{a'_2} \ldots \Lambda_{a_n}{}^{a'_n} (\Lambda^{-1})_{b'_1}{}^{b_1} (\Lambda^{-1})_{b'_2}{}^{b_2} \ldots (\Lambda^{-1})_{b'_m}{}^{b_m} T^{a_1 a_2 \ldots a_n}{}_{b_1 b_2 \ldots b_m}. \quad (3.4)$$

A *contraction* of a tensor is formed by summing over a pair of indices, with one "up" index and one "down" index. Using the Einstein summation convention, this is written as a tensor with the same letter used in one of the "up" indexes and one of the "down" indexes, e.g. $T^b{}_{ba}$. Such objects are also tensors (**exercise**: show that $T^b{}_{ba}$ transforms as a covector).

Indices can be *lowered* and *raised* using the metric m and its inverse m^{-1}. To be precise: the metric can be used to define an isomorphism $T_p(\mathbb{M}) \to T_p^*(\mathbb{M})$ as follows: for a vector X, and an arbitrary vector Y

$$X \mapsto X^\flat$$
$$X^\flat(Y) = m(X, Y),$$

or, in terms of indices (in which case it is conventional to avoid the 'flat' sign)

$$X_\mu := m_{\mu\nu} X^\nu,$$

and similarly for a covector η and an arbitrary vector Y, we have

$$\eta \mapsto \eta^\sharp$$
$$m(\eta^\sharp, Y) = \eta(Y),$$

or in terms of indices

$$\eta^\mu := (m^{-1})^{\mu\nu} \eta_\nu.$$

3.6 Tensor fields

A *tensor field* is an assignment of a tensor to all points in spacetime.

One example of a tensor field is the metric m: this defines a rank $(0, 2)$ tensor field whose action on the vector fields X, Y is given by

$$m(X, Y) := m_{ab} X^a Y^b$$

where, on the right hand side, we are working in an inertial coordinate system, and $m_{ab} = \mathrm{diag}(-1, 1, 1, 1)$. Note that, because of the special properties of the metric tensor, the metric takes this form in *all* inertial coordinate systems.

Another example is the *identity* or *Kronecker delta*: this is a $(1, 1)$ tensor field whose action on a vector field X and a covector field η is given by

$$\delta(X, \eta) := \eta(X).$$

Note that, in *any* coordinate system (not just inertial ones!) the components of δ are given by[3]

$$\delta^b_a = \mathrm{diag}(1, 1, 1, 1).$$

Next, consider a function $f : \mathbb{M}^4 \to \mathbb{R}$, and consider the covector field $\mathrm{d}f$ whose components in the coordinates x^a are

$$(\mathrm{d}f)_a = \frac{\partial f}{\partial x^a} = \partial_a f.$$

[3]It is conventional *not* to stagger the indices on the Kronecker delta, unlike other tensor fields.

By construction this defines a covector field: its action on the vector X is

$$\mathrm{d}f(X) = (\partial_a f)X^a.$$

Note that, if we work in another set of inertial coordinates $y^{a'} = \Lambda_a{}^{a'} x^a$, then (using the usual transformation law for covector fields) the components of $\mathrm{d}f$ are

$$\begin{aligned}(\mathrm{d}f)'_{a'} &= (\Lambda^{-1})_{a'}{}^{a}(\mathrm{d}f)_a \\ &= (\Lambda^{-1})_{a'}{}^{a}\frac{\partial f}{\partial x^a} \\ &= (\Lambda^{-1})_{a'}{}^{a}\frac{\partial y^{b'}}{\partial x^a}\frac{\partial f}{\partial y^{b'}} \\ &= (\Lambda^{-1})_{a'}{}^{a}\Lambda_a{}^{b'}\frac{\partial f}{\partial y^{b'}} \\ &= \frac{\partial f}{\partial y^{a'}},\end{aligned}$$

so in fact, in *any* inertial coordinates, the components of $\mathrm{d}f$ are given by an expression of the form $\frac{\partial f}{\partial x^a}$.

Finally, consider a more general rank (n, m) tensor field with components in the x^a coordinate system $T^{a_1\ldots a_n}{}_{b_1\ldots b_m}$. We can construct the $(n, m+1)$ tensor field which has components given by the derivatives of the components of T, i.e.

$$\partial_c T^{a_1\ldots a_n}{}_{b_1\ldots b_m}.$$

It is easy to check that this definition is actually independent of the inertial coordinates in which we work, i.e. this expression transforms as an $(n, m + 1)$ tensor field.

3.6.1 Integral curves

If X is a vector field, then we can define the *integral curves of the vector field* X. In some inertial coordinates x^a, the integral curve of the vector field X through the point with coordinates x^a_0 is the curve defined by the ODE

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}x^a(\lambda) &= X^a\big|_{x^a(\lambda)} \\ x^a(0) &= (x_0)^a.\end{aligned}$$

Standard ODE theory ensures that this equation has a unique solution, if the vector field X is smooth and has bounded components.

3.7 Worldlines of particles

Suppose a particle moves along a curve with coordinates $x^a(\lambda)$ in some inertial frame. For a massive particle, we can parametrize this curve by the proper time τ instead of the parameter λ.

We define the *velocity* of the particle v as the tangent vector to its worldline, parametrized by proper time. In terms of inertial coordinates, we have

$$v^a = \frac{\mathrm{d}x^a(\tau)}{\mathrm{d}\tau}.$$

Note that this defines a vector along the curve γ, but it does not define a genuine vector field since there is not prescription for the vector away from the its wordline.

Note that, by the definition of proper time, we have $m_{ab}v^a v^b = -1$. Thus we can write

$$v = \gamma \begin{pmatrix} 1 \\ \boldsymbol{v} \end{pmatrix}$$
$$\gamma = \frac{1}{\sqrt{1 - |\boldsymbol{v}|^2}}.$$

The *four-momentum* of the particle is defined as the covector $p_a := \mu m_{ab} v^b$, where μ is the rest mass of the particle. Thus we have

$$p = \begin{pmatrix} -\mu\gamma \\ \mu\gamma\boldsymbol{v} \end{pmatrix} = \begin{pmatrix} -\mu - \frac{1}{2}\mu|\boldsymbol{v}|^2 + \mathcal{O}(|\boldsymbol{v}|^4) \\ \mu\boldsymbol{v} + \mathcal{O}(|\boldsymbol{v}|^3) \end{pmatrix}.$$

So for $|\boldsymbol{v}| \ll 1$, i.e. for velocities much slower than the speed of light p_0 is (up to an additive constant) the usual expression for the kinetic energy, while p_i are the usual expressions for the components of the momentum. Thus we write

$$p = \begin{pmatrix} -E \\ \boldsymbol{p} \end{pmatrix}.$$

On the other hand, we can choose inertial coordinates so that, at some time τ_0, the velocity vector has components $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$. In these coordinates, at this instant of time, we have

$$p = \begin{pmatrix} -\mu \\ 0 \end{pmatrix}.$$

Now, the quantity $-(m^{-1})^{ab}p_a p_b$ is a scalar quantity, so its value does not depend on which inertial coordinates we use to evaluate it. So we have Einstein's famous formula

$$E^2 - |\boldsymbol{p}|^2 = \mu^2,$$

or, restoring the speed of light using dimensional arguments (recall that we set $c = 1$),

$$E^2 = \mu^2 c^4 + |\boldsymbol{p}|^2 c^2.$$

The *acceleration* of a massive particle is the four-vector

$$a^a := \frac{\mathrm{d}^2}{\mathrm{d}\tau^2} x^a(\tau).$$

Exercise: show that the acceleration of a particle is orthogonal to its four-velocity in the Lorentzian sense, i.e. $m(v, a) = m_{ab}v^a a^b = 0$.

3.8 The energy-momentum tensor

For a continuous distribution of matter, the energy density, energy flux, momentum density and pressure are encoded in a *symmetric* rank $(2, 0)$ tensor field $T^{\mu\nu}$.

If v^a is the tangent vector of the worldline of an observer moving through spacetime, then

- The vector $j^a := -T^{ab}v_b$ is the *four-momentum density*.
- The scalar $\rho := -v_a j^a = T^{ab}v_a v_b$ is the *energy density*. For normal matter $\rho \geq 0$ (the *weak energy condition*).

Moreover, if n and N are spacelike vectors defined along the worldline of an observer which are normalised so that $m(n, n) = m(N, N) = 1$, and which are orthogonal to the velocity of the observer (i.e. $m(n, v) = m(N, v) = 0$), then

- The scalar $p = T^{ab}n_a n_b$ is the pressure measured by the observer in the n direction.
- The stress in the n direction across a surface orthogonal to N is $S = T^{ab}n_a N_b$.

For example, for a *perfect fluid* moving along the integral curves of a vector field u (normalised so that $m(u,u) = -1$) the energy-momentum tensor is

$$T^{ab} := (\rho + p)u^a u^b + p(m^{-1})^{ab},$$

where ρ and p are the fluid density and pressure in its rest frame. The *equation of state* specifies the scalar field p (the pressure of the fluid) in terms of the scalar field ρ (the density of the fluid), i.e. $p = p(\rho)$.

A second example is given by a *massless scalar field*, whose the energy-momentum tensor is

$$T^{ab} := (\partial^a \phi)(\partial^b \phi) - \frac{1}{2}(m^{-1})^{ab}(m^{-1})^{cd}(\partial_c \phi)(\partial_d \phi)$$

where here ϕ is the scalar field, and $\partial^a \phi = (m^{-1})^{ab}\partial_b \phi$.

For continuous distributions of matter, the conservation of energy and momentum is ensured by the conservation of the energy-momentum tensor, that is, by its being divergence free: using inertial coordinates:

$$\partial_a T^{ab} = 0.$$

To see how this is connected with conservation, consider some region S_0, with smooth boundary ∂S_0, in the "time slice" $t = 0$. Write S_t for the time translation of this surface. Let n be the outwards pointing unit normal to ∂_S (see figure 3.3). Then, by Stoke's theorem, we have

$$\int_{S_{t'}} T^{a0}\mathrm{d}x^1\mathrm{d}x^2\mathrm{d}x^3 = \int_{S_0} T^{a0}\mathrm{d}x^1\mathrm{d}x^2\mathrm{d}x^3 + \int_{t=0}^{t'}\int_{\partial S_t} T^{ab}n_b \,\mathrm{d}\Sigma\mathrm{d}t$$

where $\mathrm{d}\Sigma$ is the surface element of ∂S_t. Choosing $a = 0$ we find that the energy in the region S_t (or "in the region S at the time t) is equal to the initial energy in the region S_0, plus the integral of the flux of energy through the bounary ∂_S. Similarly, choosing $a = 1, 2, 3$ we obtain the same conclusion for the momentum in the region S.

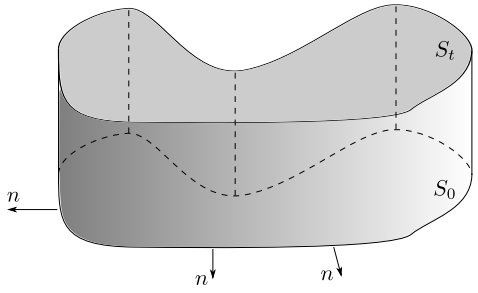

Figure 3.3

Chapter 4

Differential geometry

> Riemann has shewn that as there are different kinds of lines and surfaces, so there are different kinds of space of three dimensions; and that we can only find out by experience to which of these kinds the space in which we live belongs. In particular, the axioms of plane geometry are true within the limits of experiment on the surface of a sheet of paper, and yet we know that the sheet is really covered with a number of small ridges and furrows, upon which (the total curvature not being zero) these axioms are not true. Similarly, he says although the axioms of solid geometry are true within the limits of experiment for finite portions of our space, yet we have no reason to conclude that they are true for very small portions; and if any help can be got thereby for the explanation of physical phenomena, we may have reason to conclude that they are not true for very small portions of space.
>
> William Kingdon Clifford, *On the Space-Theory of Matter*, Proceedings of the Cambridge Philosophical Society (1876)

General relativity predicts that spacetime is curved, but contrary to the quote from Clifford above, in most places in the universe it is not significantly curved on small scales, but instead on large scales. In fact, it is curved on the scale at which gravitational effects become relevant. In order to understand the curvature of spacetime, we will need a fair amount of the mathematics of *differential geometry*. We will be able to cover this subject in its full glory – instead, we will concentrate on the parts of the subject which will come in useful later, covering them in as much rigour as we have time for.

4.1 Manifolds and coordinate charts

The basic object of study in differential geometry is a *manifold*. A manifold $\mathcal{M}$ is a topological space[1] (i.e. we can talk about open sets in $\mathcal{M}$), where sufficiently small open sets "look like" $\mathbb{R}^n$.

This is made precise as follows: for every point $p \in \mathcal{M}$, there is an open neighbourhood U of p and a map $\phi_U : U \to \mathbb{R}^n$ (called a *chart* or *coordinate chart* - the set U is called a *coordinate patch*). These charts are required to be *bijections* between U and the image $\phi_U(U)$, they are also *continuous* and they have *continuous inverses*. Consequently, $\phi_U(U)$ is an open subset of $\mathbb{R}^n$. n is some fixed natural number, called the *dimension* of the manifold.

We can use the chart ϕ_U to define *local coordinates* x^a in the set U. These are defined as the 'pull-back' of the standard coordinates on $\mathbb{R}^n$: in a slight abuse of notation, for each $a \in \{0, 1, 2, \ldots n-1\}$ we set

$$x^a(p) = x^a(\phi_U(p)),$$

[1] It is also required to be *second countable* and *Hausdorf*, but these technical details will not concern us.

where, on the right hand side, $x^a(\phi_U(p))$ is just the value of the standard coordinate x^a in $\mathbb{R}^n$ at the point $\phi_U(p)$. See figure 4.1.

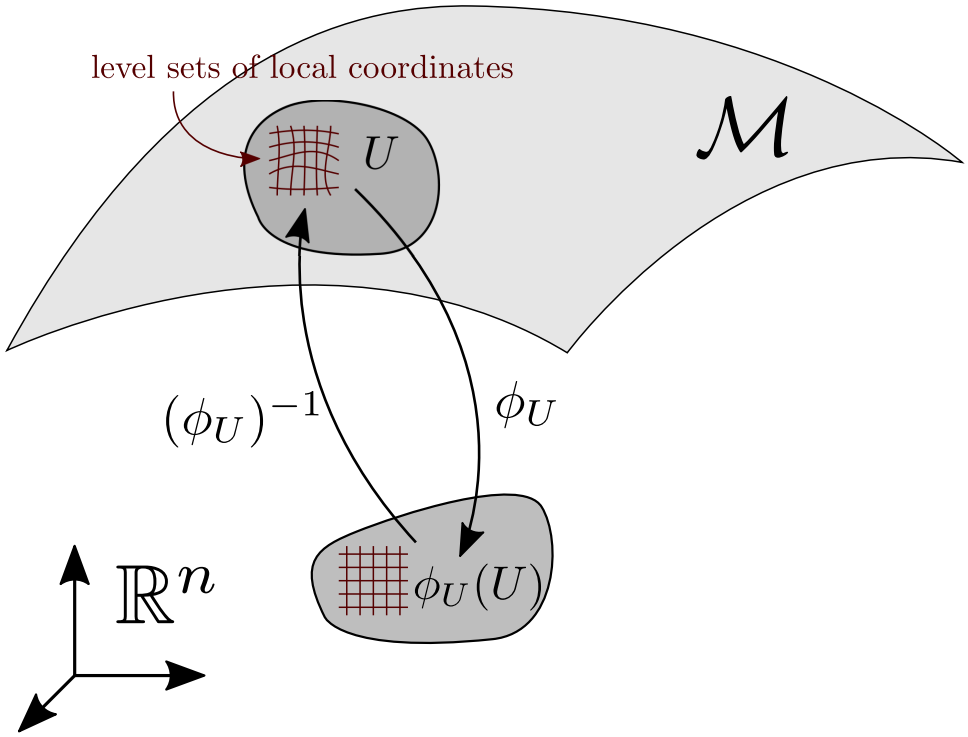

Figure 4.1: A manifold $\mathcal{M}$, with a coordinate patch U and a chart ϕ_U. Local coordinates are defined in the patch ϕ_U by using the usual coordinates on $\mathbb{R}^n$ and the chart ϕ_U.

Generally, we will need more than one chart to cover the manifold $\mathcal{M}$. An *atlas* is a collection of charts which covers the entire manifold.

It can happen that two charts overlap - that is, we can have charts ϕ_U and ϕ_V with $U \cap V \neq \emptyset$. On the overlap, we can define the *transition functions*:

$$\begin{aligned} \phi_{U,V} : \mathbb{R}^n &\to \mathbb{R}^n \\ x &\mapsto \phi_V \circ (\phi_U^{-1})(x) \end{aligned}$$

(see figure 4.2).

Since these transition functions are simply maps from some open set of $\mathbb{R}^n$ to another open set of $\mathbb{R}^n$, we can make sense of, for example, the differentiability of these maps. We will always work with *smooth manifolds*, meaning that the transition functions are C^∞.

4.2 Curves and tangent vectors

As before curve on the manifold $\mathcal{M}$ is a map

$$\gamma : [0,1] \text{ (or } \mathbb{R}) \to \mathcal{M}.$$

How can we make sense of tangent vectors? Unlike before, we don't have a map from differences of points to a vector space. However, we can still differentiate functions along a curve: given $f : \mathcal{M} \to \mathbb{R}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} f \circ \gamma := V(f).$$

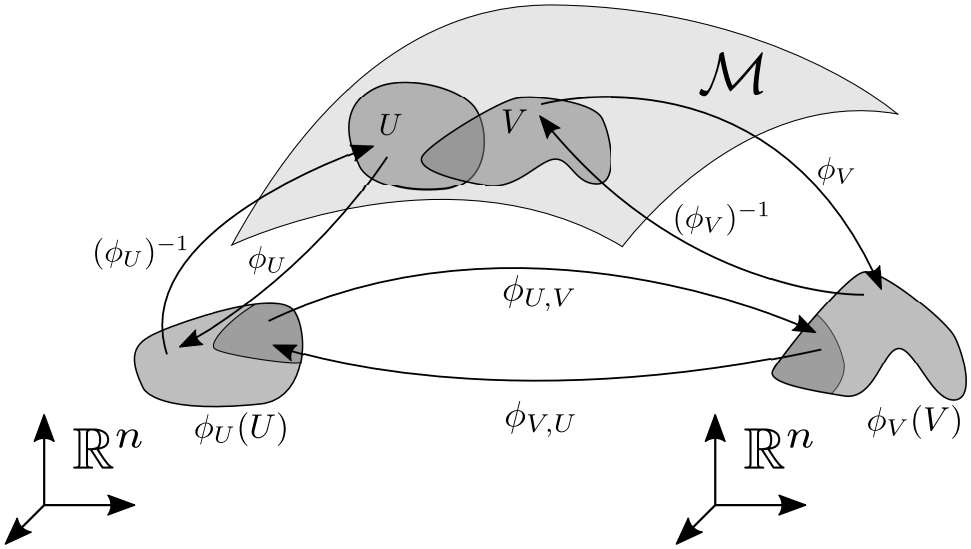

Figure 4.2: Here the coordinate patches U and V overlap, allowing us to define the transition functions $\phi_{U,V}$ and $\phi_{V,U}$. For a smooth manifold, these transition functions are smooth.

Here we *define* V to be the tangent vector to the curve γ. It satisfies the following two important properties: for constants a, $b \in \mathbb{R}$ and functions f, $g : \mathcal{M} \to \mathbb{R}$

1. *Linearity*:
$$V(af + bg) = aV(f) + bV(g).$$
2. *The Leibniz rule*:
$$V(fg) = gV(f) + fV(g).$$

In terms of local coordinates x^a, we can set

$$\begin{aligned} V(f)\big|_p &= V(f \circ \phi_U^{-1} \circ \phi_U)\big|_p \\ &= V\left(\tilde{f}(x^a)\right)\big|_p, \end{aligned}$$

where $\tilde{f} = f \circ \phi_U^{-1}$. Then, using the chain rule we have

$$\begin{aligned} V(f)\big|_p &= V(x^a)\big|_p \frac{\partial \tilde{f}}{\partial x^a}\bigg|_{x^a(p)} \\ &= V^a \partial_a \tilde{f}. \end{aligned}$$

Since this formula holds in *all* local coordinates, we write $V = V^\mu \partial_\mu$. By a common abuse of notation, people often write f for $\tilde{f} = f \circ \phi_U^{-1}$, although these are two different objects: f is a function on the manifold, while $\tilde{f}$ is a function of the local coordinates x^a (of course, they take the same value at corresponding points!).

4.3 Vectors, covectors, tensors and their transformation laws

Given a point $p \in \mathcal{M}$, a *vector at* p is just the tangent vector to some curve[2] through p, at the point p.

[2]Strictly speaking we need to talk about equivalence classes, because there are multiple curves with the same tangent vector. Two curves γ and γ', with tangent vectors V and V' at p are said to define the same vector if $V(f) = V'(f)$ for all f.

The *tangent space at* p, $T_p(\mathcal{M})$ is simply the set of all vectors at p. It is not hard to show that $T_p(\mathcal{M})$ is a vector space with the same dimension as the dimension of the manifold (**exercise** - hint: read the next sentence!). In fact, given some local coordinates x^a, we can define the vectors $\partial_a = \frac{\partial}{\partial x^a}$ as the vectors tangent to the curves along which x^a changes while x^b, $b \neq a$ remain constant, parametrised by x^a (see figure 4.3). Such vector fields are sometimes called *coordinate induced vector fields*.

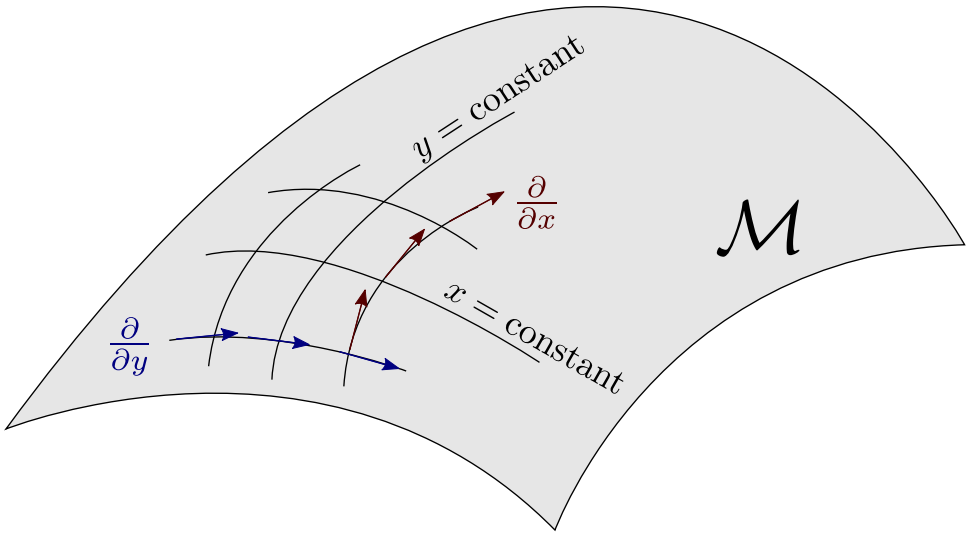

Figure 4.3: The *coordinate induced vector field* $\frac{\partial}{\partial x}$ points in the direction where x changes while all the other coordinates (here, the coordinate y) remain the same. Similarly, $\frac{\partial}{\partial y}$ points in the direction where y changes while x remains constant.

As before, the *cotangent space* $T_p^*(\mathcal{M})$ is the dual space of the vector space $T_p(\mathcal{M})$, i.e. it consists of all linear maps (called *covectors*) from the tangent space to the reals.

A *tensor of rank* (n, m) is an element of $(T_p(\mathcal{M}))^n \times \left(T_p^*(\mathcal{M})\right)^m$. We can also view this as a linear map from $(T_p(\mathcal{M}))^m \times \left(T_p^*(\mathcal{M})\right)^n$ to the reals.

Now, suppose we have some local coordinates x^a. Then the components of the vector X are

$$X^a := X(x^a).$$

Note that, by the chain rule, we have

$$X = X^a \partial_a.$$

In particular, the components of the vector ∂_b are

$$(\partial_b)^a = \partial_b(x^a) = \delta_b^a.$$

Suppose we change coordinates in a neighbourhood of the point p, from the coordinates x^a to coordinates $y^{a'}$. Then the new components of the vector X are, using the chain rule,

$$(X')^{a'} = X(y^{a'}) = \frac{\partial y^{a'}}{\partial x^a} X(x^a) = \frac{\partial y^{a'}}{\partial x^a} X^a.$$

This is the *transformation law for vectors*.

Let η be a covector. Then the components of η are defined to be

$$\eta_a := \eta(\partial_a)$$
$$\Leftrightarrow \eta = \eta_a \mathrm{d}x^a.$$

Note that

$$\eta(X) = \eta(X^a \partial_a) = X^a \eta_a.$$

Since this holds in *any* coordinate system, we can write $\eta(X) = \eta_\mu X^\mu$. Now, under a change of coordinates as above, we have

$$\eta(X) = \eta_a X^a = (\eta')_{a'} X^{a'} = (\eta')_{a'} \frac{\partial y^{a'}}{\partial x^a} X^a,$$

so we must have

$$\begin{aligned} \frac{\partial y^{a'}}{\partial x^a} (\eta')_{a'} &= \eta_a \\ \Rightarrow (\eta')_{a'} &= \frac{\partial x^a}{\partial y^{a'}} \eta_a \end{aligned}$$

using the inverse function theorem. This is the *covector transformation law.*

More general tensors transform in the obvious way:

$$(T')^{a'_1 a'_2 \ldots a'_n}{}_{b'_1 b'_2 \ldots b'_m} = \frac{\partial y^{a'_1}}{\partial x^{a_1}} \frac{\partial y^{a'_2}}{\partial x^{a_2}} \cdots \frac{\partial y^{a'_n}}{\partial x^{a_n}} \frac{\partial x^{b_1}}{\partial y^{b'_1}} \frac{\partial x^{b_2}}{\partial y^{b'_2}} \cdots \frac{\partial x^{b_m}}{\partial y^{b'_m}} T^{a_1 a_2 \ldots a_n}{}_{b_1 b_2 \ldots b_m}.$$

4.4 Tensor fields and examples

As before, a *tensor field* is an assignment of a tensor to all points in spacetime[3]. Sometimes we may define a vector field only on some open subset of the manifold.

As usual, we can also consider a rank (n, m) tensor field as a linear operator at each point p from $\left(T^*_p(\mathcal{M})\right)^m \times \left(T^*_p(\mathcal{M})\right)^n$ to the reals. This means that it is C^∞*-linear* in its arguments. For example, a covector field η is a function from vector fields to the reals, satisfying

$$\eta(aX + bY) = a\eta(X) + b\eta(Y) \qquad \text{for all scalar fields } a, b \text{ and all vector fields } X, Y.$$

Note that a and b are allowed to vary (smoothly) from point to point – they do not have to be constant!

The *tangent bundle* $T(\mathcal{M})$ is the union of all of the tangent spaces of the manifold:

$$T(\mathcal{M}) = \bigcup_{p \in \mathcal{M}} T_p(\mathcal{M})$$

An element of the tangent bundle is a pair (p, X), where p is a point in the manifold and X is a vector at p. In an exactly analogous way, we can define the *cotangent bundle* as the union of all the cotangent spaces.

4.4.1 Some examples of tensor fields

Suppose that $U \subset \mathcal{M}$ is covered by a coordinate chart, with local coordinates x^a. Then, in the set U, we can define the vector fields $\partial_a = \frac{\partial}{\partial x^a}$ as above.

Given a smooth function $f : \mathcal{M} \to \mathbb{R}$ (a *scalar field*), we can define the covector field $\mathrm{d}f$, the *differential of* f, by its action on an arbitrary vector field X:

$$\mathrm{d}f(X) := X(f).$$

Since, for all p, this defines a linear map from $T_p(\mathcal{M})$ to the reals, this defines a covector field.

[3]We will always work with smooth, i.e. C^∞ tensor fields. To check the differentiability of a tensor field we can simply examine its components in a chart. Since the transition functions are restricted to be smooth, this is a coordinate-independent notion.

As before, we can define the *Kronecker delta*, which is a $(1,1)$ tensor field, defined by its action on an arbitrary vector field X and covector field η:

$$\delta(X,\eta) = \eta(X).$$

This defines a linear map from $T_p(\mathcal{M}) \times T_p^*(\mathcal{M})$ to the reals, and so δ is a tensor field.

Given two vector fields X and Y, we can form their product XY, which is a rank $(2,0)$ tensor field with components $(XY)^{ab} = X^a Y^b$. Similarly we can form a scalar field by contracting the indices of a $(1,1)$ tensor field: $T^\mu{}_\mu$ is a scalar field, with values (*in any local coordinates*) given by $T^a{}_a$ (**exercise**: show that this defines a tensor field.). The same applies to tensors of higher rank: we can form new tensors by taking products of tensors (in which case we increase the overall rank), or by contracting indices (in which case we lower the rank).

If $T_{\mu\nu}$ is a rank $(0,2)$ tensor field, then we can define its symmetric and antisymmetric parts:

$$T_{(\mu\nu)} := \frac{1}{2}\left(T_{\mu\nu} + T_{\nu\mu}\right)$$
$$T_{[\mu\nu]} := \frac{1}{2}\left(T_{\mu\nu} - T_{\nu\mu}\right).$$

4.4.2 The metric tensor

Finally, we introduce *the metric tensor*, g. Manifolds which come equipped with a metric tensor are called[4] *Lorentzian manifolds*.

The metric is a symmetric, rank $(0,2)$ tensor field. For any vector field X, we define the covector field $X^\flat$ by

$$X^\flat(Y) := g(X,Y) \quad \text{for all vector fields } Y.$$

Then the metric is *non-degenerate*: $X^\flat = 0$ if and only if $X = 0$. In components we can write

$$(X^\flat)_a = g_{ab}X^b = X_a,$$

where we adopt the notational convention that the metric g lowers indexes.

The metric also has signature $(-,+,+,+)$. This means that, in any coordinate system, at any point in the manifold, the matrix $g_{ab} = g(\partial_a, \partial_b)$ has signature $(-,+,+,+)$ (**exercise**: show that the notion of *signature* is invariant under a change of coordinates).

The metric g will play the same role as the Minkowski metric m did in special relativity. In short:

- A nonzero vector X is *timelike* if $g(X,X) < 0$, spacelike if $g(X,X) > 0$ and null if $g(X,X) = 0$.
- Curves are timelike/spacelike/null if their tangent vector is everywhere timelike/spacelike/null.
- On a timelike curve we define the proper time as the parameter such that the tangent vector has norm -1. Similarly, on a spacelike curve the proper distance is defined so that the tangent vector has norm 1.

There are several common notations in use for the metric tensor. In terms of local coordinates x^a, we can write

$$g = g_{ab}\mathrm{d}x^a\mathrm{d}x^b = \mathrm{d}s^2.$$

Often it is taken for granted that the metric is symmetric, and so for brevity a non-symmetric expression is written down, with the understanding that the true metric is found by symmetrising. For example, we might write

$$g = \mathrm{d}x\mathrm{d}y,$$

[4] If the metric has Lorentzian signature. If the metric had signature $(+,+,+,+)$, then we would call it a *Riemannian manifold*.

which should be understood as

$$g = \frac{1}{2}\mathrm{d}x\mathrm{d}y + \frac{1}{2}\mathrm{d}y\mathrm{d}x.$$

The quantity $\mathrm{d}s^2$, which is really just the metric tensor, is sometimes called the *line element*.

We can also define the *inverse metric* g^{-1}. This is a rank $(2,0)$ metric defined by the relation

$$g^{-1}(X^\flat, \eta) = \eta(X).$$

for all vector fields X and covector fields η.

In components, this reads

$$(g^{-1})^{ab}X_a\eta_b = (g^{-1})^{ab}g_{ac}X^c\eta_b = X^a\eta_a.$$

Since this holds for all X and η, it follows that $(g^{-1})^{ab}g_{ac} = \delta^b_c$, i.e. the matrix $(g^{-1})^{ab}$ is the inverse of the matrix g_{ab}.

We can use the inverse metric to raise indices: for a covector field η, define the vector field $\eta^\sharp$ by

$$g(\eta^\sharp, Y) = \eta(Y),$$

for all vector fields Y. In components

$$(\eta^\sharp)^a = (g^{-1})^{ab}\eta_b = \eta^a.$$

4.5 Calculus on manifolds

Before we can understand dynamics on manifolds, we need to know how to do calculus. Specifically, we need to know how to take derivatives[5] of functions, vector fields, tensor fields etc.

Actually, we already know one way to differentiate scalar fields. Recall that, given a scalar field f, we defined the covector field $\mathrm{d}f$ such that, for vector fields X, $\mathrm{d}f(X) = X(f)$. This is the generalization of the 'gradient of a function' to manifolds, and $X(f)$ is the 'directional derivative' of f in the direction of the vector X.

The components of the covector field $\mathrm{d}f$ in the coordinate system x^a are given by

$$(\mathrm{d}f)_a = (\mathrm{d}f)(\partial_a) = \partial_a f.$$

Since this holds in *any* coordinate system, we can write

$$(\mathrm{d}f)_\mu = \partial_\mu f.$$

How about vector fields? There is one obvious way to differentiate a vector field: choose some local coordinates and differentiate the components of the vector field. Unfortunately this won't work: consider a change of coordinates $x^a \to y^{a'}$. Then

$$\begin{aligned}\partial_a X^b &= \frac{\partial y^{a'}}{\partial x^a}\partial'_{a'}\left(\frac{\partial x^b}{\partial y^{b'}}(X')^{b'}\right)\\ &= \frac{\partial y^{a'}}{\partial x^a}\frac{\partial x^b}{\partial y^{b'}}\partial'_{a'}(X')^{b'} + \frac{\partial y^{a'}}{\partial x^a}\left(\frac{\partial^2 x^b}{\partial y^{a'}\partial y^{b'}}\right)(X')^{b'}.\end{aligned}$$

The first term is the expected term for the transformation of a $(1,1)$ tensor field, but the second term is anomalous. Therefore, the quantity $\partial_a X^b$ can't be a $(1,1)$ tensor field. The geometric reason why this

[5]Integration on manifolds is also important for a variety of applications, including analysing the Einstein equations, however it is beyond the scope of this course.

doesn't work is that there is no canonical way to identify the different tangent spaces at different points, i.e. $T_p(\mathcal{M})$ and $T_q(\mathcal{M})$. Trying to differentiate the a vector field involves the difference of a vector field at two different points, but we cannot add or subtract vectors at different spacetime points!

A similar discussion can be had using index-free notation. Suppose that we want to differentiate the vector field X in the direction of the vector field Y. Then we might be tempted to try to define a vector field $Y(X)$ which acts on scalar fields f as

$$Y(X)(f) := Y(X(f)).$$

Although this obeys linearity, it does not obey the Leibniz rule, and so does not define a vector field. The reason is fairly obvious: the expression above depends on the second derivatives of f, whereas a vector field is supposed to only take first derivatives of f.

How do we resolve this issue? There are actually three different approaches, but we will only pursue one of these in this course (one alternative approach is introduced example sheet 2, and the third approach will be covered in the *GR2* course). The most important approach for our purposes is through the introduction of an *affine connection*.

4.5.1 Affine connections

An affine connection is something that is more or less *defined* to do the job for us, so that we can differentiate vector fields. You could worry that there might not actually be such an object, but, as we will see later, on a Lorentzian manifold we can use the metric to construct a *natural, unique* affine connection. But we will return to that later.

An affine connection is a map from a pair of vector fields to a vector field

$$\Gamma : (X, Y) \mapsto \nabla_X Y$$

with the following properties:

1. ∇ is C^∞-*linear* in the first variable: for all scalar fields f and vector fields X,

$$\nabla_{fX} Y = f \nabla_X Y.$$

2. ∇ satisfies the Leibniz rule in the second variable:

$$\nabla_X (fY) = f \nabla_X Y + (X(f))Y.$$

We can make sense of the *components of a connection*, which are also called *Christoffel symbols*. These are defined, with respect to the local coordinates x^a, as follows:

$$\begin{aligned} &\nabla_{\partial_b} \partial_c := \Gamma^a_{bc} \partial_a \\ &\Leftrightarrow \Gamma^a_{bc} = (g^{-1})^{ad} g(\nabla_{\partial_b} \partial_c \,,\, \partial_d). \end{aligned}$$

We will usually write ∇_a instead of ∇_{∂_a}.

Note these are *not the components of a tensor field*: under a change of coordinates, we have

$$\begin{aligned}
(\Gamma')^{a'}_{b'c'} &= \mathrm{d}y^{a'}\left(\nabla_{b'}\partial'_{c'}\right) \\
&= \frac{\partial y^{a'}}{\partial x^a}\mathrm{d}x^a\left(\nabla_{b'}\partial'_{c'}\right) \\
&= \frac{\partial y^{a'}}{\partial x^a}\mathrm{d}x^a\left(\nabla_{\frac{\partial x^b}{\partial y^{b'}}\partial_b}\left(\frac{\partial x^c}{\partial y^{c'}}\partial_c\right)\right) \\
&= \frac{\partial y^{a'}}{\partial x^a}\mathrm{d}x^a\left(\frac{\partial x^b}{\partial y^{b'}}\nabla_{\partial_b}\left(\frac{\partial x^c}{\partial y^{c'}}\partial_c\right)\right) \\
&= \frac{\partial y^{a'}}{\partial x^a}\mathrm{d}x^a\left(\left(\frac{\partial x^b}{\partial y^{b'}}\partial_b\left(\frac{\partial x^c}{\partial y^{c'}}\right)\right)\partial_c + \frac{\partial x^b}{\partial y^{b'}}\frac{\partial x^c}{\partial y^{c'}}(\nabla_{\partial_b}\partial_c)\right) \\
&= \frac{\partial y^{a'}}{\partial x^a}\frac{\partial^2 x^a}{\partial y^{b'}\partial y^{c'}} + \frac{\partial y^{a'}}{\partial x^a}\frac{\partial x^b}{\partial y^{b'}}\frac{\partial x^c}{\partial y^{c'}}\Gamma^a_{bc},
\end{aligned}$$

or equivalently

$$\Gamma^a_{bc} = \frac{\partial x^a}{\partial y^{a'}}\frac{\partial y^{b'}}{\partial x^b}\frac{\partial y^{c'}}{\partial x^c}(\Gamma')^{a'}_{b'c'} - \frac{\partial y^{b'}}{\partial x^b}\frac{\partial y^{c'}}{\partial x^c}\left(\frac{\partial^2 x^a}{\partial y^{b'}\partial y^{c'}}\right).$$

The first term transforms like a $(1,2)$ tensor, but the second term does not, so the Christoffel symbols do not define a tensor field. However, note that the anomalous transformation of the Christoffel symbols can exactly cancel the anomalous transformation of the object $\partial_a X^b$!

4.5.2 Covariant derivatives of vectors and tensors

The affine connection allows us to take derivatives - called *covariant derivatives* - of vector fields as well as more general tensor fields. The covariant derivative of the vector field X is a $(1,1)$ tensor field, defined as a linear map from a vector field Y and a covector field η to the reals, as follows:

$$\nabla X : (Y,\eta) \mapsto \eta(\nabla_Y X).$$

In abstract index notation, we can write this $(1,1)$ tensor field as $\nabla_\mu X^\nu$. The components of the tensor field ∇X with respect to some local coordinates are written $\nabla_a X^b$. We can write these components in terms of the Christoffel symbols:

$$\begin{aligned}
\nabla_Y X &= Y^a\nabla_a(X^b\partial_b) \\
&= Y^a(\partial_a X^b)\partial_b + Y^a X^b\Gamma^c_{ab}\partial_c \\
&= Y^a\left(\partial_a X^b + \Gamma^b_{ac}X^c\right)\partial_b,
\end{aligned}$$

where in the last line we have relabelled some of the dummy indices. Hence, the components of ∇X are

$$\nabla_a X^b = \partial_a X^b + \Gamma^b_{ac}X^c.$$

Note that this expression holds in *any* coordinate system[6].

You should not take the position of these indices too seriously! For example, suppose we have chosen some specific coordinate system, and we want to calculate the component $\nabla_0 X^1$. Then, according to the formulae above, this is

$$\begin{aligned}
\nabla_0 X^1 &= \partial_0 X^1 + \Gamma^1_{0c}X^c \\
&= \partial_0 X^1 + \Gamma^1_{00}X^0 + \Gamma^1_{01}X^1 + \Gamma^1_{02}X^2 + \Gamma^1_{03}X^3.
\end{aligned}$$

[6]With this in mind, you might be tempted to write this expression in abstract indices, i.e. $\nabla_\mu X^\nu = \partial_\mu X^\nu + \Gamma^\nu_{\mu\rho}X^\rho$. But this would be a mistake: without any coordinate system, we cannot make sense of the notation $\partial_\mu X^\nu$, which does not represent any $(1,1)$ tensor field. Nor does $\Gamma^\nu_{\mu\rho}$ represent a $(1,2)$ tensor field - it is only the sum $\partial_a X^b + \Gamma^b_{ac}X^c$ which *does* represent the components of a $(1,1)$ tensor field.

In general (i.e. if most of the Christoffel symbols don't vanish) this is not an operator acting on the component X^1 - instead, it depends on *all* of the components of X. So we cannot think of an expression like $\nabla_0 X^1$ as the operator ∇_0 acting on the vector field components X^1 – instead, it should be thought of as the $(0,1)$ component of the $(1,1)$ tensor field ∇X.

Derivatives of general tensor fields

We can also take the covariant derivative of scalar fields, covector fields and higher rank tensor fields. We define the covariant derivative of a scalar field to be the same thing as differential, that is, for a scalar field f,

$$\nabla f = \mathrm{d}f.$$

We can now extend the definition of the covariant derivative to covector fields by requiring that the Leibniz rule holds. So, for a covector field η and an arbitrary vector field X, in some arbitrary coordinate system we have

$$\begin{aligned} \mathrm{d}_a\left(\eta_b X^b\right) &= \left(\nabla_a \eta_b\right) X^b + \eta_b \left(\nabla_a X^b\right) \\ \Rightarrow (\partial_a \eta_b) X^b + \eta_b (\partial_a X^b) &= \left(\nabla_a \eta_b\right) X^b + \eta_b \partial_a X^b + \eta_b \Gamma^b_{ac} X^c \\ \Rightarrow \left(\nabla_a \eta_b\right) X^b &= (\partial_a \eta_b - \Gamma^c_{ab} \eta_c) X^b. \end{aligned}$$

Since this holds for *all* vector fields X, we must have

$$\nabla_a \eta_b = \partial_a \eta_b - \Gamma^c_{ab} \eta_c.$$

Following the same kind of reasoning, we can write out the formula for the covariant derivative of a tensor of general rank:

$$\begin{aligned} \nabla_a T^{b_1 b_2 \ldots b_n}{}_{c_1 c_2 \ldots c_m} &= \partial_a T^{b_1 b_2 \ldots b_n}{}_{c_1 c_2 \ldots c_m} \\ &\quad + \Gamma^{b_1}_{ad} T^{d b_2 \ldots b_n}{}_{c_1 c_2 \ldots c_m} + \Gamma^{b_2}_{ad} T^{b_1 d \ldots b_n}{}_{c_1 c_2 \ldots c_m} \cdots + \Gamma^{b_n}_{ad} T^{b_1 b_2 \ldots d}{}_{c_1 c_2 \ldots c_m} \\ &\quad - \Gamma^d_{a c_1} T^{b_1 b_2 \ldots b_n}{}_{d c_2 \ldots c_m} - \Gamma^d_{a c_2} T^{b_1 b_2 \ldots b_n}{}_{c_1 d \ldots c_m} \cdots - \Gamma^d_{a c_m} T^{b_1 b_2 \ldots b_n}{}_{c_1 c_2 \ldots d}. \end{aligned}$$

4.5.3 The Levi-Civita connection

We have derived all of these properties of affine connections, but we have not shown that such an object really exists. It turns out that, in general, many such objects do exist, and on a Lorentzian manifold there is in fact a unique, natural affine connection. This is called the *Levi-Civita connection*.

The Levi-Civita connection is *torsion free*. To define the torsion, we first need to define the commutator of two vector fields: given vector fields X, Y, their commutator is the vector field $[X,Y]$ which acts on scalar fields as

$$[X,Y](f) := X(Y(f)) - Y(X(f)).$$

This *does* define a vector field, since it satisfies linearity and the Leibniz rule (**exercise**). In terms of components, we have

$$[X,Y]^a = X^b \partial_b Y^a - Y^b \partial_b X^a.$$

You can check that this *does* transform as a tensor field. Note that the commutator can be defined without using an affine connection!

The *torsion* of the connection ∇ is defined as the $(1,2)$ tensor field T, whose action on the covector field η and the vector fields X, Y is

$$T(\eta, X, Y) = \eta\left(\nabla_X Y - \nabla_Y X - [X,Y]\right).$$

In terms of components, this is (**exercise**)

$$T^a_{bc} = \Gamma^a_{bc} - \Gamma^a_{cb}.$$

The Levi-Civita connection is torsion free, so $T^{\mu}_{\nu\rho} = 0$. In other words, in any local coordinate system, the Christoffel symbols are symmetric in their lower indices.

The other defining feature of the Levi-Civita connection is that it is *compatible with the metric*, which means that

$$\nabla_\mu g_{\nu\rho} = 0.$$

How does this lead to a unique "metric compatible" connection? We can calculate

$$\begin{aligned}\nabla_a g_{bc} + \nabla_b g_{ac} - \nabla_c g_{ab} &= \partial_a g_{bc} + \partial_b g_{ac} - \partial_c g_{ab} \\ &\quad - \Gamma^d_{ab} g_{dc} - \Gamma^d_{ac} g_{bd} - \Gamma^d_{ba} g_{dc} - \Gamma^d_{bc} g_{ad} + \Gamma^d_{ca} g_{db} + \Gamma^d_{cb} g_{ad} \\ &= \partial_a g_{bc} + \partial_b g_{ac} - \partial_c g_{ab} - 2\Gamma^d_{ab} g_{dc}\end{aligned}$$

$$\Rightarrow \Gamma^c_{ab} = \frac{1}{2}(g^{-1})^{cd}\left(\partial_a g_{bd} + \partial_b g_{ad} - \partial_d g_{ab}\right).$$

The components of the Levi-Civita connection are given by this expression in *all* coordinate systems. If you like, you can check that this expression does *not* transform as a $(1,2)$ tensor field, but instead as the components of an affine connection.

From this point onwards, we will *always* work with the Levi-Civita connection.

4.6 Normal coordinates

One important way in which a connection differs from a vector field is that we can choose coordinates so that, at some point, the connection vanishes. That is, given a point $p \in \mathcal{M}$, we can choose some local coordinates x^a in a neighbourhood of p such that

$$\Gamma^a_{bc}\big|_p = 0.$$

These coordinates can be further chosen so that the components of the metric at p are

$$g_{ab}\big|_p = \mathrm{diag}(-1, 1, 1, 1).$$

These coordinates are called *normal coordinates at p*. Using normal coordinates can simplify a lot of computations - you should remember that, if some equation holds in a particular coordinate system, *and* if that equation can be written entirely in terms of tensors or tensor fields, then the equation must hold in all coordinate systems. But you should be careful when using normal coordinates to remember that the expressions above *only hold at a single point in spacetime.*

The proof that such coordinates exist can be found in appendix A.

4.7 Parallel transport

We can use a connection to define *parallel transport*. A tensor field T is said to be *parallel transported* (or "parallely transported") along the integral curves of the vector field X if it obeys the equation

$$\nabla_X T = 0,$$

or, if you prefer abstract indices

$$X^\rho \nabla_\rho T^{\mu_1\mu_2\ldots\mu_n}{}_{\nu_1\nu_2\ldots\nu_m} = 0.$$

This is the closest we can get to saying that T remains "parallel to itself" when moved in the direction X (see figure 4.4). Note, however, that this doesn't mean that the values of the *components* of T in any

particular coordinate system remain constant! In normal coordinates at the point p the derivative of the components of T in the direction X vanishes:

$$X^c \partial_c T^{a_1 a_2 \ldots a_n}{}_{b_1 b_2 \ldots b_m} = 0,$$

but this *only* holds at the point p. Note also that this is *not* a 'tensorial' equation, since it involves the operator ∂_c which does not transform as a tensor, so it is not true in a general coordinate system.

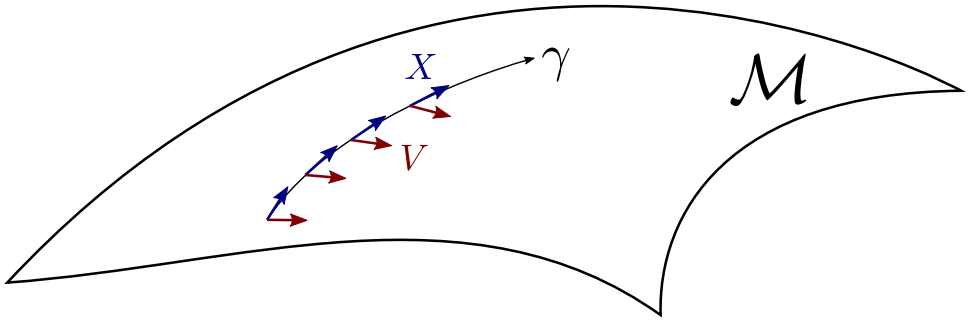

Figure 4.4: X is the tangent vector to the curve γ. Here, the vector field V is parallel transported along γ, i.e. $\nabla_X V = 0$.

4.8 Geodesics

In earlier versions of spacetime, *straight lines* played a very important role: they represented the paths of inertial observers, as well as the paths of particles which are experiencing no external forces. How can we generalise these notions to curved spacetimes?

One way to define "straight lines" which can easily be generalised to curved manifolds is that they *extremise the "distance" between two points*. On a Lorentzian manifold this has to be interpreted appropriately: for timelike curves, the proper time is extremised, while for spacelike curves it is the proper distance that is extremised.

Let γ be a timelike curve through the points p and q on the manifold $\mathcal{M}$, parametrised so that $\gamma(0) = p$ and $\gamma(1) = q$. We work in some local coordinates x^a, where the curve γ has coordinates $x^a(\lambda)$. Then the proper time interval along the curve is[7]

$$\tau(p,q)[\gamma] := \int_0^1 \sqrt{-g_{ab}(x) \frac{\mathrm{d}x^a(\lambda)}{\mathrm{d}\lambda} \frac{\mathrm{d}x^b(\lambda)}{\mathrm{d}\lambda}}\, \mathrm{d}\lambda.$$

The Euler-Lagrange equations can be used to find the curves which extremise this integral. Defining

$$\mathscr{L} := \sqrt{-g_{ab}(x) \frac{\mathrm{d}x^a(\lambda)}{\mathrm{d}\lambda} \frac{\mathrm{d}x^b(\lambda)}{\mathrm{d}\lambda}}.$$

Varying the path $x^a(\lambda)$, the Euler-Lagrange equations give (**exercise**)

$$\begin{aligned} \frac{\mathrm{d}^2 x^a}{\mathrm{d}\lambda^2} + \frac{1}{2}(g^{-1})^{ad}\left(\partial_b g_{cd} + \partial_c g_{bd} - \partial_d g_{bc}\right) \frac{\mathrm{d}x^b}{\mathrm{d}\lambda} \frac{\mathrm{d}x^c}{\mathrm{d}\lambda} &= \mathscr{L}^{-1} \frac{\mathrm{d}\mathscr{L}}{\mathrm{d}\lambda} \frac{\mathrm{d}x^a}{\mathrm{d}\lambda} \\ \Leftrightarrow \frac{\mathrm{d}^2 x^a}{\mathrm{d}\lambda^2} + \Gamma^a_{bc} \frac{\mathrm{d}x^b}{\mathrm{d}\lambda} \frac{\mathrm{d}x^c}{\mathrm{d}\lambda} &= \mathscr{L}^{-1} \frac{\mathrm{d}\mathscr{L}}{\mathrm{d}\lambda} \frac{\mathrm{d}x^a}{\mathrm{d}\lambda}. \end{aligned} \tag{4.1}$$

The Levi-Civita connection arises naturally! In fact, deriving the Euler-Lagrange equations is often the easiest way to derive the components of the Levi-Civita connection in a given spacetime, particularly

[7] To see that this is the proper time, we can change parametrisation from λ to the proper time τ along the curve. Then, using the chain rule, the one-form in the integrand becomes simply $\mathrm{d}\tau$.

if the spacetime in question has some symmetries, since these symmetries lead to conserved quantities along the geodesics which can often simplify a lot of algebra.

If we wish, we can choose the variable λ to be the proper time τ, in which case $\mathscr{L} = 1$, and so we obtain

$$\frac{\mathrm{d}^2 x^a}{\mathrm{d}\tau^2} + \Gamma^a_{bc} \frac{\mathrm{d}x^b}{\mathrm{d}\tau} \frac{\mathrm{d}x^c}{\mathrm{d}\tau} = 0. \tag{4.2}$$

Equation (4.2) is the *geodesic equation*, and curves which satisfy it are called *geodesics*. Exactly the same equation can be derived for spacelike geodesics, which extremise the proper length.

This equation is a second order ODE for the coordinates $x^a(\tau)$ of the geodesic. To solve it, we need both the initial position of the geodesic (i.e. $x^a(0)$, the coordinates of the point from which the geodesic originates) and its initial tangent vector $\frac{\mathrm{d}x^a}{\mathrm{d}\tau}$, i.e. the initial direction of the geodesic.

Equation (4.2) can also be written in terms of the tangent vector to the curve γ. Writing $X^a = \frac{\mathrm{d}x^a}{\mathrm{d}\tau}$, we have

$$\frac{\mathrm{d}X^a}{\mathrm{d}\tau} + \Gamma^a_{bc} X^b X^c = 0. \tag{4.3}$$

One final way to write this equation which will be particularly useful for us is to recall that, along the curve γ (parametrised by proper time) with tangent vector X, for any function f we have

$$\frac{\mathrm{d}}{\mathrm{d}\tau} f = X^a \partial_a f = X(f),$$

so the geodesic equation can be written as

$$\begin{aligned} X^b \partial_b X^a + \Gamma^a_{bc} X^b X^c &= 0 \\ \Leftrightarrow X^b \nabla_b X^a &= 0. \end{aligned}$$

The geodesic equation can therefore be written in the 'tensorial' manner:

$$\nabla_X X = 0. \tag{4.4}$$

in other words, X is parallel transported along its own integral curve.

Equation (4.4) gives us a new way to think about geodesics: *a geodesic is a curve along which the tangent vector to the curve remains parallel to itself.* We can think of ordinary straight lines like this: pick some vector, and extend a curve in the direction of this vector, making sure that, at every point along the curve, the tangent to the curve is parallel to the tangent at the preceding point (c.f. figure 4.4).

This equation also tells us how to define *null geodesics*: a null geodesic is simply a curve whose tangent vector X is both null and satisfies equation (4.4). In all cases (timelike spacelike or null), such a curve is said to be *affinely parametrised* - recall that the tangent to a curve depends on the parametrisation. Note that the character of a geodesic cannot change: if a geodesic is initially timelike/spacelike/null, then it will always be so, since along the cuve γ we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\lambda} \left(X^\mu X_\mu \right) &= \nabla_X \left(X^\mu X_\mu \right) \\ &= 2 g_{\mu\nu} X^\nu \nabla_X X^\mu \\ &= 0, \end{aligned}$$

where we have used the fact that $\nabla g = 0$. So $g(X, X)$ is constant along a geodesic.

4.9 Curvature

We are now (finally!) at a point where we can investigate the most important difference between curved and flat spacetimes: namely, the curvature.

The curvature measures, in a sense, the deviation away from flat space. There are several ways to approach this, but we will do it by means of *geodesic deviation*. The idea is that curvature can cause nearby geodesics to converge or diverge (see figure 4.5).

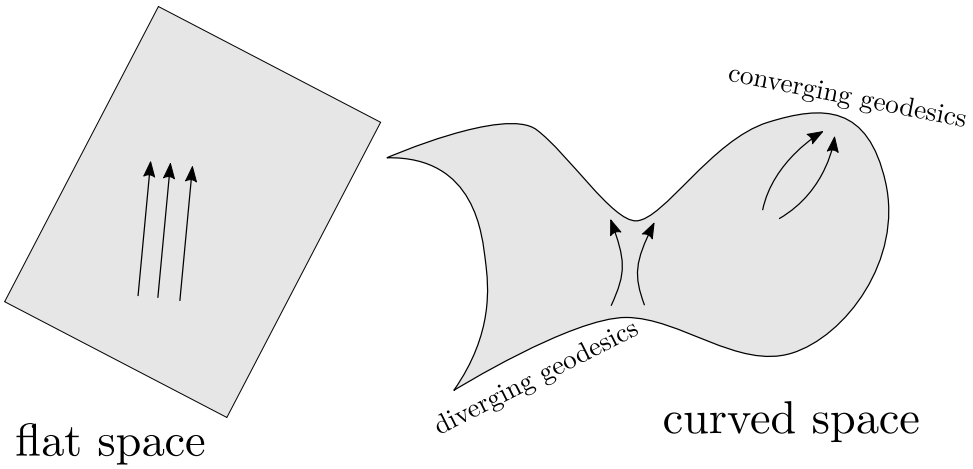

Figure 4.5: Curvature can cause nearby "parallel" geodesics to either converge or diverge.

Suppose that we have a one-parameter family of timelike geodesics, given in local coordinates by $x^a(\tau) = \gamma^a(\tau, s)$. Here τ is the proper time along each geodesic, while the geodesics are labelled by a continuous parameter s (figure 4.6).

The tangent vector to the curve γ is X, with components X^a. We can also define the vector J. This is not a vector field, but a vector defined along each of the curves γ, with components

$$J^a = \left.\frac{\partial \gamma^a}{\partial s}\right|_\tau .$$

J is sometimes called a *deviation vector* or a *Jacobi field*.

Along each geodesic we are free to choose the origin of the proper time, i.e. the point at which $\tau = 0$. Under the reparametrisation $\tau(s) \mapsto \tau(s) + b(s)$, we have

$$\begin{aligned} X &\mapsto X \\ J &\mapsto J + b' X. \end{aligned}$$

Hence we can use our ability to "slide" the point $\tau = 0$ up or down each geodesic γ to ensure that, at $\tau = 0$, we have $g(J, X) = 0$. Another way to think of this process is to consider taking a "slice" through the two dimensional surface made up of the geodesics γ, and to choose the slide so that, at every point, we are slicing orthogonally to the vector field X; we then choose $\tau = 0$ along every geodesic to be the point at which the geodesic cuts through the slice.

Note that the vectors X and J commute: we have

$$\begin{aligned} [X, J]^a &= X^b \partial_b J^a - J^b \partial_b X^a \\ &= \frac{\partial}{\partial \tau} J^a - \frac{\partial}{\partial s} X^a \\ &= \frac{\partial^2 x^a}{\partial \tau \partial s} - \frac{\partial^2 x^a}{\partial s \partial \tau} \\ &= 0. \end{aligned}$$

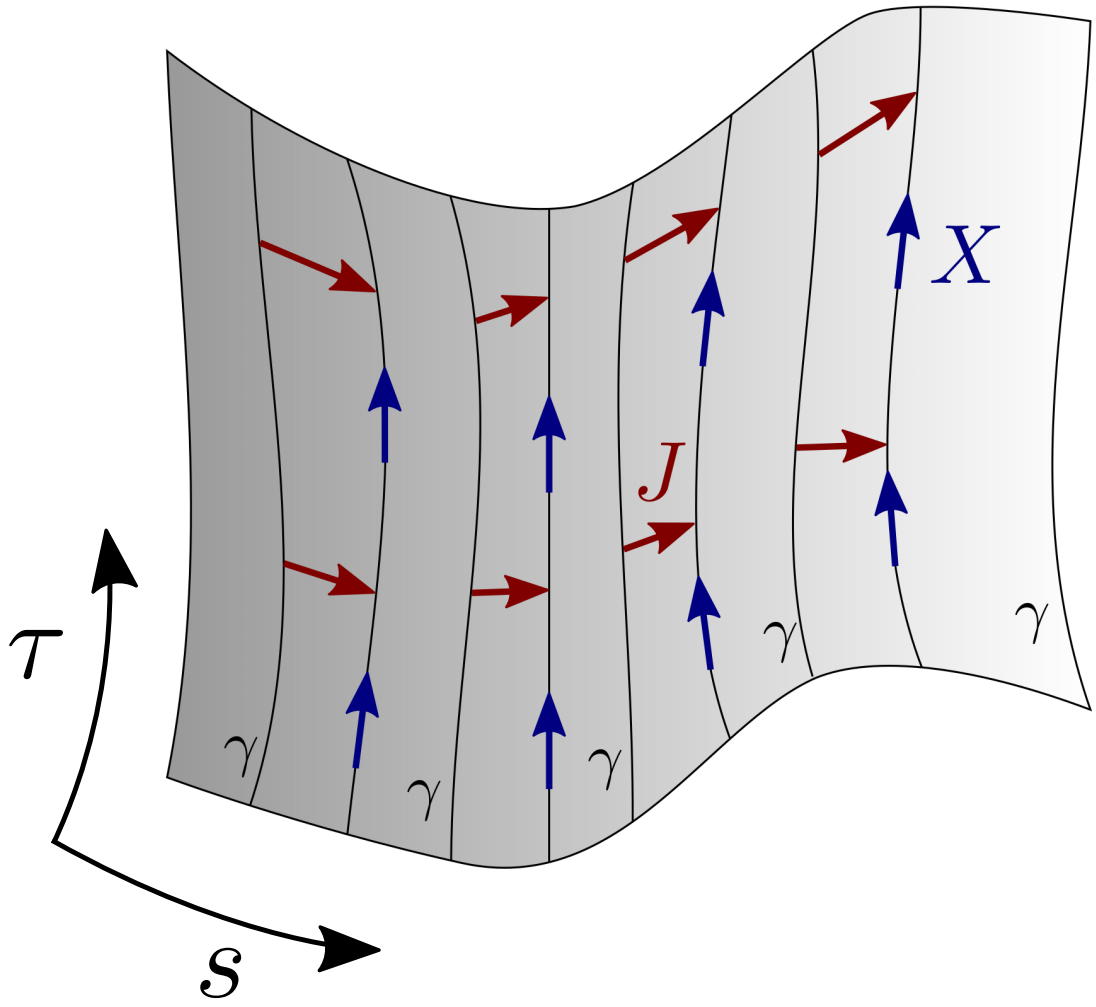

Figure 4.6: A congruence of timelike geodesics $\gamma(\tau, s)$, where τ is the proper time along a geodesic and s labels the different geodesics. Note that this congruence (locally) defines a 2 dimensional surface in spacetime. At each point on this surface, X is the tangent vector to the geodesic and J is a Jacobi field, which commutes with X. The acceleration of J along a timelike geodesic measures the *geodesic deviation*.

Note also that the value of $g(X, J)$ is invariant along each of the geodesics:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}g(X,J) &= X\left(g_{\mu\nu}X^\mu J^\nu\right)\\ &= \nabla_X\left(g_{\mu\nu}X^\mu J^\nu\right)\\ &= X_\mu\nabla_X J^\mu\\ &= X_\mu\nabla_J X^\mu + X_\mu[X,J]^\mu\\ &= \frac{1}{2}\nabla_J(X_\mu X^\mu)\\ &= \frac{1}{2}\nabla_J(-1) = 0,\end{aligned}$$

and so X and J will remain orthogonal along each of the geodesics. With this in mind, we can think of J as a "connecting vector", measuring the infinitesimal displacement of the geodesic $\gamma(\cdot, s+\epsilon)$ from the geodesic $\gamma(\cdot, s)$ (figure 4.6).

Now, we can compute the "acceleration" of the Jacobi field along each of the geodesics. We can think of this as measuring the acceleration of an infinitesimally displaced geodesic. We find[8]

$$\begin{aligned}\frac{\partial^2}{\partial\tau^2}J &= \nabla_X\nabla_X J\\ &= \nabla_X\nabla_J X \qquad \text{(using the torsion-free property and the fact that } J \text{ and } X \text{ commute)}\\ &= \left(\nabla_X\nabla_J - \nabla_J\nabla_X - \nabla_{[X,J]}\right)X \qquad \text{(using the geodesic equation } \nabla_X X = 0 \text{ and the fact}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{that } J \text{ and } X \text{ commute).}\end{aligned} \tag{4.5}$$

Why have we added these additional terms which vanish? The point is that the object in the final line can be used to define a tensor field. We first define the vector field

$$R(X,Y)Z := \left(\nabla_X\nabla_Y - \nabla_Y\nabla_X - \nabla_{[X,Y]}\right)Z,$$

where X, Y and Z are vector fields. For fixed X, Y, $R(X,Y)$ can be thought of as a map from vector fields to vector fields, taking Z to $R(X,Y)Z$.

This map is C^∞-linear in each of its arguments. First, note that it is antisymmetric in its first two arguments: we have

$$R(X,Y)Z = -R(Y,X)Z,$$

so we only need to check linearity in one of these arguments. For smooth vector fields X, X', Y, Z and smooth scalar fields a, b, we have

$$\begin{aligned}R(aX+bX',Y)Z &= \left(a\nabla_X\nabla_Y + b\nabla_{X'}\nabla_Y - \nabla_Y(a\nabla_X + b\nabla_{X'}) - \nabla_{[aX+bX',Y]}\right)Z\\ &= \Big(a\nabla_X\nabla_Y + b\nabla_{X'}\nabla_Y - a\nabla_Y\nabla_X - b\nabla_Y\nabla_{X'} - (Y(a))\nabla_X - (Y(b))\nabla_{X'}\\ &\qquad - \nabla_{a[X,Y]-Y(a)X+b[X',Y]-Y(b)X'}\Big)Z\\ &= \left(a\nabla_X\nabla_Y - a\nabla_Y\nabla_X - a\nabla_{[X,Y]} + b\nabla_{X'}\nabla_Y - b\nabla_Y\nabla_{X'} - b\nabla_{[X',Y]}\right)Z\\ &= aR(X,Y)Z + bR(X',Y)Z.\end{aligned}$$

so $R(X,Y)Z$ is C^∞ linear in X and Y.

We can also check that $R(X,Y)$ is a C^∞ linear map from vector fields to vector fields. It is easy to

[8] It is not clear how to interpret the first term in this equation, $\partial_\tau^2 J$. If J were a scalar field then the first line would hold, since then $X(J) = \partial_\tau J$. However, J is a vector field: we can choose a certain kind of "frame", with respect to which the components of J satisfy $X(X(J^a)) = (\nabla_X\nabla_X J)^a$, but this will not be true in all coordinate systems. To avoid unnecessary details, perhaps it is better to think of this formula as being *suggestive* of the acceleration of the vector J in the X direction.

see that $R(X,Y)(Z+Z') = R(X,Y)Z + R(X,Y)Z'$, so we just need to check

$$\begin{aligned}
R(X,Y)(aZ) &= (\nabla_X\nabla_Y - \nabla_Y\nabla_X - \nabla_{[X,Y]})(aZ) \\
&= a\nabla_X\nabla_Y Z + X(a)\nabla_Y Z + Y(a)\nabla_X Z + X(Y(a))Z \\
&\quad - a\nabla_Y\nabla_X Z - Y(a)\nabla_X Z - X(a)\nabla_Y Z - Y(X(a))Z - a\nabla_{[X,Y]}Z - ([X,Y](a))\,Z \\
&= a\left(\nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z\right) + (X(Y(a)) - Y(X(a)) - [X,Y](a)) \\
&= aR(X,Y)Z.
\end{aligned}$$

so $R(X,Y)$ is a linear map from vector fields to vector fields, which is also linear in both X and Y.

We can use this to define a $(1,3)$ tensor field, called the *Riemann curvature tensor* (or just the *Riemann tensor*). Acting on a covector field η and three vector fields X, Y, Z, the Riemann curvature tensor R is defined[9] as

$$R(\eta, Z, X, Y) = \eta\left(R(X,Y)Z\right).$$

Returning now to the equation governing Jacobi fields (4.5), we find that a Jacobi field satisfies the ODE

$$\frac{\partial^2}{\partial\tau^2}J = R(X,J)X,$$

so it is the Riemann curvature which governs the deviation of nearby geodesics. In flat space, the Riemann curvature vanishes!

You might prefer to work in abstract indices. Working through everything in components, we find that, for any vector field X

$$(\nabla_\mu\nabla_\nu - \nabla_\nu\nabla_\mu)X^\alpha = R^\alpha{}_{\beta\mu\nu}X^\beta.$$

What happens if we commute covariant derivatives and apply them to tensors of various types? We know what happens when we apply $[\nabla_\mu, \nabla_\nu]$ to a vector field – we get out the Riemann curvature tensor. When we apply this operator to a scalar field f, working in local coordinates we obtain

$$\begin{aligned}
[\nabla_a, \nabla_b]f &= \nabla_a\partial_b f - \nabla_b\partial_a f \\
&= \partial_a\partial_b f - \partial_b\partial_a f + \Gamma^c_{ab}\partial_c f - \Gamma^c_{ba}\partial_c f \\
&= 0,
\end{aligned}$$

using the torsion-free property of the connection. We can use this to work out the corresponding expression for a covector η (**exercise**)

$$[\nabla_\mu, \nabla_\nu]\eta_\rho = -R^\sigma{}_{\rho\mu\nu}\eta_\sigma,$$

and for a general tensor

$$\begin{aligned}
[\nabla_\mu, \nabla_\nu]T^{\rho_1\rho_2\ldots\rho_n}{}_{\sigma_1\sigma_2\ldots\sigma_m} &= R^{\rho_1}{}_{\kappa\mu\nu}T^{\kappa\rho_2\ldots\rho_n}{}_{\sigma_1\sigma_2\ldots\sigma_m} + R^{\rho_2}{}_{\kappa\mu\nu}T^{\rho_1\kappa\ldots\rho_n}{}_{\sigma_1\sigma_2\ldots\sigma_m} \\
&\quad + \ldots + R^{\rho_n}{}_{\kappa\mu\nu}T^{\rho_1\rho_2\ldots\kappa}{}_{\sigma_1\sigma_2\ldots\sigma_m} \\
&\quad - R^\kappa{}_{\sigma_1\mu\nu}T^{\rho_1\rho_2\ldots\rho_n}{}_{\kappa\sigma_2\ldots\sigma_m} - R^\kappa{}_{\sigma_2\mu\nu}T^{\rho_1\rho_2\ldots\rho_n}{}_{\sigma_1\kappa\ldots\sigma_m} \\
&\quad - \ldots - R^\kappa{}_{\sigma_m\mu\nu}T^{\rho_1\rho_2\ldots\rho_n}{}_{\sigma_1\sigma_2\ldots\kappa}.
\end{aligned}$$

4.10 Symmetries of the Riemann tensor

4.10.1 Algebraic symmetries

We already know, from the definition of the Riemann tensor, that it is antisymmetric in its final two indices:

$$R^\mu{}_{\nu\rho\sigma} = -R^\mu{}_{\nu\sigma\rho}.$$

[9] You should be careful when consulting references, as there are different conventions regarding the sign of the Riemann tensor.

From the metric compatibility condition we also find

$$\begin{aligned} 0 &= -[\nabla_\mu, \nabla_\nu] g_{\rho\sigma} \\ &= R^\alpha{}_{\rho\mu\nu} g_{\alpha\sigma} + R^\alpha{}_{\sigma\mu\nu} g_{\rho\alpha} \\ &= R_{\sigma\rho\mu\nu} + R_{\rho\sigma\mu\nu}, \end{aligned}$$

so the Riemann tensor is also antisymmetric in its first two indices.

Next consider the following expression, for some scalar field f

$$\nabla_\mu \nabla_\sigma \nabla_\nu f + \nabla_\sigma \nabla_\nu \nabla_\mu f + \nabla_\nu \nabla_\mu \nabla_\sigma f - \nabla_\sigma \nabla_\mu \nabla_\nu f - \nabla_\mu \nabla_\nu \nabla_\sigma f - \nabla_\nu \nabla_\sigma \nabla_\mu f$$

We can group these terms in two different ways: first,

$$\nabla_\mu \left([\nabla_\sigma, \nabla_\nu] f \right) + \nabla_\nu \left([\nabla_\mu, \nabla_\sigma] f \right) + \nabla_\sigma \left([\nabla_\nu, \nabla_\mu] f \right) = 0$$

since $[\nabla_\alpha, \nabla_\beta] f = 0$. On the other hand, we can group these terms as follows:

$$[\nabla_\sigma, \nabla_\nu] \nabla_\mu f + [\nabla_\mu, \nabla_\sigma] \nabla_\nu f + [\nabla_\nu, \nabla_\mu] \nabla_\sigma f = \left(R^\alpha{}_{\mu\nu\sigma} + R^\alpha{}_{\nu\sigma\mu} + R^\alpha{}_{\sigma\mu\nu} \right) \nabla_\alpha f$$

where we have used antisymmetry in the last pair of indices. But since this holds for *all* scalar functions f, we have

$$R^\alpha{}_{\mu\nu\sigma} + R^\alpha{}_{\nu\sigma\mu} + R^\alpha{}_{\sigma\mu\nu} = 0 \tag{4.6}$$

This equation is sometimes known as the *first Bianchi identity* or the *algebraic Bianchi identity*.

There is another useful symmetry of the Riemann tensor which follows from the symmetries derived above. Using the first Bianchi identity and cyclicly permuting indices, we have

$$\begin{aligned} R_{\mu\nu\rho\sigma} + R_{\mu\rho\sigma\nu} + R_{\mu\sigma\nu\rho} &= 0 \\ -R_{\nu\rho\sigma\mu} - R_{\nu\sigma\mu\rho} - R_{\nu\mu\rho\sigma} &= 0 \\ -R_{\rho\sigma\mu\nu} - R_{\rho\mu\nu\sigma} - R_{\rho\nu\sigma\mu} &= 0 \\ R_{\sigma\mu\nu\rho} + R_{\sigma\nu\rho\mu} + R_{\sigma\rho\mu\nu} &= 0 \end{aligned}$$

Adding these four equations together, and using antisymmetry in the first and last pair of indices, we find that

$$R_{\mu\nu\rho\sigma} = R_{\rho\sigma\mu\nu}$$

In summary, the Riemann tensor has the following algebraic symmetries:

$$\begin{aligned} R_{\mu\nu\rho\sigma} &= -R_{\mu\nu\sigma\rho} \\ R_{\mu\nu\rho\sigma} &= -R_{\nu\mu\rho\sigma} \\ R_{\mu\nu\rho\sigma} &= R_{\rho\sigma\mu\nu} \\ R_{\mu\nu\rho\sigma} + R_{\mu\rho\sigma\nu} + R_{\mu\sigma\nu\rho} &= 0 \end{aligned}$$

4.10.2 The (second) Bianchi identity

There is also an important symmetry of the *derivatives* of the Riemann tensor, called the *second Bianchi identity* or simply the *Bianchi identity*.

We prove this identity in a similar way to the first Bianchi identity. Consider the following expression, for some arbitrary covector η

$$\nabla_\mu \nabla_\rho \nabla_\nu \eta_\sigma + \nabla_\rho \nabla_\nu \nabla_\mu \eta_\sigma + \nabla_\nu \nabla_\mu \nabla_\rho \eta_\sigma - \nabla_\mu \nabla_\nu \nabla_\rho \eta_\sigma - \nabla_\nu \nabla_\rho \nabla_\mu \eta_\sigma - \nabla_\rho \nabla_\mu \nabla_\nu \eta_\sigma$$

Grouping the terms in one way we obtain

$$\begin{aligned}[\nabla_\mu,\nabla_\rho]\nabla_\nu\eta_\sigma+[\nabla_\rho,\nabla_\nu]\nabla_\mu\eta_\sigma+[\nabla_\nu,\nabla_\mu]\nabla_\rho\eta_\sigma &= R^\alpha{}_{\nu\rho\mu}\nabla_\alpha\eta_\sigma+R^\alpha{}_{\sigma\rho\mu}\nabla_\nu\eta_\alpha+R^\alpha{}_{\mu\nu\rho}\nabla_\alpha\eta_\sigma\\ &\quad+R^\alpha{}_{\sigma\nu\rho}\nabla_\mu\eta_\alpha+R^\alpha{}_{\rho\mu\nu}\nabla_\alpha\eta_\sigma+R^\alpha{}_{\sigma\mu\nu}\nabla_\rho\eta_\alpha\\ &=\left(R^\alpha{}_{\mu\nu\rho}+R^\alpha{}_{\nu\rho\mu}+R^\alpha{}_{\rho\mu\nu}\right)\nabla_\alpha\eta_\sigma\\ &\quad+R^\alpha{}_{\sigma\nu\rho}\nabla_\mu\eta_\alpha+R^\alpha{}_{\sigma\rho\mu}\nabla_\nu\eta_\alpha+R^\alpha{}_{\sigma\mu\nu}\nabla_\rho\eta_\alpha\\ &=R^\alpha{}_{\sigma\nu\rho}\nabla_\mu\eta_\alpha+R^\alpha{}_{\sigma\rho\mu}\nabla_\nu\eta_\alpha+R^\alpha{}_{\sigma\mu\nu}\nabla_\rho\eta_\alpha\end{aligned}$$

where we have used the first Bianchi identity. But, grouping the same terms in an alternative way, we have

$$\begin{aligned}\nabla_\mu[\nabla_\rho,\nabla_\nu]\eta_\sigma+\nabla_\nu[\nabla_\mu,\nabla_\rho]\eta_\sigma+\nabla_\rho[\nabla_\nu,\nabla_\mu]\eta_\sigma &=\nabla_\mu\left(R^\alpha{}_{\sigma\nu\rho}\eta_\alpha\right)+\nabla_\nu\left(R^\alpha{}_{\sigma\rho\mu}\eta_\alpha\right)+\nabla_\rho\left(R^\alpha{}_{\sigma\mu\nu}\eta_\alpha\right)\\ &=R^\alpha{}_{\sigma\nu\rho}\nabla_\mu\eta_\alpha+R^\alpha{}_{\sigma\rho\mu}\nabla_\nu\eta_\alpha+R^\alpha{}_{\sigma\mu\nu}\nabla_\rho\eta_\alpha\\ &\quad+\nabla_\mu R^\alpha{}_{\sigma\nu\rho}\eta_\alpha+\nabla_\nu R^\alpha{}_{\sigma\rho\mu}\eta_\alpha+\nabla_\rho R^\alpha{}_{\sigma\mu\nu}\eta_\alpha\end{aligned}$$

Combining these two equations and reordering the indices a bit using the symmetries of the Riemann tensor, we find that

$$\left(\nabla_\mu R_{\nu\rho}{}^\alpha{}_\sigma+\nabla_\nu R_{\rho\mu}{}^\alpha{}_\sigma+\nabla_\rho R_{\mu\nu}{}^\alpha{}_\sigma\right)\eta_\alpha=0$$

Since this holds for *all* covectors η (and since the connection is metric-compatible), we have proved the *second Bianchi identity*

$$\nabla_\mu R_{\nu\rho\alpha\beta}+\nabla_\nu R_{\rho\mu\alpha\beta}+\nabla_\rho R_{\mu\nu\alpha\beta}=0 \tag{4.7}$$

4.10.3 The Ricci and Einstein tensors and the contracted Bianchi identity

We can contract a pair of indices in the Riemann tensor to form the *Ricci curvature tensor* (or simply *Ricci tensor*), which is also conventionally notated with the letter R:

$$R_{\mu\nu}:=R^\alpha{}_{\mu\alpha\nu}$$

The symmetries of the Riemann tensor imply that the Ricci tensor is symmetric (**exercise**).

We can contract the indices of the Ricci tensor to form the *scalar curvature*, which is also conventionally notated[10] by the letter R:

$$R:=(g^{-1})^{\mu\nu}R_{\mu\nu}$$

If we contract the indices μ and α in the second Bianchi identity (4.7) and then relabel indices, we obtain the identity

$$\nabla^\alpha R_{\alpha\mu\nu\rho}-\nabla_\nu R_{\mu\rho}+\nabla_\rho R_{\mu\nu}=0$$

Contracting again, this time with the indices μ and ρ (and relabelling indices and dividing by two), we obtain the *contracted Bianchi identity*

$$\nabla^\mu\left(R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}\right)=0$$

This leads us to define the *Einstein tensor*:

$$G_{\mu\nu}:=R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}$$

From the calculations above, we have shown that the Einstein tensor is *divergence free*

$$\nabla^\mu G_{\mu\nu}=0$$

[10] Because of these conventional notations, when dealing with the curvature it is particularly useful to use abstract index notation rather than index-free notation.

4.11 Curvature in terms of the metric

There is one final aspect of the curvature which will be important for us, and that is its relationship to the metric tensor. If we work in some local coordinates x^a, then we can write

$$\begin{aligned}
[\nabla_a, \nabla_b]X^c &= R^c{}_{dab}X^d \\
&= \partial_a \nabla_b X^c - \Gamma^d_{ab}\nabla_d X^c + \Gamma^c_{ad}\nabla_b X^d \quad - (a \leftrightarrow b) \\
&= \partial_a\partial_b X^c + \partial_a\left(\Gamma^c_{bd}X^d\right) - \Gamma^d_{ab}\partial_d X^c + \Gamma^c_{ad}\partial_b X^d - \Gamma^d_{ab}\Gamma^c_{de}X^e + \Gamma^c_{ad}\Gamma^d_{be}X^e \quad - (a \leftrightarrow b) \\
&= \left(\partial_a\Gamma^c_{bd} - \partial_b\Gamma^c_{ad} + \Gamma^c_{ae}\Gamma^e_{bd} - \Gamma^c_{be}\Gamma^e_{ad}\right)X^d
\end{aligned}$$

and so the components of the Riemann tensor can be written in terms of the Christoffel symbols and their derivatives as follows:

$$R^a{}_{bcd} = \partial_c\Gamma^a_{bd} - \partial_d\Gamma^a_{bc} + \Gamma^a_{ce}\Gamma^e_{bd} - \Gamma^a_{de}\Gamma^e_{bc} \tag{4.8}$$

Recalling the expression for the Christoffel symbols of the Levi-Civita connection in terms of the metric components

$$\Gamma^a_{bc} = \frac{1}{2}(g^{-1})^{ad}\left(\partial_b g_{cd} + \partial_c g_{bd} - \partial_d g_{bc}\right)$$

We can substitute this expression into equation (4.8) to obtain a long and not very enlightening equation for the components of the Riemann tensor.

The important thing to notice about this expression is the following: ***the Riemann tensor depends on the metric*** g ***and its first two derivatives***.

Chapter 5

The Einstein equations and physics in curved spacetimes

We now have all the mathematical machinery needed to understand the Einstein equations, as well as to adapt other physical laws to curved spacetimes.

First, the Einstein equations. Remember that there are many hints that matter causes spacetime to curve, but, to be consistent with special relativity, it should not be just the matter density which affects curvature, but the energy density. This appears in the energy-momentum tensor $T_{\mu\nu}$.

With this in mind, in October 1915 Einstein tried the equation

$$R_{\mu\nu} = CT_{\mu\nu}$$

for some constant C. But there is a problem with this equation: the conservation of energy-momentum means that the energy momentum tensor $T_{\mu\nu}$ is divergence-free[1], while the divergence of the Ricci tensor generally does not vanish.

By November, Einstein had remedied this problem in the obvious way, replacing the Ricci tensor with the Einstein tensor. This yields the equation

$$G_{\mu\nu} = CT_{\mu\nu}.$$

We still need to fix the constant C. This can be done by taking the *weak field limit*, where we take the metric to have the form

$$g_{ab} = m_{ab} + \epsilon h_{ab}$$

in some coordinates. We then expand the Einstein equations up to first order in ϵ and then compare the Einstein equations with Newtonian gravity (see the *GR2 course* for the details). The upshot of this calculation is that $C = 8\pi$. This leads to the *Einstein equations* (or the *Einstein field equations*)

$$G_{\mu\nu} = 8\pi T_{\mu\nu}. \tag{5.1}$$

Restoring the speed of light and Newton's constant, this is

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}.$$

[1] In special relativity, the energy momentum tensor satisfied $\partial^a T_{ab} = 0$, working in inertial coordinates. This was motivated by an argument involving integrating over a region of spacetime. This argument can be repeated in a curved spacetime, but this would require us to develop the theory of integration on manifolds. The upshot is that, as might be expected, the partial derivative should be replaced by a covariant derivative.

5.1 Uniqueness of the Einstein equations and the cosmological constant

There is a sense in which the Einstein equations are "unique". This is given by *Lovelock's theorem* (the proof of which is well beyond the scope of this course)

Theorem (Lovelock's theorem)**.** *In four spacetime dimensions, the only tensor fields constructed entirely from the metric tensor together with its first and second derivatives which are symmetric and divergence-free are of the form*

$$aG_{\mu\nu} + bg_{\mu\nu},$$

where a and b are constants.

This suggests the following alternative for the Einstein equations, which Einstein published in 1917

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi T_{\mu\nu}. \tag{5.2}$$

The constant Λ in this equation is called the *cosmological constant*. Consequently, equation (5.2) are sometimes called the *Einstein equations with cosmological constant*.

Einstein originally included the cosmological constant because, without it, he couldn't find cosmological solutions[2] of the Einstein equations which didn't either expand or contract. When later observations showed that the universe *is* expanding, Einstein called it his "greatest mistake". More recent observations indicate that the cosmological constant is nonzero, but with an incredibly small positive value: in Planck units, $\Lambda \approx 7.26 \times 10^{-121}$.

5.2 The Einstein equations as a system of PDEs

In a local system of coordinates, the Riemann curvature tensor can be written in terms of the metric and its first two derivatives. Hence the Einstein equations can be viewed as a second order system of PDEs for the metric components g_{ab}.

Usually we have to supplement these equations with the equations of motion for the matter in order to obtain a closed system of equations. But there is a special case, called the *Einstein vacuum equations*, where there is no matter present:

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 0. \tag{5.3}$$

In Newtonian gravity, when there is no matter present, the gravitational potential ϕ vanishes and there are no dynamics. The situation is very different in GR, where the Einstein vacuum equations (5.3) are a second order system of equations for the metric, which have many nontrivial solutions!

How do we solve these equations, in general? As with other second order PDEs, it is useful to try to characterise these equations as *elliptic* (like the Laplace equation), *parabolic* (like the heat equation) or *hyperbolic* (like the wave equation). In the first case we would expect to have to specify boundary conditions, while in the second or third case we would expect to specify initial conditions and then treat the PDEs as evolution equations.

Unfortunately, the Einstein equations are not of any specific type. What's more, if we try to view these equations as evolution equations, e.g. specifying the metric components g_{ab} and their time derivatives everywhere on some initial time surface, then we find that *there is no unique solution.* Disaster!

What could have gone wrong? Remember that we are viewing the equations as a system of PDEs for the components of the metric g_{ab} in some coordinate system x^a. But consider another coordinate system

[2] See chapter 7.

$y^{a'}$, *which agrees with the coordinate system* x^a *in a neighbourhood of the initial hypersurface.* The metric in these coordinates has components $g'_{a'b'}$, which agree with the components g_{ab} in a neighbourhood of the initial surface, but which will generally *differ* away from this neighbourhood (see figure 5.1).

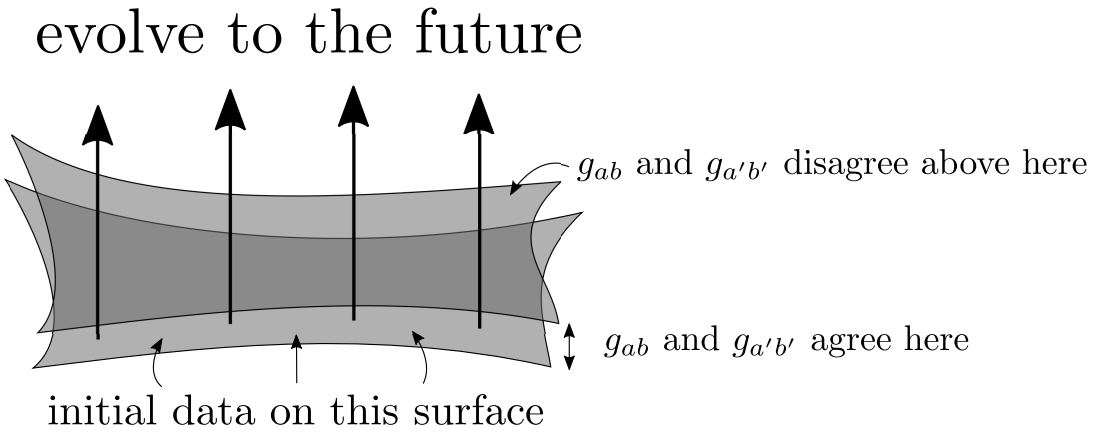

Figure 5.1: Initial data for the metric is given on the bottom surface, and we try to solve 'upwards'. One solution to the Einstein equations, which agrees with this initial data, is given by the metric g_{ab}. Another metric is given by $g'_{a'b'}$, whose components agree with the components of g_{ab} on the initial hypersurface and for some amount of 'time', but then disagree at later times. However, they only differ by a coordinate transformation, so $g'_{a'b'} = \frac{\partial x^a}{\partial y^{a'}} \frac{\partial x^b}{\partial x^{b'}} g_{ab}$, where $x^a = y^a$ near the initial hypersurface. Since the Einstein equations are invariant under coordinate transformations, both g_{ab} and $g_{a'b'}$ are solutions to the Einstein equations – so the solutions cannot be unique!

The Einstein equations are tensorial in nature. This means that they have the same form in every coordinate system. Consequently, if g_{ab} is a solution to the Einstein equations, then so is $g_{a'b'}$, where this solution differs from the first one by a coordinate transformation. From this point of view it is obvious that the Einstein equations cannot have a unique solution!

The solution to our problem is now obvious: we have to make some specific choice of coordinates! One convenient choice is to choose the coordinate functions to satisfy wave equations themselves ("wave coordinates"), in which case the Einstein equations become a *system of nonlinear wave equations.* Now we see that the Einstein equations are hyperbolic, and we can think about solving the *Cauchy problem*: pose initial data on some initial hypersurface, and then use the Einstein equations to evolve this data into the future.

Amazingly, it wasn't until almost 40 years after the Einstein equations were published when these issues were fully understood and solved (by Choquet-Bruhat in 1952). Although the details of these calculations are not necessary for our purposes, this point of view – viewing the Einstein equations as a system of evolution equations – is crucial both for a rigorous approach to GR, and for numerical GR.

5.3 Other physical laws in curved spacetimes

How do we generalise other physical laws to a curved spacetime? First, the kinematics of point particles is governed by the *geodesic postulate*: *"test particles" move along timelike geodesics in the absence of external forces.* In general, however, we should avoid working with point masses in general relativity: if you take some matter distribution with a fixed mass and then squeeze it into a smaller and smaller volume, GR predicts that at some point it will form a black hole! Also, if the test particle had mass or energy, then this should be included in the energy momentum tensor on the right hand side of the Einstein equations (this is sometimes called *back reaction*) and so it would affect the metric tensor and hence the geodesics. For these reasons, "test particles" in GR really just *mean* timelike geodesics (sometimes we talk about "massless test particles", which follow null geodesics).

To generalise other physical laws, we recall that, at any point p on the manifold, we can choose to work in normal coordinates. In these coordinates, the components of the metric g take the same values (at the point p) as the Minkowski metric m, and the Christoffel symbols vanish so partial derivatives and covariant derivatives are the same. So, at the point p, working in normal coordinates, physics should look like the physics of special relativity.

As an example, consider Maxwell's equations, which can be written (in inertial coordinates) in special relativity as

$$\begin{aligned} \partial_a F_{bc} + \partial_b F_{ca} + \partial_c F_{ab} &= 0 \\ (m^{-1})^{ab} \partial_a F_{bc} &= 0. \end{aligned}$$

These equations should take an identical form form in normal coordinates at the point p. But, in normal coordinates, these equations are equivalent to

$$\begin{aligned} \nabla_a F_{bc} + \nabla_b F_{ca} + \nabla_c F_{ab} &= 0 \\ (g^{-1})^{ab} \nabla_a F_{bc} &= 0, \end{aligned}$$

since partial derivatives are covariant derivatives and $g_{ab} = m_{ab}$ at p. But now, these equations only involve tensors, so we can write them using abstract indices

$$\begin{aligned} \nabla_\mu F_{\nu\rho} + \nabla_\nu F_{\rho\mu} + \nabla_\rho F_{\mu\nu} &= 0 \\ \nabla^\mu F_{\mu\nu} &= 0. \end{aligned}$$

This is how to generalise Maxwell's equations to a curved spacetime.

This illustrates the following general point: to generalise a physical law from special relativity to general relativity, we should

1. express the physical law in terms of tensors in Minkowski space,
2. replace all partial derivatives with covariant derivatives, and
3. replace the Minkowski metric m with the metric g.

The principle that all physical laws should be expressed as tensorial equations, encapsured by these rules, is sometimes called the *principle of covariance*.

Chapter 6

The Schwarzschild metric

> Hence, according to article 10, if the emi-diameter of a phær of the ame denity with the un were to exceed that of the un in the proportion of 500 to 1, a body falling from an infinite height towards it, would have acquired at its urface a greater velocity than that of light, and conequently, uppoing light to be attracted by the ame force in proportion to its vis inertiæ, with other bodies, all light emitted from uch a body would be made to return towards it, by its own proper gravity.
>
> John Michelle, *letter to the Royal Society*, 1784.

One solution to the Einstein vacuum equations (with $\Lambda = 0$) is Minkowski space, where we set $g_{ab} = m_{ab}$. But we already know everything about this spacetime - this is just the spacetime of special relativity!

What other solutions does the Einstein equations have? We have already described a way of generating many solutions to the equations, by choosing some initial data and treating the Einstein equations as evolution equations. Unfortunately, in most cases it is impossible to solve these equations explicitly, due to the highly nonlinear nature of the equations.

Alternatively, we can look for solutions in particular symmetry classes, where the Einstein equations simplify. There is a long and distinguished history of this approach in physics – its first major success came just one year after the Einstein equations were published, with the discovery of the *Schwarzschild solution*.

This is a solution to the Einstein vacuum equations without a cosmological constant, i.e. $G_{\mu\nu} = 0$. It is *static*, *spherically symmetric* and *asymptotically flat*. This means that we can write the metric in coordinates (t, r, θ, ϕ), where

1. the components of the metric are independent of t (*stationarity*).
2. The metric is also invariant under $t \mapsto -t$ (*staticity*).
3. The components of the metric are invariant under a family of transformations that can be parametrised by $SO(3)$ matrices, whose orbits are topological spheres (*spherical symmetry*).
4. As $r \to \infty$, the metric components approach[1] the components of the Minkowski metric written in spherical polar coordinates (*asymptotic flatness*).

It might not be obvious that point (2) adds anything new to point (1), i.e. that "static" means anything more than "stationary'. To see how these two points are different, consider the two dimensional metric

$$g = -\mathrm{d}t\mathrm{d}\theta$$

[1] Asymptotic flatness actually requires that the metric approaches the Minkowski metric at a certain rate, but there are various possible rates and we will not go into the messy details here.

This metric is stationary: the nonzero metric components are just $g_{t\theta} = -\frac{1}{2}$ and $g_{\theta t} - \frac{1}{2}$, which are obviously independent of t. However, under the map $t \mapsto -t$, the metric transforms as $g \mapsto -g$, so the metric is *not* invariant under time reversal!

For a more physical example, consider sphere rotating at a constant speed in an otherwise empty universe. This situation is stationary: it looks the same at all points in time. But if we reverse the direction of time, then the situation is not invariant, since then the sphere rotates in the opposite direction!

There is a famous theorem, which shows that the Schwarzschild solution is, in a sense, unique:

Theorem (Birkhoff's theorem)**.** *Every spherically symmetric solution to the Einstein vacuum equations is locally isometric to either Schwarzschild spacetime or to Minkowski space.*

Locally isometric means that for all sufficiently small open sets can be mapped onto an open set in the Schwarzschild spacetime (or Minkowski space) by an isometry – that is, the map preserves the metric, so that all inner products between vector fields are preserved. Basically, Birkhoff's theorem says that the only spherically symmetric solutions to the Einstein equations are Minkowski space and the Schwarzschild spacetime, up to topology.

This theorem is important because it means that the Schwarzschild metric characterises spacetime outside of *any* spherically symmetric matter distribution, regardless of the interior structure of the matter (e.g. the density profile of some fluid). In this case, the metric will *not* agree with the Schwarzschild metric *inside* the matter distribution (where $T_{\mu\nu} \neq 0$), where the metric will generally depend on the specific details of the matter.

6.1 The metric

The metric, written in the canonical coordinates described above, is

$$g = -\left(1 - \frac{2M}{r}\right) \mathrm{d}t^2 + \left(1 - \frac{2M}{r}\right)^{-1} \mathrm{d}r^2 + r^2 \left(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2\right) \tag{6.1}$$

here $M > 0$ is a constant (called the *mass*), and the ranges of the coordinates are as follows:

- $t \in \mathbb{R}$
- $0 < \theta < \pi$
- $0 < \phi < 2\pi$
- The range of the coordinate r is a bit more complicated, but for now we'll avoid difficulties and say $2M < r < \infty$. For example, we might be considering the region outside a spherically symmetric star, where the radius of the star is much larger[2] than $2M$.

It is easy to check that this metric satisfies the properties claimed above: the metric components only depend on the coordinates r and θ, and as $r \to \infty$ the metric approaches that of Minkowski space. Checking that this metric is actually a solution to the Einstein equations is extremely tedious, but it is a calculation that everyone should do at one point in their lives (and then never again!).

There are well-understood degeneracies in the metric at $\theta = 0$, π and also at $\phi = 0$, 2π. These, of course, are the usual coordinate issues with the sphere, and reflect the fact that, technically, we need to use two coordinate charts to cover the sphere. On the other hand, something odd is clearly going on with the metric at $r = 0$ and at $r = 2M$. We'll return to this point later.

[2]Modelling the sun as spherically symmetric, the surface $r = 2M$ is around 1km from its centre – well inside the region where there is matter. For the Earth, this surface is around 1cm from the centre!

6.2 Gravitational redshift

The first phenomena we'll investigate is called *gravitational redshift*. Suppose there are two observers, Alice and Bob, who move along integral curves of the vector field ∂_t, at two different radii but with the same angular coordinates. Their worldlines, parametrised by the coordinate t, are

$$\begin{aligned} \text{Alice:} \quad & (t, r_A, \theta_0, \phi_0) \\ \text{Bob:} \quad & (t, r_B, \theta_0, \phi_0) \end{aligned}$$

where θ_0 and ϕ_0 are constants, and $2M < r_A < r_B$.

Suppose that Alice sends Bob regular signals, using light rays. Light rays travel along null geodesics, which we can also parametrise by the coordinate time t. Then, for radial light rays, since they are null we have

$$\begin{aligned} 0 &= g_{ab} \frac{\mathrm{d}x^a}{\mathrm{d}t} \frac{\mathrm{d}x^b}{\mathrm{d}t} \\ &= -\left(1 - \frac{2M}{r}\right) + \left(1 - \frac{2M}{r}\right)^{-1} \left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 \\ \Rightarrow \frac{\mathrm{d}r}{\mathrm{d}t} &= \left(1 - \frac{2M}{r}\right), \end{aligned}$$

so, integrating from $r = r_A$ when $t = t_A$ to $r = r_B$ when $t = t_B$, we have

$$\int_{r_A}^{r_B} \left(1 - \frac{2M}{r}\right)^{-1} \mathrm{d}r = t_A - t_B.$$

Importantly, the coordinate time difference $t_A - t_B$ is itself is a constant,*independent of the initial time when the signal was transmitted* (you could just read this off directly from the fact that the metric is stationary). So, if Alice sends repeated signals at (coordinate) time intervals Δt, then they will be received by Bob at (coordinate) time intervals Δt.

But we should always remember that t is just a coordinate, and has no intrinsic meaning. The amount of time that Alice and Bob *experience* as passing – the amount of time measured by their clocks – is the *proper time* along their worldlines, not the coordinate time.

What is the proper time along Alice and Bob's worldlines? Along a worldline where r, θ, ϕ are constants, we calculate

$$\begin{aligned} -1 &= g_{ab} \frac{\mathrm{d}x^a}{\mathrm{d}\tau} \frac{\mathrm{d}x^b}{\mathrm{d}\tau} \\ &= -\left(1 - \frac{2M}{r}\right) \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 \\ \Rightarrow \frac{\mathrm{d}t}{\mathrm{d}\tau} &= \left(1 - \frac{2M}{r}\right)^{-\frac{1}{2}}. \end{aligned}$$

So, along Alice and Bob's worldlines, differences in proper times satisfy

$$\Delta\tau_{A/B} = \left(1 - \frac{2M}{r_{A/B}}\right)^{\frac{1}{2}} \Delta t,$$

and the ratio of the emission frequency to the received frequency is

$$\frac{\Delta\tau_B}{\Delta\tau_A} = \frac{\left(1 - \frac{2M}{r_B}\right)^{\frac{1}{2}}}{\left(1 - \frac{2M}{r_A}\right)^{\frac{1}{2}}}.$$

Since $r_B > r_A$, $\Delta\tau_B > \Delta\tau_A$. So less time passes for Alice than for Bob: Bob receives the signals at a lower frequency than the emitted frequency. *Clocks run slower in a gravitational field.* This is *gravitational redshift*. Note that, as $r_A \to 2M$ the ratio of frequencies tends to infinity – a kind of *infinite redshift*, which we will interpret later.

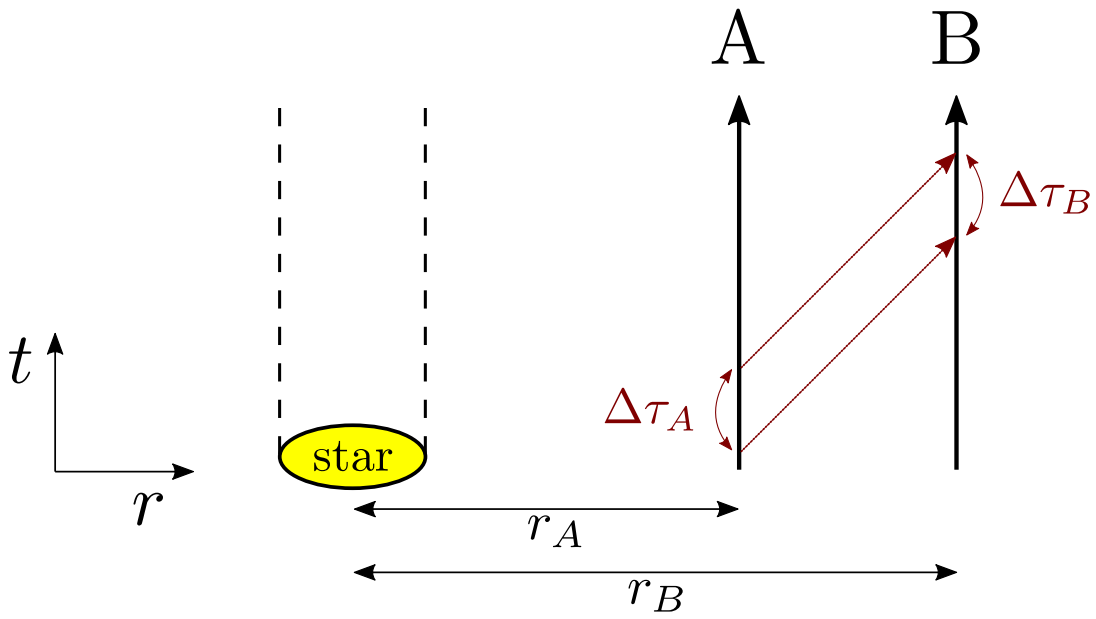

Figure 6.1: $\Delta\tau_B > \Delta\tau_A$: Bob receives signals from Alice at a slower rate than they are emitted.

6.3 Geodesics in the Schwarzschild metric

Before we can investigate other important phenomena in the Schwarzschild spacetime, we need to derive the equations for geodesics in this metric and examine some of their properties.

Rather than deriving the Christoffel symbols by differentiating the metric components directly, it is easier to start from the Lagrangian for point particles. If the Lagrangian itself is constant – which it is if we parametrise curves using an affine parameter – then we can either use the original Lagrangian L or alternatively L^2, since they will result in the same Euler-Lagrange equations[3] (**exercise**). So we can take the Lagrangian to be

$$\begin{aligned} L &= g_{ab}\frac{\mathrm{d}x^a}{\mathrm{d}\lambda}\frac{\mathrm{d}x^b}{\mathrm{d}\lambda} \\ &= -\left(1-\frac{2M}{r}\right)\dot{t}^2 + \left(1-\frac{2M}{r}\right)^{-1}\dot{r}^2 + r^2\dot{\theta}^2 + r^2\sin^2\theta\dot{\phi}^2, \end{aligned}$$

where the 'dots' represent derivatives with respect to some *affine* parameter λ.

As usual with Lagrangians, it is useful to start with conserved quantities. Since the Lagrangian is independent of t, we have the conserved 'energy'

$$\begin{aligned} E &:= -\frac{1}{2}\frac{\partial L}{\partial \dot{t}} \\ &= \left(1-\frac{2M}{r}\right)\dot{t}. \end{aligned}$$

[3]Using the 'squared' form of the Lagrangian is particularly useful for dealing with null geodesics, since varying this Lagrangian leads to the geodesic equation, while varying the original Lagrangian leads to problems due to the vanishing of the Lagrangian.

Similarly, the Lagrangian is independent of ϕ, so we have the conserved angular momentum about the z axis

$$\begin{aligned}\Omega &:= \frac{1}{2}\frac{\partial L}{\partial \dot{\phi}} \\ &= r^2 \sin^2\theta \dot{\phi}.\end{aligned}$$

The Lagrangian itself is constant: in the timelike case (i.e. for a massive particle) we can choose the affine parameter λ to be the proper time τ, while in the spacelike case we can choose the proper distance s, and so we have

$$-\left(1-\frac{2M}{r}\right)\dot{t}^2 + \left(1-\frac{2M}{r}\right)^{-1}\dot{r}^2 + r^2\dot{\theta}^2 + r^2\sin^2\theta\dot{\phi}^2 = -K = \begin{cases} -1 & \text{(timelike)} \\ 0 & \text{(null).}\end{cases}$$

Finally, we can use spherical symmetry to rotate the manifold so that the particle moves only in the equatorial plane $\theta = \frac{\pi}{2}$. To be more precise: we can use the $SO(3)$ isometries to rotate so that the particle is initially in the equatorial plane, and so its initial velocity (i.e. the tangent vector of the geodesic) is initially in the equatorial plane. Then the equation of motion for θ is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\left(r^2\dot{\theta}\right) - r^2\sin\theta\cos\theta\dot{\phi}^2 &= 0 \\ \Rightarrow r^2\ddot{\theta} + 2r\dot{r}\dot{\theta} - r^2\sin\theta\cos\theta\dot{\phi}^2 &= 0.\end{aligned}$$

So, if $\theta\big|_{\lambda=0} = \frac{\pi}{2}$ and $\dot{\theta}\big|_{\lambda=0} = 0$, we have $\ddot{\theta}\big|_{\lambda=0} = 0$. From this it follows that $\theta = \frac{\pi}{2}$ always.

Putting this all together, we find the evolution equation for the r coordinate

$$\begin{aligned}\left(1-\frac{2M}{r}\right)^{-1}\dot{r}^2 + \frac{\Omega^2}{r^2} + K &= \left(1-\frac{2M}{r}\right)^{-1}E^2 \\ \Rightarrow \frac{1}{2}\dot{r}^2 + \frac{\Omega^2}{2r^2}\left(1-\frac{2M}{r}\right) - \frac{MK}{r} &= \frac{E^2-K}{2}.\end{aligned} \tag{6.2}$$

This is the equation of motion of a particle with energy $\frac{1}{2}(E^2-K)$, moving in an *effective potential*

$$V(r) = -\frac{MK}{r} + \frac{\Omega^2}{2r^2} - \frac{Mm^2}{r^3}.$$

6.3.1 Timelike geodesics

In this case $K = 1$ and the effective potential is

$$V(r) = -\frac{M}{r} + \frac{\Omega^2}{2r^2} - \frac{Mm^2}{r^3}.$$

The first term is the Newtonian gravitational potential and the second term is the angular momentum barrier. The third term does not appear in Newtonian theory – it is a correction due to GR. See figure 6.2 for sketches of this potential.

At large r, $V \sim -Mr^{-1}$, and at $r = 2M$ we have $V = -\frac{1}{2}$ (remember that we are only working in the region $r > 2M$ for now).

The extrema of the potential are at

$$V' = 0 \Rightarrow r = \frac{\Omega^2}{2M}\left(1 \pm \sqrt{1 - 12\frac{M^2}{\Omega^2}}\right),$$

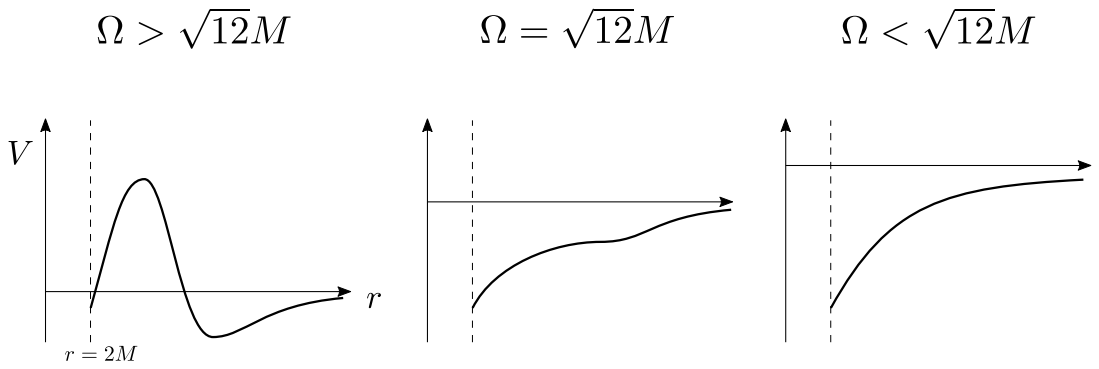

Figure 6.2: The effective potential in a Schwarzschild spacetime, for various values of the conserved angular momentum Ω.

so if $\Omega > \sqrt{12}M$ there are two local extrema, at $\Omega = \sqrt{12}M$ these two extrema collide (leaving a single inflection point), and at $\Omega < \sqrt{12}M$ there are no real extrema (see figure 6.2).

Timelike circular orbits

First, let's look for circular orbits. These have $\dot{r} = 0$ and $\ddot{r} = 0$ – the latter means that we must be at a local extrema of the effective potential.

Labelling these extrema by r_- and r_+, with $r_+ > r_-$, we find that the extrema at r_- is always *unstable* (i.e. it is a local maximum) while that at $r = r_+$ is *stable*. The *innermost* (marginally) *stable circular orbit* (ISCO[4]) is obtained when $\Omega = \sqrt{12}M$, when $r = 6M$. The energy of these orbits can be calculated: e.g. for a stable circular orbit,

$$\frac{E^2 - 1}{2} = V(r_+).$$

Bound orbits

If $\Omega > \sqrt{12}M$ then there are bound orbits which are not circular. These have energies satisfying

$$V(r_+) < \frac{E^2 - 1}{2} < 0,$$

(see figure 6.3).

Unbound orbits

If $\Omega > \sqrt{12}M$ then there are also unbound orbits, which have energies satisfying $E^2 \geq 1$ (see figure 6.4).

For smaller angular momentums ($\Omega \leq \sqrt{12}M$) the situation is interesting: in this regime there is no local maximum of the effective potential. As before, there are unbound orbits (with $E^2 \geq 1$) – but these orbits will only reach infinity if they are outgoing *initially*. All other orbits – that is, orbits with $E^2 < 1$ or with $\dot{r} < 0$ initially – will eventually reach the surface $r = 2M$ (figure 6.4).

[4]This is very important in astrophysics. As well as the exterior regions in generic spherically symmetric spacetimes, the Schwarzschild metric also describes (spoiler alert) a black hole. Astrophysical black holes often have *accretion disks*, consisting of matter orbiting close to the black hole, in almost circular orbits. Friction causes this matter to slowly lose energy, falling towards the black hole (and emitting light). Once it reaches the ISCO, it quickly falls into the black hole. So the 'inner edge' of the accretion disk that you might see is not at the edge of the black hole, but at $r = 6M$.

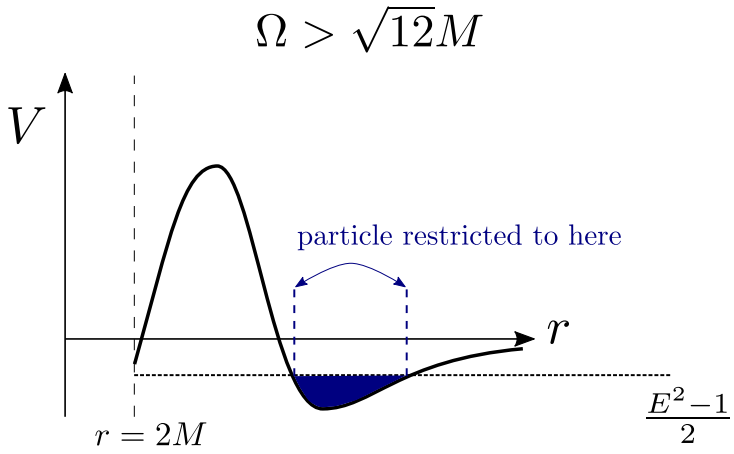

Figure 6.3: Bound orbits in Schwarzschild.

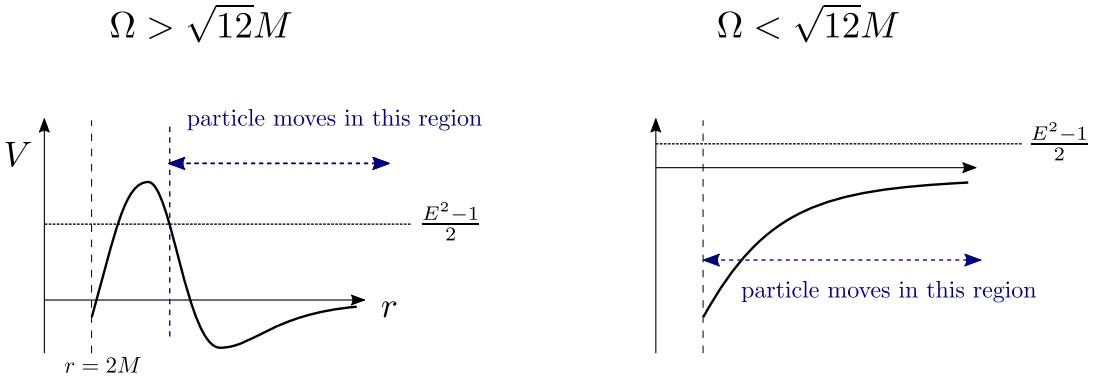

Figure 6.4: Unbound orbits in Schwarzschild.

6.4 Perihelion precession

One of the big scientific puzzles before the advent of general relativity was the *anomolous precession of the perihelion of Mercury*. The *perihelion* is the closest point of approach to the sun, and Newtonian theory predicts that planets move on ellipses, with the perihelion always occurring at the same point in space. But observations had shown that the perihelion of Mercury is *precessing* – on each orbit, the perihelion occurs at a slightly different angle (see figure 6.5).

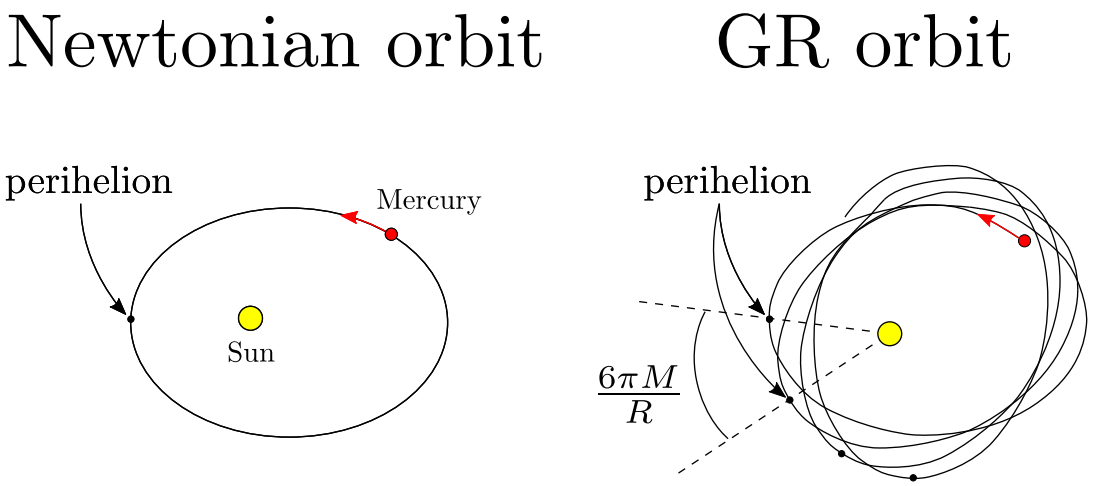

Figure 6.5

The orbits of the planets are close to circular, so our approach to this problem will be to treat Mercury as a point mass travelling on a circular orbit. We will then give this orbit small perturbations, and see what happens.

It is convenient to use the coordinate $u = \frac{M}{r}$ instead of r, and to parametrise the orbits by ϕ instead of the proper time τ. Then we have

$$\frac{\mathrm{d}u}{\mathrm{d}\phi} = \frac{\mathrm{d}u}{\mathrm{d}\tau}\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^{-1} = -\frac{M}{\Omega}\dot{r},$$

and so equation (6.2) becomes

$$\frac{1}{2}\left(\frac{\mathrm{d}u}{\mathrm{d}\phi}\right)^2 - \frac{M^2}{\Omega^2}u + \frac{1}{2}u^2 - u^3 = \frac{M^2(E^2-1)}{2\Omega^2}.$$

Differentiating with respect to ϕ

$$\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2} - \frac{M^2}{\Omega^2} + u - 3u^2 = 0.$$

Setting $u = \frac{M}{R} + \epsilon v(\phi)$, where R is the radius of a circular orbit (so $R^2 - \frac{\Omega^2}{M}R - 3\Omega^2 = 0$) we find

$$\begin{aligned} 0 &= \epsilon\frac{\mathrm{d}^2v}{\mathrm{d}\phi^2} - \frac{M^2}{\Omega^2} + \frac{M}{R} + \epsilon v - 3\left(\frac{M}{R} + \epsilon v\right)^2 \\ \Rightarrow 0 &= \frac{\mathrm{d}^2v}{\mathrm{d}\phi^2} + \left(1 - 6\frac{M}{R}\right)v - 3\epsilon v^2. \end{aligned}$$

Ignoring lower order terms, we have the equation

$$\frac{\mathrm{d}^2v}{\mathrm{d}\phi^2} + \left(1 - 6\frac{M}{R}\right)v = 0.$$

For stable circular orbits $R > 6M$. Then this equation has periodic solutions, with period

$$T_{\text{period}} = \frac{2\pi}{\left(1 - 6\frac{M}{R}\right)^{\frac{1}{2}}} \sim 2\pi + \frac{6\pi M}{R},$$

so the perihelion *precesses* by an additional $6\pi M/R$ per orbit. This matches the observed anomalous precession of Mercury!

6.5 Gravitational bending of light

One of the other classic tests of general relativity is the bending of light when it passes near a massive object.

For this purpose, we need to use the massless geodesic equation rather than the massive one, i.e. we need to take $K = 0$. The equation for r is then

$$\frac{1}{2}\dot{r}^2 + \frac{\Omega^2}{2r^2}\left(1 - \frac{2M}{r}\right) = \frac{E^2}{2}.$$

The effective potential is sketched in figure 6.6. Note that here, 'dots' mean derivatives with respect to an affine parameter along the null geodesic (not the proper time!).

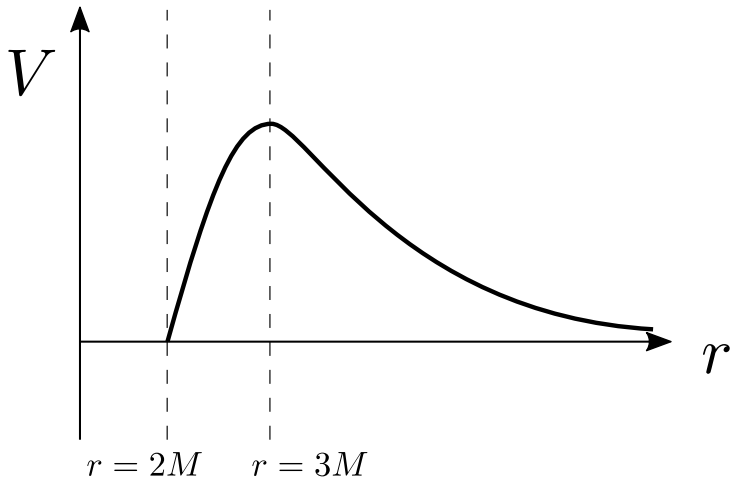

Figure 6.6: The effective potential for null geodesics moving in the Schwarzschild spacetime.

Note that, for this effective potential, there is only one local extremum, at $r = 3M$. This is called the *light ring* or *photon sphere* – on this surface, light can orbit the central object (assuming that this central object is smaller than $3M$)! However, unlike the massive case, such orbits are *always* unstable.

For the bending of light, we are interested in unbounded orbits, so the energy satisfies

$$0 < \frac{E^2}{2} < V(3M) = \frac{\Omega^2}{27M^2}.$$

As before, it is useful to set $u = \frac{M}{r}$ and use ϕ as a parameter along the curve instead of the affine parameter. Then we obtain the equation

$$\frac{\mathrm{d}^2 u}{\mathrm{d}\phi^2} + u - 3u^2 = 0.$$

In the Newtonian theory the third term is absent. So, in the Newtonian theory, the solutions are

$$u = \frac{M}{b}\sin(\phi - \phi_0),$$

where b and ϕ_0 are constants. This equation can be rewritten

$$r\sin(\phi - \phi_0) = b,$$

see figure 6.7. There is no gravitational deflection in the Newtonian theory! The constant b is called the *impact parameter*: in the Newtonian theory, it measures the closest distance of the curve to the origin.

Now, let's reintroduce the quadratic term, which is the correction from GR. We'll consider *large* impact parameters (so the light ray passes far from the central region), so we'll set $b = B\epsilon^{-1}$. We'll also set $\phi_0 = 0$ for simplicity. We write the solution as

$$u = \epsilon\frac{M}{B}\sin\phi + \epsilon^2 v(\phi),$$

and we expand in powers of ϵ. The equation for u gives

$$\epsilon^2\left(\frac{\mathrm{d}^2 v}{\mathrm{d}\phi^2} + v - 3\frac{M^2}{B^2}\sin^2\phi\right) + \mathcal{O}(\epsilon^3) = 0.$$

Ignoring lower order terms, the general solution to this equation is

$$v = \alpha\sin\phi + \beta\cos\phi + \frac{M^2}{B^2}(1+\cos\phi)^2,$$

for some constants α and β. For a particle coming in 'from the left' (see figure 6.7), both the perturbation and its derivative should vanish for $\phi = \pi$. These conditions give $\alpha = \beta = 0$.

Putting these calculations together, the solution (up to $\mathcal{O}(\epsilon^2)$ is

$$u = \epsilon\frac{M}{B}\sin\phi + \epsilon^2\frac{M^2}{B^2}(1+\cos\phi)^2 + \mathcal{O}(\epsilon^3).$$

Recall that $r = \frac{M}{u}$. To find the deflection angle, we need to find the value of ϕ such that $u = 0$, with $\phi \leq 0$.

Setting $\phi = -\epsilon(\Delta\phi)$ and expanding in powers of ϵ, we find that

$$\begin{aligned} 0 &= -\epsilon^2\frac{M}{B}(\Delta\phi) + 4\epsilon^2\frac{M^2}{B^2} + \mathcal{O}(\epsilon^3) \\ \Rightarrow (\Delta\phi) &= 4\frac{M}{B} + \mathcal{O}(\epsilon). \end{aligned}$$

So light rays *are* deflected when they pass near massive objects in general relativity!

This value matches the observations very well. In 1919, after Einstein published GR, two expeditions were sent out (from the Royal Astronomical Society and the Royal Society) to measure the deflection of light coming from stars behind the sun during a solar eclipse. They found that the apparent position of the stars changed when they were behind the sun – in perfect agreement with the prediction we have just made. This success led to a 'ticker-tape parade' for Einstein through New York City – the only such parade that has ever taken place for a scientist!

6.6 Black holes and singularities

6.6.1 Coordinates

So far we have resolutely stuck to the region $r > 2M$. This is fine so long as we are looking at spherically symmetric stars or planets, which will have some matter that modifies the geometry in some way (which we don't have to care about) for small values of r. But what if there is no matter there? The Schwarzschild

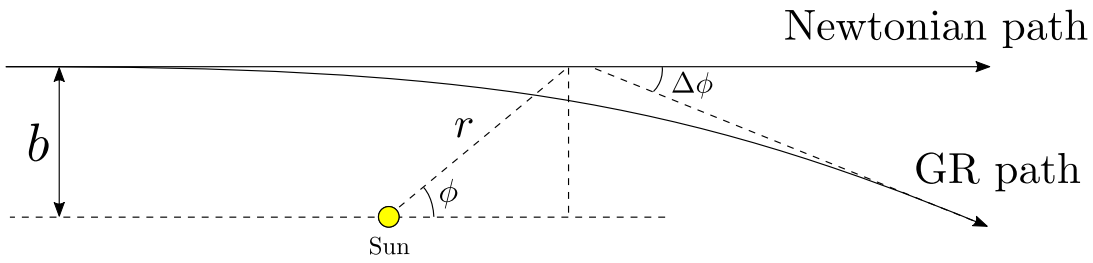

Figure 6.7: Gravitational deflection of light.

solution is still a solution to the *vacuum Einstein equations*, so sooner or later we have to understand what's going on at $r = 2M$.

Recall that the Schwarzschild metric is

$$g = -\left(1 - \frac{2M}{r}\right) \mathrm{d}t^2 + \left(1 - \frac{2M}{r}\right)^{-1} \mathrm{d}r^2 + r^2 \left(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2\right).$$

There are several places where something goes funny with this expression for the metric. At $r = 2M$ the first component vanishes and the second term becomes infinite. At $r = 0$ all of the components vanish except for the first one, which becomes infinite. Finally[5], at $\theta = 0, \pi$ the $\mathrm{d}\phi^2$ component vanishes.

This last point should give us pause for thought. The metric degenerates on the axis $\theta = 0, \pi$, but this doesn't mean that there is some kind of physical singularity there – in fact, exactly the same thing happens in flat space when it's written in spherical polar coordinates! What's happening at the poles is not that there is something wrong with the *metric*, but that there is something wrong with the *coordinates*. If we change to different coordinates – for example, changing to polar coordinates with a different pole, or to rectangular coordinates – then these points appear totally normal.

Could something similar be happening at the other places where the metric is problematic, at $r = 2M$ and at $r = 0$? Well, maybe. We can obtain hints of what might be going on my looking at *scalar* quantities, since these are invariant under coordinate transformations.

If we want to look at scalar quantities constructed out of the metric, then we cannot take contractions of the metric itself: $(g^{-1})^{\mu\nu} g_{\mu\nu} = 4$, which doesn't tell us anything. As far as first derivatives of the metric go, these are encoded in the Christoffel symbols – but these are not tensors. Our only real choice is to look at the curvature. The scalar curvature R will not work, since the Schwarzschild metric is a solution of the Einstein *vacuum* equations: contracting these equations using the metric

$$\begin{aligned} R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} &= 0 \\ \Rightarrow R - 2R = 0 \Rightarrow R &= 0. \end{aligned}$$

For this reason, the Einstein vacuum equations are equivalent to $R_{\mu\nu} = 0$.

There is a scalar that can be built out of the curvature tensor that can be useful for our purposes, called the *Kretschmann scalar* $R_{\mu\nu\rho\sigma} R^{\mu\nu\rho\sigma}$. In the Schwarzschild metric, this has the value

$$R_{\mu\nu\rho\sigma} R^{\mu\nu\rho\sigma} = \frac{12M^2}{r^6}.$$

So in a coordinate-independent sense, the curvature is finite at $r = 2M$ but infinite at $r = 0$. This suggests that $r = 2M$ might be an ordinary surface in spacetime – not a singularity – but $r = 0$ might represent a real singularity.

We should not be too hasty and immediately conclude that $r = 2M$ is a 'coordinate singularity' while $r = 0$ is a physical singularity, simply on the basis of the Kretschmann scalar (even though this

[5] Technically, there is also something unusual going on at $\phi = 0, 2\pi$. Since we are supposed to cover a manifold by open sets which are then mapped onto $\mathbb{R}^n$, these points are not covered by our chart.

will actually turn out to be the case!). There are some interesting subtleties here, that we won't be able to fully explore, but I will give some indication of the issues.

First, suppose that the curvature blows up in some coordinate-independent way. Does this necessarily mean that spacetime comes to an end? Remember that the Einstein equations are a second-order system of PDEs, so we might think that, to make sense of the Einstein equations, the curvature must be finite. However, if you've ever studies PDEs you might have come across the notion of a *weak solution*, which is a 'solution' of a PDE with less regularity than might be expected - a famous example is a shock wave in a fluid. It turns out that the Einstein equations can be made sense of in situations where the curvature is infinite (so-called *impulsive gravitational waves*), so this doesn't necessarily signal a singularity.

On the other hand, there are situations where the curvature is finite, and indeed nothing at all unusual happens *locally*, but for *global* reasons the solution to the Einstein equations cannot be continued past some surface in spacetime. The most famous such example is called a *Cauchy horizon* – these occur inside rotating black holes, and are connected with an important unproved conjecture in GR called *strong cosmic censorship*. Unfortunately, these issues also lie beyond the scope of this course.

Returning to the Schwarzschild metric, we have seen that the Kretschmann scalar *suggests* that the surface $r = 2M$ might be simply a coordinate singularity. To show that this is in fact the case, we need to transform to some different coordinates which 'pass through' the surface $r = 2M$.

First, let's examine the structure of the light cones near $r = 2M$. Since we're in spherical symmetry, we'll look only at the (t, r) plane, and we'll look at radially ingoing and outgoing null curves.

A null vector X in the (t, r) plane satisfies

$$\begin{aligned} &-\left(1-\frac{2M}{r}\right)(X^t)^2+\left(1-\frac{2M}{r}\right)^{-1}(X^r)^2=0\\ &\Rightarrow X^r=\pm\left(1-\frac{2M}{r}\right)X^t, \end{aligned}$$

and a null curve passing through the point (t_0, r_0) is given by

$$\begin{aligned} \frac{\mathrm{d}r}{\mathrm{d}t}&=\pm\left(1-\frac{2M}{r}\right)\\ \Rightarrow t-t_0&=\pm\left(r-r_0+2M\log\left(\frac{r-2M}{r_0-2M}\right)\right), \end{aligned}$$

so as $r \to 2M$, the null cones are "squeezed together" (see figure 6.8). It looks as though an ingoing null curve will never reach the surface $r = 2M$, but is this really true? Or is it an artefact of the coordinates?

This suggests that we should try to 'straightening out' the null cones. This can be achieved by changing from the coordinate t to a coordinate v which is *constant* on the ingoing null curves. With this in mind, we define

$$\begin{aligned} r^*&:=r+2M\log\left(\frac{r-2M}{2M}\right)\\ v&:=t+r^* \end{aligned}$$

It is easy to check that v is constant along ingoing null geodesics (**exercise**). Then we have

$$\begin{aligned} \mathrm{d}v&=\mathrm{d}t+\mathrm{d}r^*\\ &=\mathrm{d}t+\left(1-\frac{2M}{r}\right)^{-1}\mathrm{d}r. \end{aligned}$$

Now, if we write the metric in coordinates (v, r, θ, ϕ) it takes the form

$$g=-\left(1-\frac{2M}{r}\right)\mathrm{d}v^2+2\mathrm{d}v\mathrm{d}r+r^2\left(\mathrm{d}\theta^2+\sin^2\theta\mathrm{d}\phi^2\right).$$

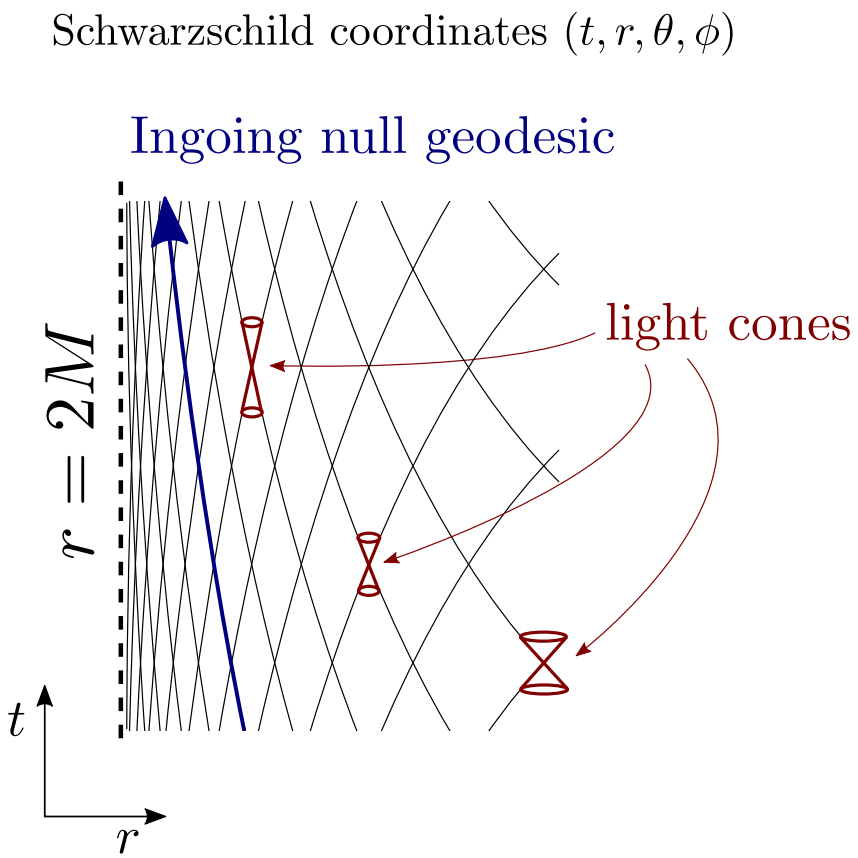

Figure 6.8: Light cones and radial light rays in the Schwarzschild spacetime, drawn in "Schwarzschild coordinates" (t, r, θ, ϕ).

The metric is no longer diagonal due to the term $2\mathrm{d}v\mathrm{d}r$. The first term still vanishes at $r = 2M$ so we might be tempted to say that the metric degenerates here. However, the matrix g_{ab} is not degenerate: it is still an invertible matrix, so it has maximal rank. To check this, we calculate

$$\begin{pmatrix} \left(1-\frac{2M}{r}\right) & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2\sin^2\theta \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & \left(1-\frac{2M}{r}\right) & 0 & 0 \\ 0 & 0 & r^{-2} & 0 \\ 0 & 0 & 0 & r^{-2}(\sin\theta)^{-2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

so the inverse metric is

$$g^{-1} = (g^{-1})^{ab}\partial_a\partial_b = 2\partial_v\partial_r + \left(1-\frac{2M}{r}\right)(\partial_r)^2 + r^{-2}(\partial_\theta)^2 + r^{-2}(\sin\theta)^{-2}(\partial_\phi)^2.$$

So the components of g and g^{-1} are both finite at $r = 2M$ (except for the usual degeneracy at $\theta = 0, \pi$ from polar coordinates)!

The coordinates (v, r, θ, ϕ) are called *ingoing Eddington-Finkelstein coordinates* (There are also *outgoing Eddington-Finkelstein coordinates* (u, r, θ, ϕ), where $u = t - r^*$). In these coordinates, the ingoing null curves are simply given by $v =$ constant, while the outgoing radial null curves (in the region $r > 2M$) are given by

$$\begin{aligned} v - v_0 &= 2\left(r^* - r_0^*\right) \\ &= 2\left(r - r_0 + 2M\log\left(\frac{r-2M}{r_0-2M}\right)\right). \end{aligned}$$

Note that, as $r \to \infty$, $v \sim t + r$. Also, in these coordinates, there is nothing stopping us from considering the 'interior' region $0 < r < 2M$ - the metric is perfectly regular at $r = 2M$. You can also check that the curve given by $r = 2M$ is itself a null curve (**exercise**). In fact, the surface $r = 2M$ acts as a one-way membrane: you can pass through it from the exterior $r > 2M$ to the interior $r < 2M$, but there are no causal curves going in the other direction! See figure 6.9.

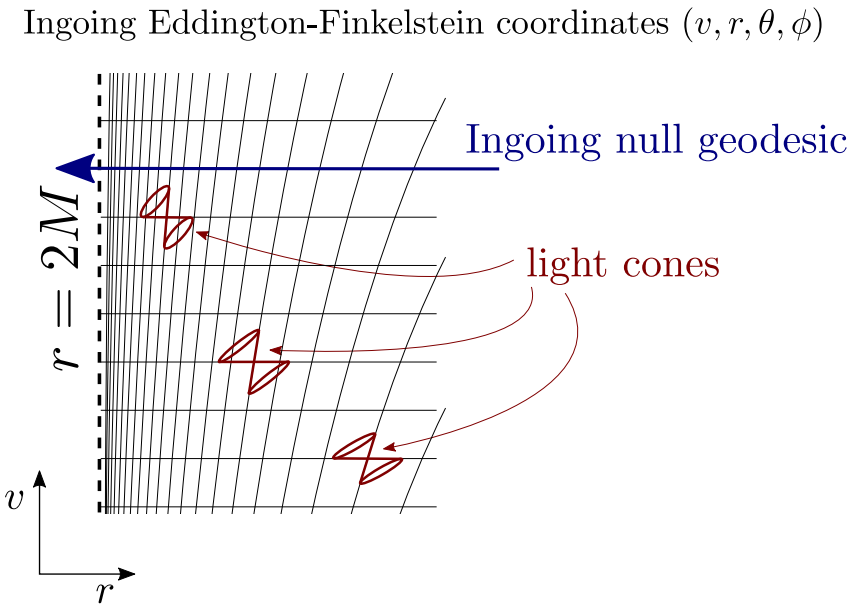

Figure 6.9: Light cones and radial light rays in the Schwarzschild spacetime, drawn in ingoing Eddington-Finkelstein coordinates (v, r, θ, ϕ).

We can also consider *outgoing* Eddington-Finkelstein coordinates (u, r, θ, ϕ). In these coordinates, the surface $r = 2M$ can be also be seen to be a null surface, but this time causal curve cross it in the opposite direction: you can pass from the interior $r < 2M$ to the exterior $r > 2M$, but you can't

enter the interior! The reason for this apparent discrepancy is that the ingoing and outgoing Eddington-Finkelstein coordinates cover different parts of the manifold (see figure 6.10). Outgoing Eddington-Finkelstein coordinates cover a *different* 'interior' region $0 < r < 2M$.

We can also find some coordinates which cover the original region ($r > 2M$) as well as the two extra regions covered by ingoing and outgoing Eddington-Finkelstein coordinates. These are called *Kruskal coordinates* (or *Kruskal-Szekeres coordinates*), and are defined by

$$U = -e^{-\frac{u}{4M}}$$
$$V = e^{\frac{v}{4M}}$$

then the metric takes the form

$$g = \frac{32M^3}{r} e^{-\frac{r}{2M}} \mathrm{d}U\mathrm{d}V + r^2 \left(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2\right)$$

where here r is defined implicitly in terms of U and V by the relationship

$$\log(-UV) = \frac{r^*}{2M}$$

This metric is regular at $r = 2M$, which is given by $U = 0$ or $V = 0$. The original region that we started in is the region $U < 0$, $V > 0$. We can also clearly extend to positive values of U and/or negative values of V. The region $V > 0$, $U > 0$ is the same as the interior region covered by the ingoing Eddington-Finkelstein coordinates, while the region $U < 0$, $V < 0$ is the same as the interior region covered by outgoing Eddington-Finkelstein coordinates. The region $U > 0$, $V < 0$ is new: it turns out that this region is isometric to the original exterior region $r > 2M$!

The full spacetime, including all four regions, is sometimes called the *maximally extended Schwarzschild spacetime*.

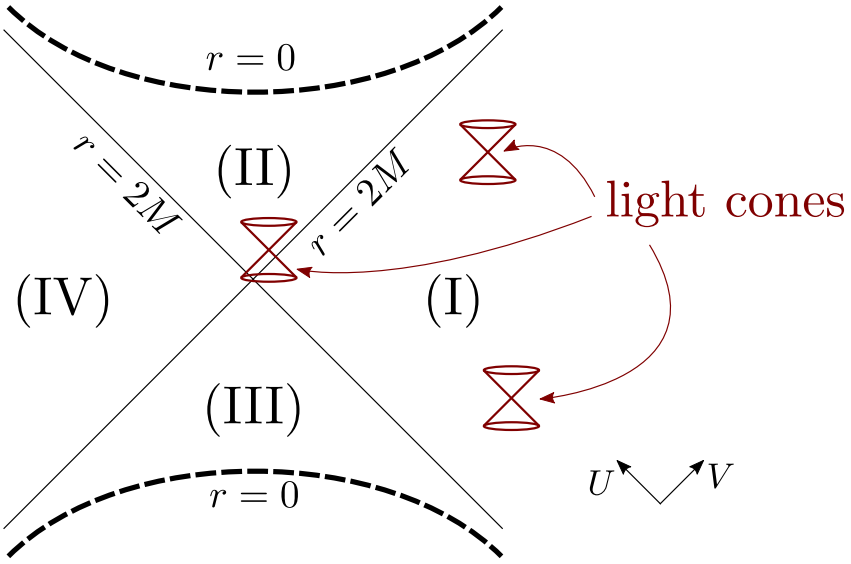

Figure 6.10: The maximally extended Schwarzschild spacetime. The original Schwarzschild coordinates only cover region (I). Ingoing Eddington-Finkelstein coordinates cover the regions (I) and (II), while outgoing Eddington-Finkelstein coordinates cover regions (I) and (III). Finally, KruskalSzekeres coordinates cover regions (I), (II), (III) and (IV).
Regions (I) and (IV) are asymptotically flat regions: they include regions of aribtrarily large r, where the metric looks like the Minkowski metric. Region (II) is the black hole region: causal worldlines can enter this region but never leave it. Region (III) is the white hole region: causal worldlines can leave this region but never enter it.

Often, rather than U and V, Kruskal coordinates are defined to be the coordinates (T, X, θ, ϕ), where

$$T = \frac{1}{2}(U + V)$$
$$X = \frac{1}{2}(V - U)$$

Exercise: write out the Schwarzschild solution in these coordinates. You may make use of the function r, which can be implicitly defined in terms of T and X.

Note the following useful fact: in each of the four regions (I), (II), (III) and (IV), the metric can be put back into coordinates that look like the original Schwarzschild coordinates. However, note that these are all *different* coordinate systems: the original coordinates only cover region (I), and none of these coordinates cover the surface $r = 2M$ (given in Kruskal coordinates by $UV = 0$). In region (II), for example, we can define a coordinate system $\left(t_{(\mathrm{II})}, r_{(\mathrm{II})}, \theta_{(\mathrm{II})}, \phi_{(\mathrm{II})}\right)$, which can be obtained from the ingoing Eddington-Finkelstein coordinates (v, r, θ, ϕ) by setting

$$\begin{aligned} r^*_{(\mathrm{II})} &:= r + 2M \log\left(\frac{2M - r}{2M}\right) \\ t_{(\mathrm{II})} &:= v - r^*_{(\mathrm{II})} \\ r_{(\mathrm{II})} &:= r \\ \theta_{(\mathrm{II})} &:= \theta \\ \phi_{(\mathrm{II})} &:= \phi \end{aligned}$$

then the metric takes the form

$$g = \left(1 - \frac{2M}{r_{(\mathrm{II})}}\right)^{-1} \mathrm{d}r^2_{(\mathrm{II})} - \left(1 - \frac{2M}{r}\right) \mathrm{d}t^2_{(\mathrm{II})} + r^2_{(\mathrm{II})} \left(\mathrm{d}\theta^2_{(\mathrm{II})} + \sin^2\theta_{(\mathrm{II})} \mathrm{d}\phi^2_{(\mathrm{II})}\right)$$

Note that, since $r < 2M$ in this region, the coefficient of $\mathrm{d}r^2_{(\mathrm{II})}$ is negative, while the coefficient of $\mathrm{d}t^2_{(\mathrm{II})}$ is positive! So, in this region, $r_{(\mathrm{II})}$ is a 'time coordinate' and $t_{(\mathrm{II})}$ is a 'space coordinate'. Actually, $r_{(\mathrm{II})}$ *decreases* towards the future in this region.

6.6.2 Interpreting the maximally extended Schwarzschild space time

Throughout the entire extended Schwarzschild spacetime, the vacuum Einstein equations $R_{\mu\nu} = 0$ hold. Hence this is a *vacuum solution to the Einstein equations*. You will sometimes hear people say things like 'there is an infinitely dense point of matter at $r = 0$ in a black hole' but this is not true. The 'point' $r = 0$ is not actually a part of the Schwarzschild manifold at all (it only extends to the open region $r > 0$) and there is no matter anywhere in sight. What's more, it has recently been shown that black holes can (theoretically – not in realistic astrophysical situations) be formed from the collision of gravitational waves, without any matter taking part!

It is useful to keep figure 6.10 in mind, and to remember that, on this diagram (radial) light rays travel at 45 degrees. Region (I) is the most familiar region: here $r > 2M$ and, at large distances, the metric approaches the flat Minkowski metric. Worldlines of observers in this region, which always have tangent vectors *inside* the light cones, can always escape to regions of arbitrarily large r.

Region (II) is called the *black hole region*. Once an observer has crossed the surface $r = 2M$ into region (II) they are stuck inside this region: no worldlines with tangent vectors inside the future light cones can leave this region. The surface $r = 2M$, called the *event horizon*, acts like a one-way membrane: once you cross $r = 2M$ there is no turning back! On the other hand, there is no local quantity which distinguishes this surface: the curvature is finite, and small observers can cross this surface without noticing anything dramatic.

Once inside the black hole region r decreases along all worldlines, and in fact *all observers will reach* $r = 0$ *in a finite affine time* (see the example sheet). Indeed, $r = 0$ is more like a time than a point in

space. However, unlike the surface $r = 2M$, $r = 0$ is a genuine singularity: the curvature is infinite here, and there is no way to extend the manifold[6] beyond $r = 0$. Point particles moving on geodesics will not experience any forces (after all, that is what is the defining feature of a geodesic) but for a realistic observer of any finite size, tidal forces become infinite as $r \to 0$, ripping the observer apart.

Because of the singularity at $r = 0$, the Schwarzschild spacetime is said to be *geodesically incomplete*. A *geodesically complete manifold* is one for which all geodesics can be extended arbitrarily far, so that their associated affine parameters can take values in $(-\infty, \infty)$. Minkowski space is geodesically complete, while the Schwarzschild spacetime is not.

Finally, we come to regions (III) and (IV). Both of these regions are considered unphysical: in a realistic black hole formed by the collapse of a star they are 'covered up' by the matter (see figure 6.11). However, for the purposes of calculations it is often helpful to work with the full extended Schwarzschild geometry. Since this geometry is invariant under time reversal, these two regions must be present. Region (III) is called the *white hole* region: it is the time-reversal of the black hole region, and has the time-reversed properties. Observers can leave this region, but they can never enter it!

Region (IV) is a 'copy' of the original region (I): it is also asymptotically flat, and looks like Minkowski space far away from the event horizon. It is sometimes said to be 'another universe'. There is no way for observers to get from region (I) into region (IV), so there is no 'wormhole' here!

[6] In fact, the manifold cannot be extended in a way such that the metric is even *continuous*, let alone differentiable. This was only proved in the last few years!

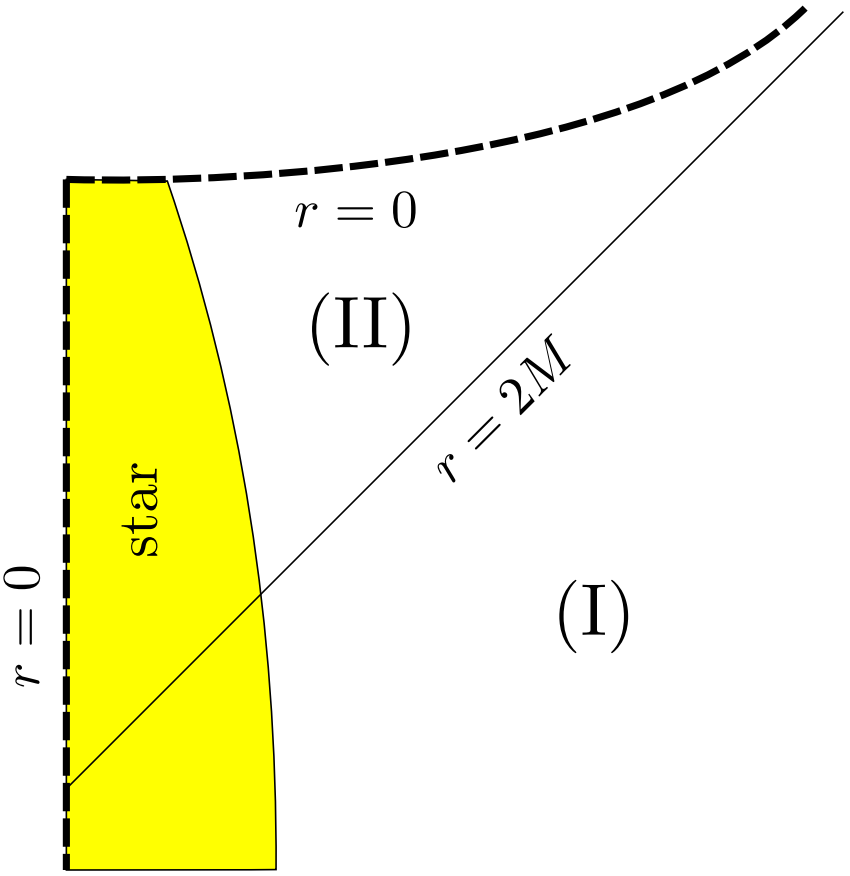

Figure 6.11: The spacetime of a "realistic" spherically symmetric collapsing star. By Birkhoff's theorem, the geometry is exactly described by the Schwarzschild metric outside of the star, but inside the star the geometry will be modified. In this case it is modified so that there is a regular 'centre' at $r = 0$ running up the left hand side of the page: unlike the singularity at $r = 0$ at the top of the page, this line does not represent anything unusual: this line is like $r = 0$ in Minkowski space. Note that the star effectively 'covers up' regions (III) and (IV), but there are still parts of regions (I) and (II) left intact. In particular, there is still a black hole region, from which no causal worldlines can escape – outside of the star, this region is bounded by the surface $r = 2M$.

Chapter 7

Cosmology

> The Cosmos is all that is or ever was or ever will be. Our feeblest contemplations of the Cosmos stir us – there is a tingling in the spine, a catch in the voice, a faint sensation of a distant memory, as if we were falling from a great height. We know we are approaching the greatest of mysteries.
>
> Carl Sagan, *Cosmos*.

So far we have only seen one solution to the Einstein equations: the Schwarzschild solution, which is a solution to the vacuum Einstein equations with zero cosmological constant. Remember that this solution was found by searching for solutions with lots of symetries: in this case, we looked for spherically symmetric, static spacetimes[1]. This symmetry class is important for astrophysical purposes, in which many important systems are approximately spherically symmetric, and it also led to the very interesting phenomena of black holes.

Another important symmetry class is *homogeneous and isotropic* spacetimes. Instead of astrophysical applications, this symmetry class is suitable for studying the entire universe on the largest scales.

Homogeneity

A *homogeneous spacetime* is one where there is a global function τ, called a *time function*, with level sets Σ_τ satisfying:

- The surfaces Σ_τ are spacelike hypersurfaces, i.e. $\mathrm{d}\tau$ is a timelike covector, $g^{-1}(\mathrm{d}\tau, \mathrm{d}\tau) < 0$ (equivalently every curve which lies entirely within Σ_τ is spacelike). By rescaling τ if necessary, we can assume that $g^{-1}(\mathrm{d}\tau, \mathrm{d}\tau) = -1$.
- The surfaces Σ_τ are *homogeneous spaces*. This means that there is a group G which acts on the surface Σ_τ *transitively* (any point can be mapped to any other point by some group element) and by *isometries* (the action of the group preserves the metric). Informally, this says that 'every point looks like every other point' (sometimes called the *Copernican principle*).

(see figure 7.1).

Isotropy

The level sets Σ_τ are also required to be *isotropic*. This means that

- for each point $p \in \Sigma_\tau$ and for each pair of unit tangent vectors $X, Y \in T_p(\Sigma_\tau)$ (that is, tangent vectors to the submanifold Σ_τ, not spacetime vectors – although such vectors can be considered

[1] From Birkhoff's theorem, we could have looked for solutions to the Einstein equations in spherical symmetry: the only such solutions turn out to also be static.

spacetime vectors that are tangent to the surface Σ_τ), there is an isometry mapping X to Y (see figure 7.1).

By a unit vector, we simply mean a vector X such that $g(X, X) = 0$. By a vector tangent to the surfaces Σ_τ, we mean that $X(\tau) = 0$.

In fact, isotropy implies homogeneity but not vice versa. For example, a torus is homogeneous but not isotropic.

These definitions also mean that there are special observers, called *isotropic observers* or *comoving observers*, whose worldlines are such that

- The worldlines are timelike.
- For every pair of unit vectors X, Y which are *orthogonal* to the tangent vector of the wordline, there is an isometry mapping X to Y.

Note that the tangent vector to these worldlines is simply $-(\mathrm{d}\tau)^\sharp$, or, in abstract index notation, $-\partial^\mu \tau$. The $-$ sign is chosen so that these tangents are future-directed.

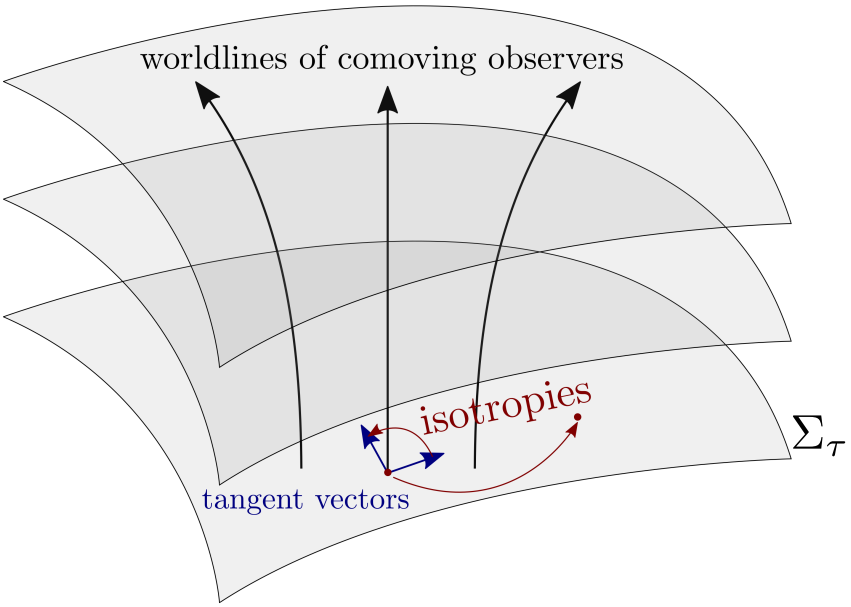

Figure 7.1: A homogeneous and isotropic spacetime. There are isometries mapping any point on a surface Σ_τ to any other point on that surface, and also isometries mapping any tangent vector to the surface Σ_τ to any other tangent vector (these isometries are shown in red). Comoving observers move along worldlines which are orthogonal to ths surfaces Σ_τ.

The worldline of the Earth is, roughly, the worldline of an isotropic observer: on large scales, the universe looks more or less the same in every direction. This would not be the case for an astronaut moving past the earth with some large relative velocity (even if they are also moving on a timelike geodesic): to them, the universe would look "blueshifted" in front of the spacecraft and "redshifted" behind it, due to the Doppler effect[2].

[2]In some ways we are almost back to the 'Atomist' view of spacetime. But there is a crucial difference: now, the existence of the special class of observers is not built into the basic structure of spacetime - instead it is due to symmetries which exist in the particular solution to equations which we happen to be living in. This is reminiscent of the way in which altitude is dealt with in the Aristotelian vs. Atomist spacetimes.

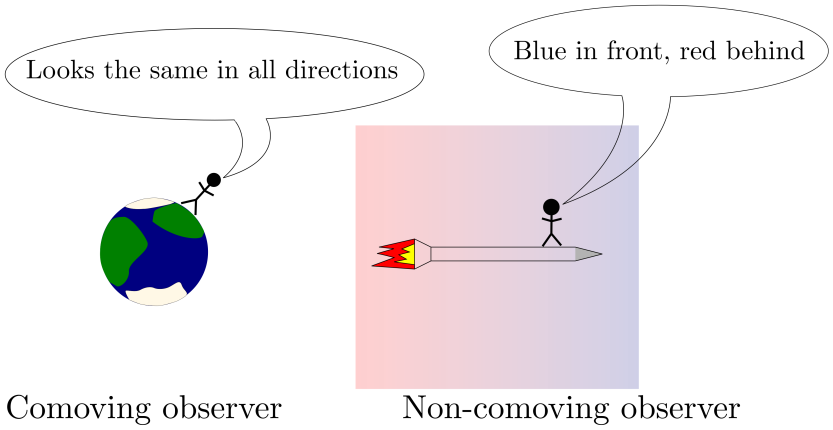

Figure 7.2: Comoving and non-comoving observers in a homogeneous and isotropic spacetime.

7.1 Geometry and matter in a homogeneous and isotropic universe

Homogeneity and isotropy imply that the spacetime metric can be written as

$$g = -\mathrm{d}\tau^2 + (a(\tau))^2\, \underline{g}.$$

Here $\underline{g}$ is the spatial part of the metric, and $a(\tau)$ is called the *scale factor*. Furthermore, the metric $\underline{g}$ is required to be the metric of a *maximally symmetric space*. There are actually only three options:

- *Flat*: the metric $\underline{g}$ is simply the Euclidean metric in three dimensions.
- *Closed*: the metric $\underline{g}$ is the standard metric on the 3-sphere $\mathbb{S}^3$.
- *Open*: the metric $\underline{g}$ is a metric of constant *negative* curvature.

In each case there are possible topological modifications – e.g. in the flat space, Σ_τ need not be homeomorphic to $\mathbb{R}^3$ but it could instead be a cylinder $\mathbb{R}^2 \times \mathbb{S}^1$ or a torus $\mathbb{S} \times \mathbb{S} \times \mathbb{S}$.

In all cases, the metric can be written in a universal form, called the *Robertson-Walker metric*:

$$\underline{g} = \frac{\mathrm{d}r^2}{1 - kr^2} + r^2 \left(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2\right),$$

where $k > 0$ is a closed universe, $k = 0$ is a flat universe and $k < 0$ is an open universe. By rescaling r and the scale factor a, we can always choose $k = 1, 0$ or -1 in the closed, flat or open cases respectively (**exercise**).

Christoffel symbols and curvature components

This symmetry class makes it relatively easy to calculate the Christoffel symbols and the components of the Riemann curvature tensor, for the following reasons:

- Scalar fields on each surface Σ_τ must be constant for consistency with homogeneity. So scalar quantities can only depend on time τ.
- All (co)vector fields on the surfaces Σ_τ must vanish for consistency with isotropy.
- The only $(1,1)$ tensor fields on the surface Σ_τ consistent with homogeneity and isotropy are those proportional to the Kronecker delta δ^i_j, and the constant of proportionality can only depend on time τ.

- The only $(0,2)$ tensor fields on the surface Σ_τ consistent with homogeneity and isotropy are those proportional to the metric $\underline{g}_{ij}$, and the constant of proportionality can only depend on time τ.
- The only $(2,0)$ tensor fields on the surface Σ_τ consistent with homogeneity and isotropy are those proportional to the inverse metric $(\underline{g}^{-1})^{ij}$, and the constant of proportionality can only depend on time τ.

Recall that we use i, j, k etc. to refer to *spatial* indices. Using these facts together with the form of the metric, some fairly tedious calculations lead to the following expressions for the Ricci tensor components:

$$\begin{aligned} R_{00} &= -3\frac{\ddot{a}}{a} \\ R_{0i} &= 0 \\ R_{ij} &= \left(a\ddot{a} + 2\dot{a}^2 + 2k\right)\underline{g}_{ij}, \end{aligned}$$

where 'dots' represent derivatives with respect to τ. From these we can calculate

$$\begin{aligned} G_{00} + \Lambda g_{00} &= 3\frac{\dot{a}^2 + k}{a^2} - \Lambda \\ G_{ij} + \Lambda g_{ij} &= \left(-2a\ddot{a} - \dot{a}^2 - k + a^2\Lambda\right)\underline{g}_{ij}. \end{aligned}$$

The energy momentum tensor must also respect the symmetries imposed by homogeneity and isotropy. This means that we can write

$$\begin{aligned} T_{00} &:= \rho \\ T_{ij} &:= pa^2\underline{g}_{ij}, \end{aligned}$$

where these equations *define* the 'density' and 'pressure' – we are not necessarily assuming that the matter is a fluid.

7.2 The Friedmann equations

The Einstein equations in a homogeneous and isotropic spacetime are called the *Friedmann equations*. They are the following system of ODEs, which are derived from the expressions above for the components of the Einstein tensor and the energy momentum tensor:

$$3\frac{\dot{a}^2 + k}{a^2} - \Lambda = 8\pi\rho \tag{7.1}$$

$$2a\ddot{a} + \dot{a}^2 + k - a^2\Lambda = -8\pi p a^2. \tag{7.2}$$

Sometimes the following equation, following from the two above, is useful:

$$\frac{\ddot{a}}{a} = -\frac{4}{3}\pi(\rho + 3p) + \frac{1}{3}\Lambda.$$

If we supplement these equations with an equation of state, expressing the pressure p as a function of the density ρ, then this forms a closed system of ODEs, which can be solved as follows: first we write p in terms of ρ in equation (7.2), and then we substitute for ρ in terms of a and $\dot{a}$ using equation (7.1). This will lead to a nonlinear second order ODE for the scale factor a.

An evolution equation for the density ρ

By differentiating equation (7.1) with respect to τ, we find

$$\begin{aligned} 8\pi\dot{\rho} &= \frac{3\dot{a}}{a^3}\left(2a\ddot{a} - 2\dot{a}^2 - 2k\right) \\ &= \frac{3\dot{a}}{a^3}\left(3\dot{a}^2 - 3k + a^2\Lambda - 8\pi a^2 p\right) \\ &= -24\pi\frac{\dot{a}}{a}(p+\rho), \end{aligned}$$

and so we obtain the equation for the derivative of the density

$$\dot{\rho} = -3\frac{\dot{a}}{a}(p+\rho). \tag{7.3}$$

Equations of state

Often equations of state of the following form are considered:

$$p = w\rho,$$

where w is a constant. In this case the density evolves as

$$\begin{aligned} \frac{\dot{\rho}}{\rho} &= -3(1+w)\frac{\dot{a}}{a} \\ \Rightarrow \rho &\propto a^{-3(1+w)}. \end{aligned}$$

There are certain particular values of the constant w which have physical meanings:

1. **Dust**: $w = 0$, $\rho \propto a^{-3}$. In this case the pressure vanishes for any value of the density. This is often used to model "ordinary" matter – stars, galaxies, dark matter etc. – since on very large scales there is negligible pressure between these objects. Note that in this case the energy is proportional to the volume element: as the universe expands (and a increases), matter simply dilutes.

2. **Radiation**: $w = \frac{1}{3}$, $\rho \propto a^{-4}$. If you have studies statistical physics then you will know that radiation can be treated as a perfect fluid with $p = \frac{1}{3}\rho$. In this case the energy density decreases both due to the increase in the volume element *and* due to redshift of the photons (see section 7.3).

3. **Dark energy**: $w = -1$, $\rho = $ const. In this case both the pressure and density are constant, independent of the behaviour of the scale factor a. In fact, in this case the energy-momentum tensor is just proportional to the metric g (**exercise**). This allows us to reinterpret the cosmological constant Λ as a component of "matter" rather than "gravity" – moving it onto the right hand side of the Einstein equations. However, there is no clear microscopic understanding of the origin of this "matter": it is usually thought of as the energy of the vacuum, i.e. the ground state energy of the quantum fields describing matter. However, cosmological observations place the value of Λ as almost zero, whereas calculations using quantum field theory predict that is should be around 10^{120} times larger[3]!

7.3 Cosmological redshift and the Hubble constant

Suppose Alice and Bob are both comoving observers. In terms of the coordinates (τ, r, θ, ϕ), suppose that Alice follows the worldline $(\tau, 0, 0, 0)$ and Bob the worldline $(\tau, r_B, \theta_B, \phi_B)$, where r_B, θ_B and ϕ_B are constants. Suppose that Bob sends light signals at regular intervals $\Delta\tau_B$ to Alice, who receives them at time intervals $\Delta\tau_A$.

[3]This has been called "the worst theoretical prediction in the history of physics." - M.P. Hobson, G.P. Efstathiou & A.N. Lasenby (2006), *General Relativity: An introduction for physicists.*

Radial null lines in a cosmological spacetime are null geodesics (this follows from isotropy, but **exercise**: check it explicitly). Hence the tangent to an affinely parametrised (ingoing) radial null geodesic is

$$X = \partial_\tau - \frac{\sqrt{1-kr^2}}{a}\partial_r$$

and the path of a null geodesic is $(\tau, r(\tau), \theta_0, \phi_0)$ where

$$\frac{\mathrm{d}r}{\mathrm{d}\tau} = -\frac{\sqrt{1-kr^2}}{a}$$

so if a light ray leaves Bob at time $\tau = \tau_B$ and arrives at Alice at $\tau = \tau_A$, then

$$\int_{r_B}^{0} -\frac{\mathrm{d}r}{\sqrt{1-kr^2}} = \int_{\tau_B}^{\tau_A} \frac{\mathrm{d}\tau}{a(\tau)}$$

Performing the same calculation for the subsequent signal, and noting that the left hand side is independent of τ, we have

$$\int_{\tau_B+\Delta\tau_B}^{\tau_A+\Delta\tau_A} \frac{\mathrm{d}\tau}{a(\tau)} = \int_{\tau_B}^{\tau_A} \frac{\mathrm{d}\tau}{a(\tau)}$$

If we suppose that $\Delta\tau_A$ and $\Delta\tau_B$ are very small compared to $\tau_B - \tau_A$, then we find that, to leading order,

$$\frac{\Delta\tau_A}{a(\tau_A)} = \frac{\Delta\tau_B}{a(\tau_B)}$$

so the ratio of the received frequency to the emitted frequency is

$$\frac{\Delta\tau_A}{\Delta\tau_B} = \frac{a(\tau_A)}{a(\tau_B)}$$

In an expanding universe, the scale factor a grows over time. Since $\tau_A > \tau_B$, $a(\tau_A) > a(\tau_B)$ and so $\Delta\tau_A > \Delta\tau_B$. So Alice sees the signals at a *lower* frequency than they are emitted by Bob: this is *cosmological redshift*.

Next suppose that Alice and Bob are close together (relative to the time scale on which a varies). Then, expanding the expression above, we have

$$\begin{aligned}\frac{\Delta\tau_A}{\Delta\tau_B} &= a(\tau_A)\Big(a(\tau_A) - (\tau_A-\tau_B)\dot{a}(\tau_A)\Big)^{-1} + \mathcal{O}\left((\tau_A-\tau_B)^2\right)\\ &= 1 + (\tau_A-\tau_B)\frac{\dot{a}(\tau_A)}{a(\tau_A)} + \mathcal{O}\left((\tau_A-\tau_B)^2\right)\\ &= 1 + (\tau_A-\tau_B)H(\tau_A) + \mathcal{O}\left((\tau_A-\tau_B)^2\right)\end{aligned}$$

where $H(\tau) := \frac{\dot{a}(\tau)}{a(\tau)}$ is the "Hubble constant". Note that the Hubble constant is not actually constant, but depends on time!

The expression $\frac{\Delta\tau_A}{\Delta\tau_B} \approx 1 + (\tau_A - \tau_B)H$ is known as *Hubble's law*. The quantity $(\tau_A - \tau_B)$ measures, in a natural sense, the *distance* from Bob to Alice: it is the time taken for light to travel from Bob to Alice. So Hubble's law says that light from nearby galaxies (assumed to move, roughly, along the worldlines of comoving observers) is redshifted in a way which scales linearly with the distance to that galaxy. Hubble's discovery of this law, with $H > 0$, provided the first clear evidence in favour of an expanding universe.

7.4 Solutions to the Friedmann equations

So far we have derived the Friedmann equations, discussed a few different models for matter, and shown that, in the case of an expanding universe, comoving observers will observe cosmological redshift. Our last task will be to actually *solve* the Friedmann equations, and to examine some of the properties of the solutions.

7.4.1 The Einstein static universe

The original motivation for introducing the cosmological constant Λ was to find a cosmological model which is static, so $\dot{a} = \ddot{a} = 0$. This solution is also supposed to model the universe now, when the matter is well approximated by 'dust', so we have $p = 0$. Substituting into the Friedmann equations (7.2), (7.1) we obtain

$$\begin{aligned} \frac{3k}{a^2} - \Lambda &= 8\pi\rho \\ k - a^2\Lambda &= 0 \end{aligned}$$

from which it follows that

$$\begin{aligned} k &= a^2\Lambda \\ \Lambda &= 4\pi\rho \end{aligned}$$

since we are dealing with ordinary matter we have $\rho > 0$, so $\Lambda > 0$ and hence we can set $k = +1$. This means that the solution is given, in terms of the energy density ρ, as

$$\begin{aligned} k &= 1 \\ \Lambda &= 4\pi\rho \\ a &= \frac{1}{\sqrt{4\pi\rho}} \end{aligned}$$

Unfortunately, this solution is dynamically unstable: small perturbations lead to a solution which either rapidly expands or collapses (see example sheet 4).

7.4.2 Matter dominated universes with $\Lambda = 0$

Observations now show conclusively that the universe is not static, but expanding, so let's look for such a solution. Again, we'll look for a "matter dominated universe", where the matter is well-approximated by dust, and for simplicity we'll take the cosmological constant $\Lambda = 0$.

In this case the Friedmann equations are

$$\begin{aligned} 3\frac{\dot{a}^2 + k}{a^2} &= 8\pi\rho \\ 2a\ddot{a} + \dot{a}^2 + k &= 0 \end{aligned}$$

Using this second equation, we find that

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(a\dot{a}^2 + ka\right) = \dot{a}\left(2a\ddot{a} + \dot{a}^2 + k\right) = 0$$

and so

$$\dot{a}^2 = \frac{C}{a} - k$$

where C is constant. We can identify this constant using the other Friedmann equation:

$$C = \frac{8\pi}{3}\rho a^3$$

Recall that, in the case of dust, $\rho \propto a^{-3}$ so C is indeed a constant.

First, let's consider the flat case $k = 0$. Then we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau}(a^{\frac{3}{2}}) &= \frac{3}{2}C^{\frac{1}{2}} \\ \Rightarrow a &= \left(\frac{3}{2}C^{\frac{1}{2}}\right)^{\frac{2}{3}}\tau^{\frac{2}{3}} \\ \rho &= \frac{1}{6\pi}\tau^{-2} \end{aligned}$$

where we have chosen the solution where $\dot{a} > 0$ and where $a = 0$ at $\tau = 0$. This universe expands from a 'big bang' at $\tau = 0$, where the scale factor goes to zero and the density becomes infinite as τ^{-2}.

Note that there is no "place" where the big bang happens: the solution remains homogeneous and isotropic at all times. Rather, the scale factor goes to zero, meaning that the proper distance on a surface Σ_τ between any two comoving observers goes to zero as $\tau \to 0$, and the density $\rho \to \infty$ *everywhere* as $\tau \to 0$. Nor is there a time 'before' the big bang: the spacetime as a whole is a solution to the Einstein equations for all $\tau > 0$, and spacetime terminates in a singularity at $\tau = 0$.

What about closed ($k = 1$) or open ($k = -1$) universes? Let's take $k = 1$ first. Then we have

$$\dot{a}^2 = \frac{C}{a} - 1$$

First we substitute $a = Cb^2$. Then we have

$$\frac{b^2}{\sqrt{1-b^2}} \frac{\mathrm{d}b}{\mathrm{d}\tau} = \pm \frac{1}{2C}$$

Now substituting $b = \sin u$ and doing the integral and a bit of algebra, we find

$$C\left(\arcsin\sqrt{\frac{a}{C}} - \sqrt{\frac{a}{C}}\sqrt{1-\frac{a}{C}}\right) = \pm\tau + \text{const.}$$

We can choose the constant to be zero, so that again $a = 0$ when $\tau = 0$. In this case we should also choose the + sign, so that for small positive values of τ, the scale factor a is positive.

For small values of τ, we find (**exercise**) to leading order

$$a = \left(\frac{3}{2}\right)^{\frac{2}{3}} C^{\frac{1}{3}} \tau^{\frac{2}{3}}$$

so that the scale factor approaches zero as $t^{\frac{2}{3}}$. As before, $\rho \sim a^{-3}$ so the energy density becomes infinite as t^{-2}.

There is an important difference in this case, however: at the time $\tau = C\pi$ we again have $a = 0$. So, in the closed case, there is a big bang followed by a *big crunch*, when the scale factor decreases to zero in the future and the universe recollapses!

Finally, consider the open case $k = -1$. Following similar algebra as in the closed case, we obtain the solution

$$C\left(\sqrt{\frac{a}{C}}\sqrt{1+\frac{a}{C}} - \operatorname{arsinh}\sqrt{\frac{a}{C}}\right) = \tau$$

As $\tau \to 0$ this solution behaves in the same way as the closed and flat cases, but this time, for large values of τ we have $a \sim \tau$, so these universes grow faster than their flat counterparts.

7.4.3 Radiation dominated universes

So far we have only looked at "matter dominated" universes, where the matter content is well approximated by dust, or pressure-free fluid. We can also consider the case where the primary matter content of the universe is described by radiation, so that the pressure is given by $p = \frac{1}{3}\rho$ instead of $p = 0$.

Again, setting $\Lambda = 0$ and setting

$$B = \frac{8\pi}{3} a^4 \rho$$

we find that B is a constant (so $\rho \sim a^{-4}$).

Following similar calculations as in the matter dominated case, the solutions are found to be (**exercise**)

$$k = 0 \qquad \Rightarrow \qquad a = \sqrt{2} B^{\frac{1}{4}} \tau^{\frac{1}{2}}$$

$$k = 1 \qquad \Rightarrow \qquad a = \sqrt{B} \sqrt{1 - \left(1 - \frac{\tau}{\sqrt{B}}\right)^2}$$

$$k = -1 \qquad \Rightarrow \qquad a = \sqrt{B} \sqrt{\left(1 + \frac{\tau}{\sqrt{B}}\right)^2 - 1}$$

Qualitatively these solutions behave similarly to the matter dominated scenarios: in the flat and open cases the universe just goes on expanding, while in the closed case the universe expands initially and subsequently contracts, ending in a big crunch. The rates, however, are different from those in the matter dominated cases.

7.4.4 The de-Sitter spacetime

Now we'll add back in the cosmological constant Λ, but for simplicity we'll now consider the case in which there is no matter (or, if you prefer, the only matter content of the universe is "dark energy").

In this case, the Friedmann equations are

$$3\frac{\dot{a}^2 + k}{a^2} - \Lambda = 0$$
$$2a\ddot{a} + \dot{a}^2 + k - a^2\Lambda = 0$$

Multiplying the first equation by a^3 and then taking a derivative with respect to τ, we see that (assuming $a > 0$) any smooth solution $a(\tau)$ to the first equation will automatically solve the second equation (**exercise**). So we can concentrate entirely on the first equation.

The solutions to this equation are

$$k = 0 \qquad \Rightarrow \qquad a \propto e^{\pm\sqrt{\frac{\Lambda}{3}}\tau}$$

$$k = 1 \qquad \Rightarrow \qquad a = \sqrt{\frac{3}{\Lambda}} \cosh\left(\sqrt{\frac{\Lambda}{3}}\tau\right)$$

$$k = -1 \qquad \Rightarrow \qquad a = \sqrt{\frac{3}{\Lambda}} \sinh\left(\sqrt{\frac{\Lambda}{3}}\tau\right)$$

It appears that we have three different solutions as before. But in fact, all three solutions can be shown to represent different parts of the *same* spacetime – called *de Sitter spacetime*. Moreover, this spacetime has no 'big bang' – in fact, in the right coordinates, it can be seen that this spacetime is actually *static*, so it doesn't evolve over time. The surface $\tau = 0$ in the open case ($k = -1$), where the metric appears to be singular (since $a = 0$) actually just represents a coordinate singularity, like the surface $r = 2M$ in Schwarzschild. Note that, in this case, $\rho \equiv 0$ so we cannot argue for a 'physical singularity' by saying that the energy density is infinite, as we could before!

7.4.5 Mixture models

A realistic model of the universe contains both 'dust' and 'radiation', as well as a cosmological constant $\Lambda > 0$ (albeit with Λ *almost* equal to zero!). In this case we can try to solve the Friedmann equations

by including two different components of the fluid, as well as a cosmological constant.

Recall the equation that we derived for the evolution of the density:

$$\frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{a}}{a}$$

from which it follows that $\rho \propto a^{-3(1+w)}$. Ultimately, this equation is derived from the conservation of the energy momentum tensor for the fluid, $\nabla_\mu T^{\mu\nu} = 0$. If we have several fluids which are *non-interacting*, then each of their energy-momentum tensors is *separately* conserved, and so for *each* fluid we will have

$$\rho_{(f)} \propto a^{-3(1+w_{(f)})}$$

where the subscript (f) labels the different fluids.

Although in general the equations cannot be solved explicitly for a general mixture of this sort, we can still obtain the following picture of what's going on:

1. At early times, when a is very small, the energy density of radiation goes as $\rho_{(\text{rad})} \sim a^{-4}$, and this is the dominant component. So at early times the universe is well approximated by a radiation dominated solution.

2. The energy density of dust behaves as $\rho_{(\text{dust})} \sim a^{-3}$. So as a grows larger, eventually the dust component dominates over the radiation component, and the universe behaves as a matter dominated solution.

3. The energy density of 'dark energy' (or the cosmological constant) is constant, but very small. Eventually, if the universe continues expanding, the scale factor a becomes so large that dark energy dominates over the matter component. At this point, the universe becomes approximately de-Sitter.

There is one important addendum to this picture which is often incorporated into modern cosmology, although it remains slightly controversial. To solve certain observational conundrums, cosmologists often suppose that there was an additional time, very early on in the evolution of the universe (before the radiation dominated era) when there was a period of exponential, de Sitter-like growth. This requires some very exotic matter which is able to mimic a large cosmological constant at early times, before falling away so that it is unobservable at the present time.

Appendix A

Normal coordinates

Let $p \in \mathcal{M}$ be a point in a Lorentzian manifold with metric g. We want to show that there are local coordinates in some region around p such that

1. $g_{ab}\big|_p = \text{diag}\,(-1, 1, 1, 1)$
2. $\Gamma^a_{bc}\big|_p = 0$

Consider the tangent space at p, $T_p(\mathcal{M})$. The idea is to match up vectors in the tangent space at p with points along the corresponding geodesics.

First, consider the metric at p, $g\big|_p$. This is a quadratic form on the vector space $T_p(\mathcal{M})$ – moreover, it is non-degenerate and has signature $(-,+++)$. By standard linear algebra, it is possible to find a basis for the vector space, say (e_0, e_1, e_2, e_3), such that the components of $g\big|_p$ with respect this basis are

$$g\big|_p(e_a, e_b) = \text{diag}\,(-1, 1, 1, 1)$$

Note that these are only the components of g with respect to the vectors e_a, which (for the time being) are not related to any kind of coordinate system, so we have not yet achieved our first goal.

Next, we introduce the *exponential map*, defined as follows:

$$\begin{aligned} \exp_p : T_p(\mathcal{M}) &\to \mathcal{M} \\ X &\mapsto \gamma_X(1) \end{aligned}$$

where γ_X is the affinely parametrised geodesic through p with tangent vector X at p, and with $\gamma_X(0) = p$. In other words, $\exp_p(X)$ is the point on the manifold reached by travelling a unit distance along the geodesic through p with initial tangent vector X.

This map is a smooth bijection in a neighbourhood of the origin. This can be seen by working in some (arbitrary) local coordinates in a neighbourhood of p: we have

$$\begin{aligned} (\phi_U \circ \exp_p) : T_p(\mathcal{M}) &\to \mathbb{R}^n \\ X &\mapsto (x^0(1), x^1(1), \ldots, x^{n-1}(1)) \end{aligned}$$

where
$$\begin{aligned} \frac{\mathrm{d}^2 x^a(s)}{\mathrm{d}s^2} + \Gamma^a_{bc}\big|_{x(s)} \frac{\mathrm{d}x^b(s)}{\mathrm{d}s}\frac{\mathrm{d}x^c(s)}{\mathrm{d}s} &= 0 \\ \Big(x^0(0), x^1(0), \ldots, x^{n-1}(0)\Big) &= \phi_U(p) \\ \left.\frac{\mathrm{d}x^a(s)}{\mathrm{d}s}\right|_{s=0} &= X^a = X\left(\phi_U^{-1}(x^a)\right) \end{aligned}$$

For any fixed vector X, the system of equations above have a unique solution for sufficiently small s – this follows from standard ODE theory. These equations are also invariant under $s \mapsto \lambda s$, $X \mapsto \lambda^{-1} s$, from which it follows that the map $(\phi_U \circ \exp_p)$ is well defined for all vectors X which are sufficiently close to the origin.

To check that this map gives a bijection in a neighbourhood of the origin we can use the inverse function theorem. It is clear that $(\phi_U \circ \exp_p)$ maps the origin of $T_p(\mathcal{M})$ to the origin of $\mathbb{R}^n$. If we write y^a for the solution to

$$\begin{aligned} \frac{\mathrm{d}^2 y^a(s)}{\mathrm{d}s^2} + \Gamma^a_{bc}\big|_{y(s)} \frac{\mathrm{d}y^b(s)}{\mathrm{d}s} \frac{\mathrm{d}y^c(s)}{\mathrm{d}s} &= 0 \\ \Big(y^0(0), y^1(0), \ldots, y^{n-1}(0)\Big) &= \phi_U(p) \\ \left.\frac{\mathrm{d}y^a(s)}{\mathrm{d}s}\right|_{s=0} &= \epsilon Y^a \end{aligned}$$

and then expanding to first order in ϵ, we see that y satisfies

$$\begin{aligned} \frac{\mathrm{d}^2 y^a(s)}{\mathrm{d}s^2} &= \mathcal{O}(\epsilon^2) \\ \Big(y^0(0), y^1(0), \ldots, y^{n-1}(0)\Big) &= \phi_U(p) \\ \left.\frac{\mathrm{d}y^a(s)}{\mathrm{d}s}\right|_{s=0} &= \epsilon Y^a \end{aligned}$$

the solution to which is $y^a(s) = \phi_U(p) + \epsilon Y^a s + \mathcal{O}(\epsilon^2)$. In particular, $y^a(1) = \epsilon Y^a + \mathcal{O}(\epsilon^2)$. From this we see that the differential of $(\phi_U \circ \exp_p)$ at the origin is the identity map, and hence (by the inverse function theorem) $(\phi_U \circ \exp_p)$ is invertible in a neighbourhood of the origin.

Finally, we note that, since ϕ_U is itself a bijection, and we have just seen that $(\phi_U \circ \exp_p)$ is a bijection, it follows that $\exp_p$ is a bijection (in a neighbourhood of the origin).

Now that we know that the exponential map is a bijection in a neighbourhood of the origin, we can use it to define some new local coordinates x^a near p. We do this using the vectors e_a as follows:

$$\begin{aligned} \phi_V(q) &= (x^0(q), x^1(q), \ldots, x^{n-1}(q)) \\ \text{where} \qquad \exp_p(x^a e_a) &= q \end{aligned}$$

Since $\exp_p$ is a bijection in a neighbourhood of the origin and the e_a form a basis, this uniquely defines the constants x^a, as long as q is sufficiently close to p.

Now, by the definition of vector fields, the vector $\left.\frac{\partial}{\partial x^a}\right|_p$ is the tangent to the curve of constant x^b, $b \neq a$ through the point p, parametrised by x^a. From the definitions above, this is the curve

$$x^a \mapsto \exp_p(x^a e_a)$$

which is simply the geodesic through p, with initial tangent vector e_a, and with affine parameter x^a. Hence we have

$$\left.\frac{\partial}{\partial x^a}\right|_p = e_a$$

and, since the e_a are orthonormal, the components of the metric g at the point p are

$$g_{ab}\big|_p = g\left(\frac{\partial}{\partial x^a}\ ,\ \frac{\partial}{\partial x^b}\right)\big|_p = g(e_a, e_b) = \mathrm{diag}\,(-1, 1, 1, 1)$$

This coordinate system $\{x^a\}$ therefore satisfies the first of our desired conditions. We now need to check the second condition, that is, that the Christoffel symbols vanish at p.

To check this, note that, for any set of constants X^a, the curve given by

$$x^a(s) = sX^a$$

is an affinely parametrised geodesic. Hence these curves satisfy the geodesic equation, i.e.

$$\begin{aligned} \frac{\mathrm{d}^2 x^a(s)}{\mathrm{d}s^2} + \Gamma^a_{bc}\big|_{x(s)} \frac{\mathrm{d}x^b(s)}{\mathrm{d}s}\frac{\mathrm{d}x^c(s)}{\mathrm{d}s} &= 0 \\ \Rightarrow \Gamma^a_{bc}\big|_p X^b X^c &= 0 \end{aligned}$$

Define $Z^a_{(b)} = \delta^a_b$. Then choosing $X^a = Z^a_{(b)}$ in the formula above, we conclude that

$$\Gamma^a_{bb}\big|_p = 0 \qquad \text{(no sum)}$$

On the other hand, if we choose $X^a = Z^a_{(b)} + Z^a_{(c)}$ then we find that

$$\Gamma^a_{bb}\big|_p + \Gamma^a_{bc}\big|_p + \Gamma^a_{cb}\big|_p + \Gamma^a_{cc}\big|_p = 0 \qquad \text{(no sum)}$$

and since the first and last terms were already shown to vanish, it follows that

$$\Gamma^a_{bc}\big|_p + \Gamma^a_{cb}\big|_p = 0$$

Finally, since the Christoffel symbols are symmetric in the lower indices, we see that the Christoffel symbols must vanish at p.