Part II Relativity

William Royce

October 31, 2024

Part II Physics, The University of Cambridge

Preface

> *"The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality."*

General relativity is the theory of space and time and gravity. The essence of the theory is simple: gravity is geometry. The effects that we attribute to the force of gravity are due to the bending and warping of spacetime, from falling cats, to orbiting spinning planets, to the motion of the cosmos on the grandest scale. The purpose of these lectures is to explain this.

Before we jump into a description of curved spacetime, we should first explain why Newton's theory of gravity, a theory which served us well for 250 years, needs replacing. The problems arise when we think about disturbances in the gravitational field. Suppose, for example, that the Sun was to explode. What would we see? Well, for 8 glorious minutes – the time that it takes light to reach us from the Sun – we would continue to bathe in the Sun's light, completely oblivious to the fate that awaits us. But what about the motion of the Earth? If the Sun's mass distribution changed dramatically, one might think that the Earth would start to deviate from its elliptic orbit. But when does this happen? Does it occur immediately, or does the Earth continue in its orbit for 8 minutes before it notices the change?

Of course, the theory of special relativity tells us the answer. Since no signal can propagate faster than the speed of light, the Earth must continue on its orbit for 8 minutes. But how is the information that the Sun has exploded then transmitted? Does the information also travel at the speed of light? What is the medium that carries this information? As we will see throughout these lectures, the answers to these questions forces us to revisit some of our most basic notions about the meaning of space and time and opens the to door to some of the greatest ideas in modern physics such as cosmology and black holes.

Contents

List of Figures

Chapter 1

Introduction

1.1 Newtonian Gravity

There is a well trodden path in physics when trying to understand how objects can influence other objects far away. We introduce the concept of a field. This is a physical quantity which exists everywhere in space and time; the most familiar examples are the electric and magnetic fields. When a charge moves, it creates a disturbance in the electromagnetic field, ripples of which propagate through space until they reach other charges. To develop a causal theory of gravity, we must introduce a gravitational field that responds to mass in some way.

It's a simple matter to cast Newtonian gravity in terms of a field theory. A particle of mass m_G experiences a force that can be written as

$$\mathbf{F} = -m_G \boldsymbol{\nabla}\Phi. \tag{1.1}$$

The quantity m_G is the *passive gravitational mass*, and it determines the gravitational force on the particle. The gravitational field $\Phi(\mathbf{r}, t)$ is determined by the surrounding matter distribution which is described by the mass density $\rho(\mathbf{r}, t)$. If the matter density is static, so that $\rho(\mathbf{r})$ is independent of time, then the gravitational field obeys

$$\nabla^2\Phi = 4\pi G\rho, \tag{1.2}$$

with Newton's constant G given by

$$G \approx 6.67 \times 10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}. \tag{1.3}$$

This equation is simply a rewriting of the usual inverse square law of Newton. For example, if a mass M is concentrated at a single point we have

$$\rho(\mathbf{r}) = M\delta^{(3)}(\mathbf{r}) \implies \Phi = -\frac{GM}{r}, \tag{1.4}$$

which is the familiar gravitational field for a point mass.

The question that we would like to answer is: how should we modify (1.2) when the mass distribution $\rho(\mathbf{r})$ changes with time? Of course, we could simply postulate that (1.2) continues to hold even in this case. A change in ρ would then immediately result in a change of Φ throughout all of space. Such a theory clearly is not consistent with the requirement that no signal can travel faster than light. Our goal is to figure out how to generalise (1.2) in a manner that is compatible with the postulates of special relativity. The end result of this goal will be a theory of gravity that is compatible with special relativity: this is the general theory of relativity.

Fixing this incompatibility will ultimately require a radical modification of how we think about gravity and, indeed, spacetime itself. Sticking with Newtonian gravity for

the moment, it is not immediately obvious that the mass density appearing in Poisson's equation should refer to the density of the passive gravitational mass. Rather, let us also introduce the active gravitational mass m_A, so that the relevant mass density for a point particle at position $\mathbf{r}'(t)$ at time t is

$$\rho(\mathbf{r}, t) = m_A \delta^{(3)}(\mathbf{r} - \mathbf{r}'(t)). \tag{1.5}$$

For the point particle, the relevant solution of Poisson's equation is

$$\Phi(\mathbf{r}, t) = -\frac{G m_A}{|\mathbf{r} - \mathbf{r}'(t)|}. \tag{1.6}$$

It follows that the force on a test particle of passive gravitational mass $m_{G,1}$ at position $\mathbf{r}_1$ at time t due to a particle of active gravitational mass $m_{A,2}$ at position $\mathbf{r}_2$ *at the same time* t is

$$\mathbf{F}_{2 \text{ on } 1} = -G m_{G,1} m_{A,2} \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2{}^3|}. \tag{1.7}$$

Similarly, the force on the second particle due to the first is

$$\mathbf{F}_{1 \text{ on } 2} = -G m_{G,2} m_{A,1} \frac{\mathbf{r}_2 - \mathbf{r}_1}{|\mathbf{r}_1 - \mathbf{r}_2{}^3|}. \tag{1.8}$$

If momentum is to be conserved, i.e., $\mathbf{F}_{2 \text{ on } 1} = \mathbf{F}_{1 \text{ on } 2}$, we must have

$$m_{G,1} m_{A,2} = m_{G,2} m_{A,1}. \tag{1.9}$$

Since this must hold for arbitrary masses, we must have that the ratio of passive to active gravitational mass is the same for all particles. Thus, we can take these masses to be equal, $m_G = m_A$, for all matter (absorbing their universal ratio in the gravitational constant).

This sort of universality is not unusual in physics – a similar thing happens in electromagnetism, for example, where the passive and active electric charges are equal. However, there is a further equality of masses in Newtonian gravity that is rather more surprising: the equality of gravitational and inertial masses. A particle acted on by a force $\mathbf{F}$ experiences an acceleration such that

$$\mathbf{F} = m_I \frac{\mathrm{d}^2 \mathbf{r}}{\mathrm{d}t^2}, \tag{1.10}$$

where m_I is the *inertial mass*. For the gravitational force, the acceleration is

$$\frac{\mathrm{d}^2 \mathbf{r}}{\mathrm{d}t^2} = -\frac{m_G}{m_I} \boldsymbol{\nabla} \Phi. \tag{1.11}$$

It is an experimental fact[1] (known since Galileo's time) that the ratio m_G/m_I is the same for all particles, so we can always take $m_G = m_I$ (further absorbing their universal ratio in the gravitational constant). This means that if two particles of different composition fall freely in a gravitational field, they have the same acceleration. This is often rephrased as the *weak equivalence principle*:

> Freely-falling particles with negligible gravitational self-interaction follow the same path through space and time if they have the same initial position and velocity, independent of their composition.

This property of gravity is in striking contrast to other forces; for example, in electromagnetism the acceleration of a point particle in a given electric field depends on the ratio of the electric charge to inertial mass, which is definitely not universal.

[1] The equality of gravitational and inertial masses is now verified to the level of one part in 10^{13}.

1.2 Implications of the Equivalence Principle

Consider an observer in a free-falling, non-rotating elevator in a uniform gravitational field. Relative to this observer, free-falling particles move on straight lines at constant velocity – the effects of the uniform gravitational field have been removed and the observer perceives that the usual laws of special relativistic kinematics hold. This idea motivates an extension of the weak equivalence principle to what is known as the *strong equivalence principle*:

> In an arbitrary gravitational field, *all* the laws of physics in a free-falling, non-rotating laboratory occupying a sufficiently small region of spacetime look locally like special relativity (with no gravity).

Note how the strong equivalence principle is supposed to apply to all laws of physics, not just the dynamics of free-falling particles. Why the qualification of observations over a sufficiently small region of spacetime?

Consider the same elevator falling freely in the non-uniform gravitational field of the earth. Free particles initially at rest in the elevator will move together over time as they follow radial trajectories towards the centre of the earth. It is these tidal effects that are the physical manifestation of the gravitational field, and that cannot be removed by passing to the free-falling frame. However, for sufficiently local measurements in space and time, these tidal effects are undetectable, and physics relative to the free-falling elevator looks just like special relativistic physics in an inertial frame of reference in the absence of gravity.

The strong equivalence principle implies the local equivalence of a gravitational field and acceleration. In particular, it implies that a constant gravitational field is unobservable – observations in a reference frame at rest in such a field would be indistinguishable from those in a uniformly-accelerating reference frame in the absence of gravity. In special relativity, physics looks simple when referred to an inertial frame, one defined by comoving, unaccelerated observers with synchronised clocks. However, with gravity, the equivalence principle tells us that physics looks equally simple *locally* in a free-falling reference frame, suggesting that we should *define* inertial reference frames locally by free-falling observers. Acceleration should be defined relative to such local inertial frames, so that a particle acted on by no other force (and so free-falling) should be regarded as unaccelerated.

1.2.1 Gravity as Spacetime Curvature

The universality of free fall suggested to Einstein that the trajectories of free-falling particles should be determined by the local structure of spacetime, rather than by the action of a gravitational force with a mysterious universal coupling to matter.

Local inertial reference frames correspond to local systems of coordinates over spacetime so that the geometry over a small region looks like that of the spacetime of special relativity. Gravity manifests itself through our inability to extend such coordinates globally, reflecting the *curvature of spacetime*.

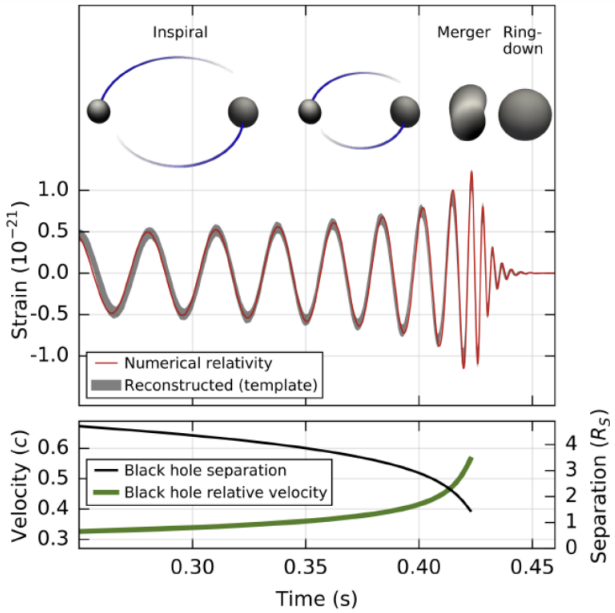

Fig. 1.1: Top: Estimated gravitational wave strain amplitude inferred from the LIGO data for their discovery event. The signal is generated from the inspiral, merger and ring-down of two massive black holes. The properties of the source can be estimated by comparing the measured waveform with detailed calculations in general relativity. Bottom: the relative speed and separation (in units of the Schwarzschild radius, $R_s = 2GM/c^2$) of the blackholes during the event. For reference, the Newtonian potential at R_S away from a mass M is $|\Phi|/c^2 = 1/2$. Figure taken from Abbot et al., Phys. Rev. Lett. 116, 061102 (2016).

General relativity abandons the idea of gravity as a force defined on the fixed spacetime of special relativity, replacing it with a geometric theory in which the geometry of spacetime determines the trajectories of free-falling particles, the geometry itself being curved by the presence of matter.

1.3 Further Motivation: Extreme Gravity

Newtonian gravity is recovered from general relativity in the limit of low relative speeds of particles, $v \ll c$, and weak gravitational fields, typically $|\Phi| \ll c^2$. Note that in situations where speeds are determined by gravity, these two regimes are generally equivalent.

To see this, consider a particle in a circular orbit of radius R around a mass M in Newtonian gravity: the speed is determined by

$$\frac{v^2}{R} = \frac{GM}{R^2}, \tag{1.12}$$

and so

$$\frac{v^2}{c^2} = \frac{GM}{Rc^2} = \frac{|\Phi|}{c^2}. \tag{1.13}$$

However, increasingly we are observing phenomena where Newtonian gravity is a very poor approximation. A striking example is the recent first detection of gravitational waves by the LIGO interferometer; see Fig. 1.1.

Gravitational waves are wavelike disturbances in the geometry of spacetime, which can be detected by looking for their characteristic quadrupole distortion (i.e., a shortening in one direction and stretching in an orthogonal direction) of the two arms of a laser interforemeter. Gravitational waves propagate at the speed of light and are a natural prediction of general relativity; they do not arise in Newtonian gravity where the potential responds instantly to distant rearrangements of mass.

The first LIGO signal was generated by a truly extreme astrophysical source: two merging black holes each with a mass around 30 times that of the Sun at a distance from us of around 2 Gly.As the blackholes orbited their common centre of mass, the system radiated gravitational waves causing the blackholes to spiral inwards and increase their speed until they merged to form a single black hole. Such sources probe the strong-field regime of general relativity during the merger phase and involve highly relativistic speeds (see Fig. 1.1). At its peak, the source was losing energy to gravitational waves at a rate of 3.6×10^{49}W, which is equivalent to 200 times the rest mass energy of the Sun per second!

Chapter 2

Recap of Special Relativity

Although Newtonian mechanics gives an excellent description of Nature, it is not universally valid. When we reach extreme conditions — the very small, the very heavy or the very fast — the Newtonian Universe that we're used to needs replacing. You could say that Newtonian mechanics encapsulates our common sense view of the world. One of the major themes of twentieth century physics is that when you look away from our everyday world, common sense is not much use.

One such extreme is when particles travel very fast. The theory that replaces Newtonian mechanics is due to Einstein. It is called *special relativity*. The effects of special relativity become apparent only when the speeds of particles become comparable to the speed of light in the vacuum. The speed of light is

$$c = 299792458 \mathrm{m\,s^{-1}} \tag{2.1}$$

This value of c is exact. It may seem strange that the speed of light is an integer when measured in meters per second. The reason is simply that this is taken to be the definition of what we mean by a meter: it is the distance travelled by light in $1/299792458$ seconds. For the purposes of this course, we'll be quite happy with the approximation $c \approx 3 \times 10^8 \mathrm{m\,s^{-1}}$.

The first thing to say is that the speed of light is fast. Really fast. The speed of sound is around $300 \mathrm{m\,s^{-1}}$; escape velocity from the Earth is around $104 \mathrm{m\,s^{-1}}$; the orbital speed of our solar system in the Milky Way galaxy is around $105 \mathrm{m\,s^{-1}}$. As we shall soon see, nothing travels faster than c.

The theory of special relativity rests on two experimental facts. (We will look at the evidence for these shortly). In fact, the first of these is simply the Galilean principle of relativity as in classical Newtonian mechanics. The second postulate is more surprising:

- The principle of relativity: the laws of physics are the same in all inertial frames.
- The speed of light in vacuum is the same in all inertial frames.

On the face of it, the second postulate looks nonsensical. How can the speed of light look the same in all inertial frames? If light travels towards me at speed c and I run away from the light at speed v, surely I measure the speed of light as $c - v$. Right? Well, no.

2.1 Newtonian Geometry of Space and Time

Newtonian theory assumes an absolute time – the same for every observer. This common sense view is encapsulated in the Galilean transformations.. Mathematically, we derive this

"obvious" result as follows: two inertial frames, S and S', in *standard configuration*: axes aligned, the same spacetime origin, which move relative to each with velocity $\mathbf{v} = (v, 0, 0)$, have Cartesian coordinates related by

$$x' = x - vt, \qquad y' = y, \qquad z' = z, \qquad t' = t \tag{2.2}$$

If a ray of light travels in the x direction in frame S with speed c, then it traces out the trajectory $x/t = c$. The transformations above then tell us that in frame S0 the trajectory if the light ray is $x'/t' = c - nv$. This is the result we claimed above: the speed of light should clearly be $c - v$. If this is wrong (and it is) something must be wrong with the Galilean transformations (2.2). But what?

Our immediate goal is to find a transformation law that obeys both postulates above. As we will see, the only way to achieve this goal is to allow for a radical departure in our understanding of time. In particular, we will be forced to abandon the assumption of absolute time, enshrined in the equation $t' = t$ above. We will see that time ticks at different rates for observers sitting in different inertial frames.

For two events A and B, the Galliliean transformation implies that

- the time difference $\Delta t = t_B - t_A$ is invariant; and
- $\Delta r^2 = \Delta x^2 + \Delta y^2 + \Delta z^2$ is invariant for simultaneous events (since Δx, Δy, and Δz are).

Space and time are separate entities in Newtonian theory.

2.2 Lorentz Transformations

We stick with the idea of two inertial frames, S and S', moving with relative speed v. For simplicity, we'll start by ignoring the directions y and z which are perpendicular to the direction of motion. Both inertial frames come with Cartesian coordinates: (x, t) for S and (x', t') for S' . We want to know how these are related. The most general possible relationship takes the form

$$x' = f(x, t), \qquad t' = g(x, t), \tag{2.3}$$

for some function f and g. However, there are a couple of facts that we can use to immediately restrict the form of these functions. The first is that the law of inertia holds; left alone in an inertial frame, a particle will travel at constant velocity. Drawn in the (x, t) plane, the trajectory of such a particle is a straight line. Since both S and S' are inertial frames, the map $(x, t) \mapsto (x', t'))$ must map straight lines to straight lines; such maps are, by definition, linear. The functions f and g must therefore be of the form

$$x' = \alpha_1 x + \alpha_2 t, \qquad t' = \alpha_3 x + \alpha_f t, \tag{2.4}$$

where $\alpha_i = 1, 2, 3, 4$ can each be a function of v.

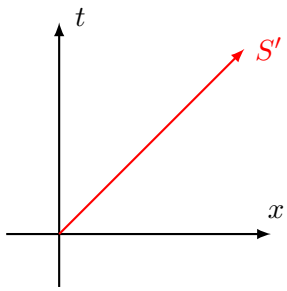

Fig. 2.1: Space-time diagram, representing the motion of a particle at the origin $x' = 0$ in S', which moves along the trajectory $x = vt$ in S.

Secondly, we use the fact that S' is travelling at speed v relative to S. This means that an observer sitting at the origin, $x' = 0$, of S' moves along the trajectory $x = vt$ in S shown in Fig. (2.1). Or, in other words, the points $x = vt$ must map to $x' = 0$. (There is actually one further assumption implicit in this statement: that the origin $x' = 0$ coincides with $x = 0$ when $t = 0$). Together with the requirement that the transformation is linear, this restricts the coefficients α_1 and α_2 above to be of the form,

$$x' = \gamma(x - vt), \tag{2.5}$$

for some coefficient γ. Once again, the overall coefficient can be a function of the velocity: $\gamma = \gamma_v$. (We've used subscript notation v rather than the more standard (v) to denote that depends on v. This avoids confusion with the factors of $(x - vt)$ which aren't arguments of but will frequently appear after like in the equation (2.5)).

There is actually a small, but important, restriction on the form of γ_v: it must be an even function, so that $\gamma_v = \gamma_{-v}$. There are a couple of ways to see this. The first is by using rotational invariance, which states that can depend only on the direction of the relative velocity $\mathbf{v}$, but only on the magnitude $v^2 = \mathbf{v} \cdot \mathbf{v}$. Alternatively, if this is a little slick, we can reach the same conclusion by considering inertial frames $\tilde{S}$ and $\tilde{S}'$ which are identical to S and S' except that we measure the x-coordinate in the opposite direction, meaning $\tilde{x} = -x$ and $\tilde{x}' = -x'$. While S is moving with velocity $+v$ relative to S', $\tilde{S}$ is moving with velocity $-v$ with respect to $\tilde{S}'$ simply because we measure things in the opposite direction. That means that

$$\tilde{x}' = \gamma_{-v}(\tilde{x} + v\tilde{t}). \tag{2.6}$$

Comparing this to (2.5), we see that we must have $\gamma_v = \gamma_{-v}$ as claimed.

We can also look at things from the perspective of S', relative to which the frame S moves backwards with velocity v. The same argument that led us to (2.5) now tells us that

$$x = \gamma(x' + vt'). \tag{2.7}$$

Now the function $\gamma_v = \gamma_{-v}$. But by the argument above, we know that v = v. In other words, the coefficient appearing in (2.7) is the same as that appearing in (2.5).

At this point, things don't look too different from what we've seen before. Indeed, if we now insisted on absolute time, so $t = t'$, we're forced to have $\gamma = 1$ and we get back to the Galilean transformations (2.2). However, as we've seen, this is not compatible with

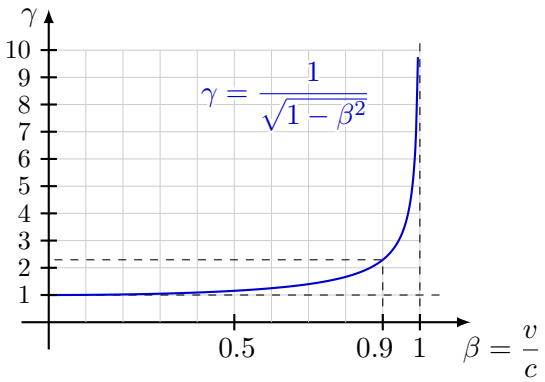

Fig. 2.2: Lorentz factor as a function of velocity $\beta = v/c$ (i.e. in units of c).

the second postulate of special relativity. So let's push forward and insist instead that the speed of light is equal to c in both S and S'. In S, a light ray has trajectory

$$x = ct. \tag{2.8}$$

While, in S', we demand that the same light ray has trajectory

$$x' = ct'. \tag{2.9}$$

Substituting these trajectories into (2.5) and (2.7), we have two equations relating t and t',

$$ct' = \gamma(c - v)t, \qquad \text{and,} \qquad ct = \gamma(c + v)t'. \tag{2.10}$$

A little algebra shows that these two equations are compatible only if γ is given by

$$\boxed{\gamma = \sqrt{\frac{1}{1 - v^2/c^2}}.} \tag{2.11}$$

We'll be seeing a lot of this coefficient γ in what follows. Notice that for $v \ll c$, we have $\gamma \approx 1$ and the transformation law (2.5) is approximately the same as the Galilean transformation (2.2). However, as $v \to c$ we have $\gamma \to \infty$. Furthermore, becomes imaginary for $v > c$ which means that we're unable to make sense of inertial frames with relative speed $v > c$.

Equations (2.5) and (2.11) give us the transformation law for the spatial coordinate. But what about for time? In fact, the temporal transformation law is already lurking in our analysis above. Substituting the expression for x' in (2.5) into (2.7) and rearranging, we get

$$t' = \gamma\left(t - \frac{v}{c^2}x\right). \tag{2.12}$$

We shall soon see that this equation has dramatic consequences. For now, however, we merely note that when $v \ll c$, we recover the trivial Galilean transformation law $t' \approx t$. Equations (2.5) and (2.12) are the *Lorentz transformations*.

2.2.1 Lorentz Transformations in Three Spatial Dimensions

In the above derivation, we ignored the transformation of the coordinates y and z perpendicular to the relative motion. In fact, these transformations are trivial. Using the above arguments for linearity and the fact that the origins coincide at $t = 0$, the most general form of the transformation is

$$y' = \kappa y, \tag{2.13}$$

But, by symmetry, we must also have $y' = \kappa y$. Clearly, we require $\kappa = 1$. (The other possibility $\kappa = -1$ does not give the identity transformation when $v = 0$. Instead, it is a reflection).

With this we can write down the final form of the Lorentz transformations. Note that they look more symmetric between x and t if we write them using the combination ct,

$$\begin{aligned} x' &= \gamma\left(x - \frac{v}{c}ct\right), \\ y' &= y, \\ z' &= z, \\ ct' &= \gamma\left(ct - \frac{v}{c}x\right), \end{aligned} \tag{2.14}$$

where γ is given by (2.11). These are also known as Lorentz boosts. Notice that for $v/c \ll 1$, the Lorentz boosts reduce to the more intuitive Galilean boosts. (We sometimes say, rather sloppily, that the Lorentz transformations reduce to the Galilean transformations in the limit $c \to \infty$).

It's also worth stressing again the special properties of these transformations. To be compatible with the first postulate, the transformations must take the same form if we invert them to express x and t in terms of x' and t', except with v replaced by $-v$. And, after a little bit of algebraic magic, they do.

Secondly, we want the speed of light to be the same in all inertial frames. For light travelling in the x direction, we already imposed this in our derivation of the Lorentz transformations. But it's simple to check again: in frame S, the trajectory of an object travelling at the speed of light obeys $x = ct$. In S', the same object will follow the trajectory $x' = \gamma(x - vt) = \gamma(ct - vx/c) = ct'$

What about an object travelling in the y direction at the speed of light? Its trajectory in S is $y = ct$. From (2.14), its trajectory in S' is $y' = ct'/\gamma$ and $x' = vt'$. Its speed in S' is therefore $v'^2 = v_x^2 x + v_y^2$, or

$$v'^2 = \left(\frac{x'}{t'}\right)^2 + \left(\frac{x'}{t'}\right)^2 = v^2 + \frac{c^2}{\gamma^2} = c^2. \tag{2.15}$$

Note how time and space are mixed by the Lorentz transformation. However, for two events, the (squared) interval

$$\boxed{\Delta s^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2.} \tag{2.16}$$

is *invariant* under any Lorentz transformation. In special relativity, space and time are united into a four-dimensional continuum called spacetime with invariant geometry characterised by Δs^2. The spacetime of special relativity is topologically $\mathbb{R}^4$. When endowed with the measure of distance (2.16), this spacetime is referred to as Minkowski space. Although topologically equivalent to Euclidean space, distances are measured differently. To stress the difference between the time and spatial directions, Minkowski space is sometimes said to have dimension $d = 1 + 3$. (For once, it's important that you don't do this sum!). In later courses – in particular General Relativity – you will see the invariant interval written as the distance between two infinitesimally close points. In practice that just means we replace all the Δ(something)s with d(something)s.

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2\,. \tag{2.17}$$

In this infinitesimal form, $\mathrm{d}s^2$ is called the *line element*.

2.2.2 Lorentz Transformations as 4D "Rotations"

Different Cartesian inertial frames S and S' simply relabel events in Minkowski spacetime, i.e., perform a coordinate transformation $(ct, x, y, z) \to (ct', x', y', z')$. It is often convenient to define the rapidity parameter ψ (which runs from $-\infty$ to ∞) by $v/c = \tanh\psi$, so that

$$\gamma = \cosh\psi, \qquad \text{and,} \qquad \gamma v/c = \tanh\psi \tag{2.18}$$

For S and S' in standard configuration, we have

$$\begin{aligned} ct' &= ct\cosh\psi - x\sinh\psi \\ x' &= -ct\sinh\psi + x\cosh\psi \\ y' &= y \\ z' &= z \end{aligned} \tag{2.19}$$

These are like a rotation in the $ct - x$ plane, but with hyperbolic rather than trigonometric functions. The hyperbolic functions are necessary to ensure the invariance of Δs^2 given the minus signs in its definition.

2.2.3 More Complicated Lorentz Transformations

More generally, the relation between two Cartesian inertial frames S and S' can differ from that for the standard configuration since[1]:

- the 4D origins may not coincide, i.e., the event at $ct = x = y = z = 0$ may not be at $ct' = x' = y' = z' = 0$;
- the relative velocity of the two frames may be in an arbitrary direction in S, rather than along the x-axis; and

[1] Lorentz transformations can be considered more formally as linear transformations that preserve the interval Δs^2. In this case, the definition admits transformations that are not continuously connected to the identity; i.e., parity transformations and/or time reversal. We shall not consider such transformations further.

- the spatial axes in S and S' may not be aligned, e.g., the components of the relative velocity in S' may not be minus those in S.

We can always deal with the origins not coinciding (known as inhomogeneous Lorentz transformations or Poincaré transformations) by appropriate temporal and spatial displacements. We can find the form of the remaining Lorentz transformation in the general case by decomposing as follows.

1. Apply a purely spatial rotation in the frame S to align the new x-axis with the relative velocity of the two frames.
2. Apply a standard Lorentz transformation as in Eq. (2.12) and (2.5).
3. Apply a spatial rotation in the transformed coordinates to align the axes with those of S'.

Given a reference frame S, the *Lorentz boost* of this frame for a general relative velocity $\mathbf{v}$ is obtained by rotating the spatial axes of S so that the relative velocity is along the new x-axis, applying the standard Lorentz transformation, and applying the inverse spatial rotation in the transformed frame.If the relative velocity is along the original x-axis, this reduces to the standard Lorentz transformation.

More generally, reference frames connected by a Lorentz boost have their spatial axes as aligned as possible given the relative velocity of the frames, i.e., they are generated by hyperbolic "rotations" in the plane defined by the ct-axis and the relative velocity.

2.2.3.1 General Velocity Lorentz Transformation

Let us now consider an inertial frame S' is related to the frame S by a boost of $\vec{v}$ whose components in S are (v_x, v_y, v_z). We aim to resolve the 3-vector position into components parallel and perpendicular to the 3-vector velocity.

Let us define

$$\vec{\beta} = \vec{v}/c = (\beta_x, \beta_y, \beta_z), \tag{2.20}$$

a natural extension of what we have already seen, and from this define the unit normal vector

$$\hat{\mathbf{n}} = \frac{1}{\sqrt{\beta_x^2 + \beta_y^2 + \beta_z^2}}(\beta_x, \beta_y, \beta_z) = \frac{1}{\left|\vec{\beta}\right|}\vec{\beta} \tag{2.21}$$

Thus the parallel component of the 3-vector position is given by

$$\begin{aligned} \vec{r}_{\|} &= (\hat{\mathbf{n}} \cdot \vec{r})\hat{\mathbf{n}} \\ &= \frac{1}{\left|\vec{\beta}\right|^2}\left(\vec{\beta} \cdot \vec{r}\right)\vec{\beta}. \end{aligned} \tag{2.22}$$

The component perpendicular to $\vec{v}$ is now most easily obtained by simply subtracting the parallel component,

$$\begin{aligned}\vec{r}_\perp &= \vec{r} - \vec{r}_\parallel \\ &= \vec{r} - \frac{1}{\left|\vec{\beta}\right|^2}\left(\vec{\beta}\cdot\vec{r}\right)\vec{\beta}.\end{aligned} \tag{2.23}$$

Now we are in a place to Lorentz transform both the perpendicular and parallel components by analogy to (2.14),

$$\vec{r}_\parallel{}' = \gamma\left(\vec{r}_\parallel - \vec{\beta}ct\right) \tag{2.24}$$

$$\vec{r}_\perp{}' = \vec{r}_\perp, \tag{2.25}$$

to give us

$$\begin{aligned}\vec{r}' &= \vec{r}_\parallel{}' + \vec{r}_\perp{}' \\ &= \gamma\left(\frac{1}{\left|\vec{\beta}\right|^2}\left(\vec{\beta}\cdot\vec{r}\right)\vec{\beta} - \vec{\beta}ct\right) + \vec{r} - \frac{1}{\left|\vec{\beta}\right|^2}\left(\vec{\beta}\cdot\vec{r}\right)\vec{\beta} \\ &= \frac{(\gamma-1)}{\left|\vec{\beta}\right|^2}\left(\vec{\beta}\cdot\vec{r}\right)\vec{\beta} - \gamma\vec{\beta}ct + \vec{r}.\end{aligned} \tag{2.26}$$

Let us introduce $\alpha = (\gamma-1)/\left|\vec{\beta}\right|^2$ to simplify this result further to

$$\vec{r}' = \alpha\left(\vec{\beta}\cdot\vec{r}\right)\vec{\beta} - \gamma\vec{\beta}ct + \vec{r}. \tag{2.27}$$

But we also need an expression for the transformation $ct \to ct'$! Let us generalise the last Lorentz transformation (2.14) to

$$ct' = \gamma(ct - \vec{\beta}\cdot\vec{r}). \tag{2.28}$$

Thus, in explicit component form we see

$$ct' = \begin{pmatrix}\gamma & -\gamma\beta_x & -\gamma\beta_y & -\gamma\beta_z\end{pmatrix}\begin{pmatrix}ct\\x\\y\\z\end{pmatrix}. \tag{2.29}$$

We can also see that Eq. (2.26) reduces to

$$\begin{aligned}\begin{pmatrix}x'\\y'\\z'\end{pmatrix} &= \alpha(\beta_x x + \beta_y y + \beta_z z)\begin{pmatrix}\beta_x\\\beta_y\\\beta_z\end{pmatrix} - \gamma ct\begin{pmatrix}\beta_x\\\beta_y\\\beta_z\end{pmatrix} + \begin{pmatrix}x\\y\\z\end{pmatrix} \\ &= \begin{pmatrix}\alpha\beta_x^2 x + \alpha\beta_x\beta_y y + \alpha\beta_x\beta_z z + x\\ \alpha\beta_x\beta_y x + \alpha\beta_y^2 + \alpha\beta_y\beta_z z + y\\ \alpha\beta_x\beta_z x + \alpha\beta_y\beta_z y + \alpha\beta_z^2 z + z\end{pmatrix} - \begin{pmatrix}\gamma\beta_x\\\gamma\beta_y\\\gamma\beta_z\end{pmatrix}ct \\ &= \begin{pmatrix}-\gamma\beta_x & 1+\alpha\beta_x^2 & \alpha\beta_x\beta_y & \alpha\beta_x\beta_z\\ -\gamma\beta_y & \alpha\beta_y\beta_x & 1+\alpha\beta_y^2 & \alpha\beta_y\beta_z\\ -\gamma\beta_z & \alpha\beta_z\beta_x & \alpha\beta_z\beta_y & 1+\alpha\beta_z^2\end{pmatrix}\begin{pmatrix}ct\\x\\y\\z\end{pmatrix}.\end{aligned} \tag{2.30}$$

Finally, it should be clear that we can simply combine Eqs. (2.29) and (2.30) to form a single matrix equation to show that the coordinates (ct', x', y', z') and (ct, x, y, z) of an event are related by

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma\beta_x & -\gamma\beta_y & -\gamma\beta_z \\ -\gamma\beta_x & 1+\alpha\beta_x^2 & \alpha\beta_x\beta_y & \alpha\beta_x\beta_z \\ -\gamma\beta_y & \alpha\beta_y\beta_x & 1+\alpha\beta_y^2 & \alpha\beta_y\beta_z \\ -\gamma\beta_z & \alpha\beta_z\beta_x & \alpha\beta_z\beta_y & 1+\alpha\beta_z^2 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}. \tag{2.31}$$

where $\vec{\beta} = \vec{v}/c$, $\gamma = (1 - |\vec{\beta}|^2)^{-1/2}$ and $\alpha = (\gamma - 1)/|\vec{\beta}|^2$.

2.2.4 The Interval

As we have seen, the interval is invariant under Lorentz transformations. This is particularly transparent using the hyperbolic form of the standard transformation:

$$\begin{aligned} \Delta s^2 &= c^2(\Delta t')^2 - (\Delta x')^2 - (\Delta y')^2 - (\Delta z')^2 \\ &= [(c\Delta t)\cosh\psi - (\Delta x)\sinh\psi]^2 - [-(c\Delta t)\sinh\psi + (\Delta x)\cosh\psi]^2 - \Delta y^2 - \Delta z^2 \\ &= c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2 \end{aligned} \tag{2.32}$$

The interval is also invariant under more general Lorentz transformations since a shift in origin does not alter the differences $(c\Delta t, \Delta x, \Delta y, \Delta z)$, and rotations preserve the spatial interval $\Delta x^2 + \Delta y^2 + \Delta z^2$.

The invariant interval provides an observer-independent characterisation of the distance between any two events. However, it has a strange property: it is not positive definite. Two events whose separation is $\Delta s^2 > 0$ are said to be *timelike* separated. They are closer together in space than they are in time. Pictorially, such events sit within each others light cone.

In contrast, events with $\Delta s^2 < 0$ are said to be *spacelike* separated. They sit outside each others light cone. Two observers can disagree about the temporal ordering of spacelike separated events. However, they agree on the ordering of timelike separated events. Note that since $\Delta s^2 < 0$ for spacelike separated events, if you insist on talking about Δs itself then it must be purely imaginary. However, usually it will be perfectly fine if we just talk about Δs^2.

Finally, two events with $\Delta s^2 = 0$ are said to be *lightlike* separated. Notice that this is an important difference between the invariant interval and most measures of distance that you're used to. Usually, if two points are separated by zero distance, then they are the same point. This is not true in Minkowski spacetime: if two points are separated by zero distance, it means that they can be connected by a light ray.

2.2.5 Space-Time Diagrams

We'll find it very useful to introduce a simple spacetime diagram to illustrate the physics of relativity. In a fixed inertial frame, S, we draw one direction of space – say x – along

the horizontal axis and time on the vertical axis. But things look much nicer if we rescale time and plot ct on the vertical instead. In the context of special relativity, space and time is called *Minkowski space*.

This is a spacetime diagram. Each point, P, represents an event. In the following, we'll label points on the spacetime diagram as coordinates (ct, x) i.e. giving the coordinate along the vertical axis first. This is backwards from the usual way coordinates but is chosen so that it is consistent with a later, standard, convention.

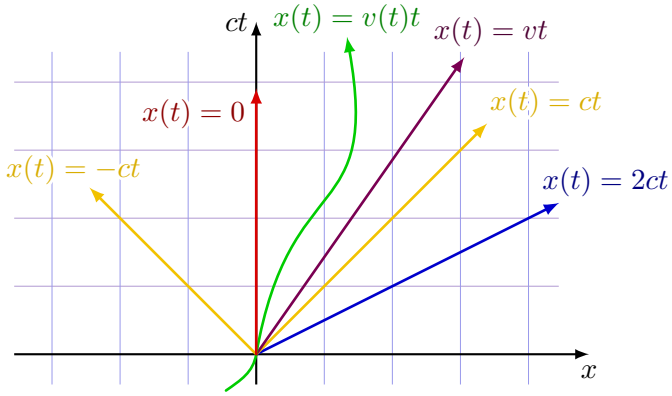

Fig. 2.3: Several spacetime worldlines with different velocities.

A particle moving in spacetime traces out a curve called a worldline as shown in Fig 2.3. Because we've rescaled the time axis, a light ray moving in the x direction moves at 45°. We'll later see that no object can move faster than the speed of light which means that the worldlines of particles must always move upwards at an angle steeper than 45°.

The horizontal and vertical axis in the spacetime diagram are the coordinates of the inertial frame S. But we could also draw the axes corresponding to an inertial frame S' moving with relative velocity $\mathbf{v} = (v, 0, 0)$. The t' axis sits at $x' = 0$ and is given by $x = \frac{v}{c}ct$. Meanwhile, the x' axis is determined by $t' = 0$ which, from the Lorentz transformation (2.14), is given by the equation $ct = \frac{v}{c}x$. It is immediately clear that the angle between the x and x' axes is the same as that between the t and t' axes, and has value $\tan^{-1}(v/c)$.

These two axes are drawn on the Fig. (2.4). They can be the thought of as the x and ct axes, rotated by an equal amount towards the diagonal light ray. The fact the axes are symmetric about the light ray reflects the fact that the speed of light is equal to c in both frames.

2.2.6 Causality and the Lightcone

We start with a simple question: how can we be sure that things happen at the same time? In Newtonian physics, this is a simple question to answer. In that case, we have an absolute time t and two events, P_1 and P_2, happen at the same time if $t_1 = t_2$. However, in the relativistic world, things are not so easy.

We start with an observer in inertial frame S, with time coordinate t. This observer

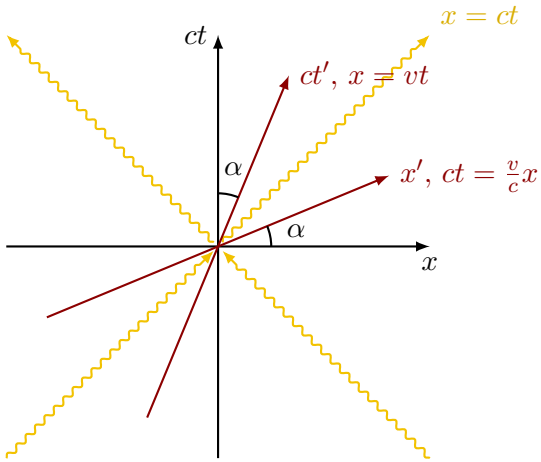

Fig. 2.4: Space-time diagram with axes corresponding to an inertial frame S' moving with a relative velocity. They can be the thought of as the x and ct axes, rotated by an equal amount towards the diagonal light ray. The fact the axes are symmetric about the light ray reflects the fact that the speed of light is equal to c in both frames.

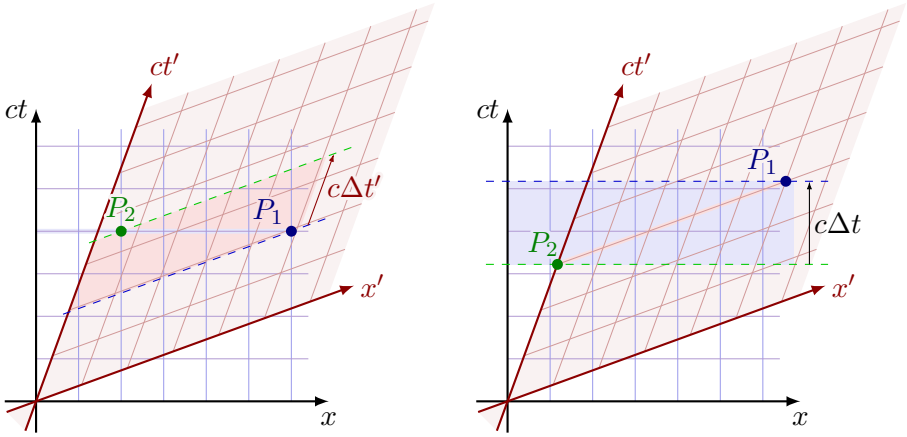

Fig. 2.5: Left: Events P_2 and P_1 are simultaneous in the rest frame S, but in the boosted frame S', P_1 happens before P_2. Right: Events P_2 and P_1 are simultaneous in the boosted frame S', but in the rest frame S, P_2 happens before P_1.

sensibly decides that two events, P_1 and P_2, occur simultaneously if $t_1 = t_2$. In the spacetime diagram in Fig. 2.5, we have drawn lines of simultaneity for this observer in light blue.

But for an observer in the inertial frame S', simultaneity of events occurs for equal t'. Using the Lorentz transformation, lines of constant t' become lines described by the equation $t - vx/c^2 =$ constant. These lines are drawn as light red on the spacetime diagrams in Fig. 2.5.

The upshot of this is that two events simultaneous in one inertial frame are not simultaneous in another. An observer in S thinks that events P_1 and P_2 happen at the same time. All other observers disagree.

We've seen that different observers disagree on the temporal ordering of two events. But where does that leave the idea of causality? Surely it's important that we can say

that one event definitely occurred before another. Thankfully, all is not lost: there are only some events which observers can disagree about.

To see this, note that because Lorentz boosts are only possible for $v < c$, the lines of simultaneity cannot be steeper than 45°. Take a point A and draw the 45° light rays that emerge from A. This is called the *light cone*. In more than a single spatial dimension, the light cone is really two cones, touching at the point A. They are known as the future light cone and past light cone.

For events inside the light cone of A, there is no difficulty deciding on the temporal ordering of events. All observers will agree that B occurred after A. However, for events outside the light cone, the matter is up for grabs: some observers will see D as happening after A; some before.

This tells us that the events which all observers agree can be causally influenced by A are those inside the future light cone. Similarly, the events which can plausibly influence A are those inside the past light cone. This means that we can sleep comfortably at night, happy in the knowledge that causality is preserved, only if nothing can propagate outside the light cone. But that's the same thing as travelling faster than the speed of light.

The converse to this is that if we do ever see particles that travel faster than the speed of light, we're in trouble. We could use them to transmit information faster than light. But another observer would view this as transmitting information backwards in time. All our ideas of cause and effect will be turned on their head. We will show later why it is impossible to accelerate particles past the light speed barrier.

There is a corollary to the statement that events outside the lightcone cannot influence each other: there are no perfectly rigid objects. Suppose that you push on one end of a rod. The other end cannot move immediately since that would allow us to communicate faster than the speed of light. Of course, for real rods, the other end does not move instantaneously. Instead, pushing on one end of the rod initiates a sound wave which propagates through the rod, telling the other parts to move. The statement that there is no rigid object is simply the statement that this sound wave must travel slower than the speed of light.

Finally, let me mention that when we're talking about waves, as opposed to point particles, there is a slight subtlety in exactly what must travel slower than light. There are at least two velocities associated to a wave: the group velocity is (usually) the speed at which information can be communicated. This is less than c. In contrast, the phase velocity is the speed at which the peaks of the wave travel. This can be greater than c, but transmits no information.

2.2.6.1 A Potential Confusion: What the Observer Observes

We'll pause briefly to press home a point that may lead to confusion. You might think that the question of simultaneity has something to do with the finite speed of propagation. You don't see something until the light has travelled to you, just as you don't hear something until the sound has travelled to you. This is not what's going on here! A look at the

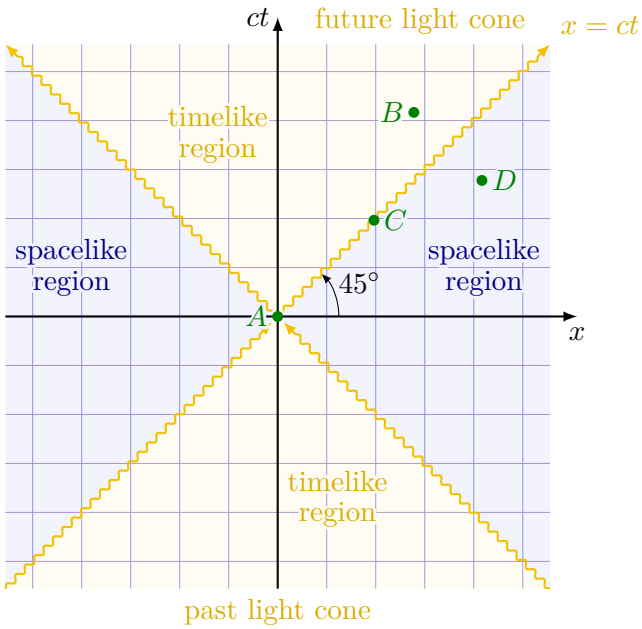

Fig. 2.6: Lightcone structure around the event A to illustrate the causal structure of Minkowski space. Events B and A are separated by a timelike interval, and B lies in the forward lightcone of A. The events could be causally connected. Events C and A are separated by a null (or lightlike) interval and could be connected by a light signal, at a 45° angle on the diagram. Events D and A are separated by a spacelike interval and cannot be causally connected.

spacetime diagram in Figure 48 shows that we've already taken this into account when deciding whether two events occur simultaneously. The lack of simultaneity between moving observers is a much deeper issue, not due to the finiteness of the speed of light but rather due to the constancy of the speed of light.

The confusion about the time of flight of the signal is sometimes compounded by the common use of the word observer to mean "inertial frame". This brings to mind some guy sitting at the origin, surveying all around him. Instead, you should think of the observer more as a Big Brother figure: a sea of clocks and rulers throughout the inertial frame which can faithfully record and store the position and time of any event, to be studied at some time in the future.

2.3 Length Contraction and Time Dilation

2.3.1 Time Dilation

We'll now turn to one of the more dramatic results of special relativity. Consider a clock sitting stationary in the frame S' which ticks at intervals of T'. This means that the tick events in frame S' occur at $(ct'_1, 0)$ then $(ct'_1 + cT', 0)$ and so on. What are the intervals between ticks in frame S?

We can answer immediately from the Lorentz transformations (2.14). Inverting this

gives

$$t = \gamma\left(t' + \frac{vx'}{c^2}\right). \tag{2.33}$$

The clock sits at $x' = 0$, so we immediately learn that in frame S, the interval between ticks is

$$T = \gamma T' \tag{2.34}$$

This means that the gap between ticks is longer in the stationary frame. A moving clock runs more slowly. But the same argument holds for any process, be it clocks, elementary particles or human hearts. The correct interpretation is that time itself runs more slowly in moving frames. This is *time dilation*.[2]

Note that, throughout this course, we shall consider only *ideal clocks* – clocks that are unaffected by acceleration – for example, the half-life of a decaying particle.

2.3.2 Length Contraction

We've seen that moving clocks run slow. We will now show that moving rods are shortened. Consider a rod of length L_0 sitting stationary in the frame S', as illustrated on the right in Fig. 2.7. What is its length in frame S?

To begin, we should state more carefully something which seems obvious: when we say that a rod has length L_0, it means that the distance between the two end points at equal times is L_0. So, drawing the axes for the frame S' in dark red with light red lines of simutaneity, the situation looks like the right diagram in Fig. 2.7. The two, simultaneous, end points in S' are A and B′. Their coordinates in S' are $(ct', x') = (0, 0)$ and $(0, L_0)$ respectively.

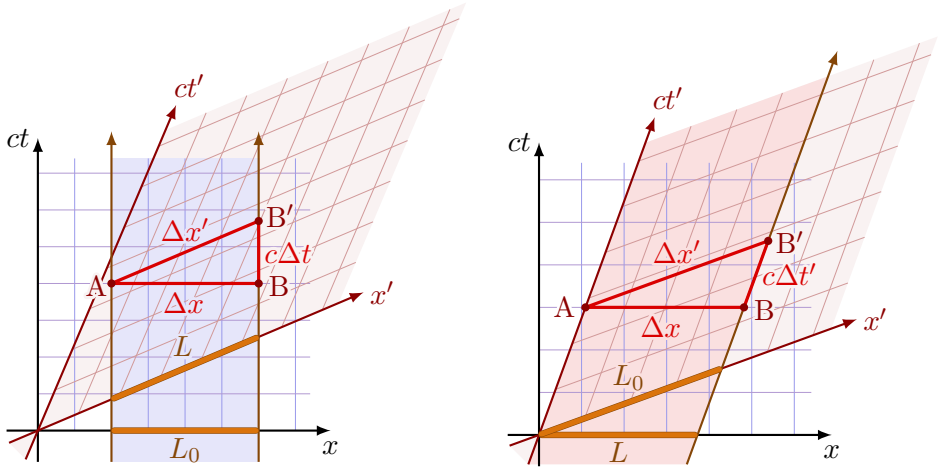

Fig. 2.7: Length contraction of a rod that is at rest (left)/ moving (right) in frame S. The rod is shorter in the boosted frame S' than in its rest frame by a factor of γ. This phenomenon is known as Lorentz contraction.

[2] I'm not sure that this is a helpful description: what exactly is dilated?? It is better, as always in Special Relativity, to fix on a precise space-time description of the situation: what events we are considering and in which frame.

The observer in frame S' initially places the rod along the x'-axis. As time evolves, the world lines traced out by the ends of the rod trace out the two parallel dark brown solid lines shown. During the time interval Δt the observer measures the rod length as $\Delta x'$ as indicate in red. The observer at rest with respect to reference frame S measures the ends of the rod at a fixed time and finds that the length of the moving rod is $\Delta x = L$. From the Lorentz transformations with $\Delta t = 0$, we have

$$\Delta x' = \gamma \Delta x \quad \Longrightarrow \quad L_0 = \gamma L. \tag{2.35}$$

This is the length contraction equation.

In a similar manner, a rod at rest with respect to frame S is depicted in the left of Fig. 2.7. The rod is initially aligned with the x-axis. The world lines for the ends of the rod are shown as two dark brown parallel lines. The observer in frame S' measures the length at a fixed time, so $\Delta t' = 0$. The Lorentz transformation gives

$$\Delta x = \gamma \Delta x' \tag{2.36}$$

This is telling us that the length L measured in frame S is

$$L = \frac{L_0}{\gamma}. \tag{2.37}$$

It is shorter than the length of the rod in its rest frame by a factor of γ. This phenomenon is known as *Lorentz contraction*. The fact that this is symmetric for both cases discussed should not be a surprise. The rod suffers no contraction in the y- and z-directions (i.e., perpendicular to its velocity)

It follows that the volume V' of a moving object is related to proper volume V by $V = V'/\gamma$. Since the total number of objects in a system is Lorentz invariant, number densities thus transform from the rest frame as $n = n'/\gamma$.

But hold on! If we look at the lengths of the rod, marked in the two diagrams in Fig. 2.7, it seems that we have got it the wrong way round: in both cases the Euclidean lengths of the $\Delta x'$ sides of each triangle appear longer than side Δx between points A and B. How can this be? This is a trap: lengths in space-time diagrams are not like lengths in the more familiar x-y plane and we must rely on our calculations.[3] Remember, these increments in spacetime are given by the invariant Δs in both systems.

2.3.3 The Ladder-and-Barn Non-Paradox

Take a ladder of length $2L$ and try to put it in a barn of length L. If you run fast enough, can you squeeze it? Here are two arguments, each giving the opposite conclusion

- From the perspective of the barn, the ladder contracts to a length $2L/\gamma$. This shows that it can happily fit inside as long as you run fast enough, with $\gamma \geq 2$.

[3] Lengths are shorter the closer the inclination to the 45° of the null cone. This is because instead of the Euclidean norm (Pythagoras), one must use the norm $|(ct, x)| = (c^2t^2 - x^2)^{1/2}$.

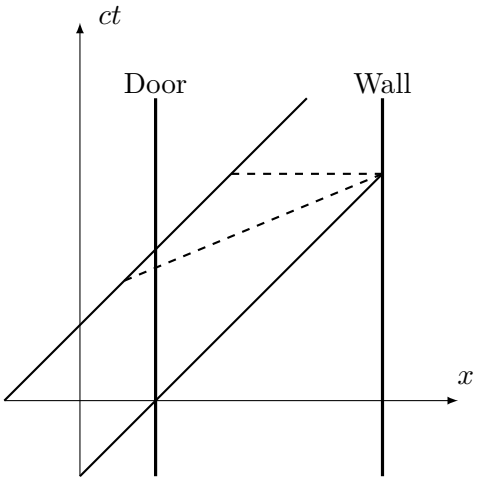

Fig. 2.8: The Ladder-and-Barn Non-Paradox: Regarded from the point of view of a space-time diagram, the paradox dissolves. One consequence of time not being invariant under Lorentz transformations is that the ladder 'fits in' the barn in one frame but does not 'fit in' in another.

- From the perspective of the ladder, the barn has contracted to length L/γ. This means there's no way you're going to get the ladder inside the barn. Running faster will only make things worse.

What's going on? The answer stems, as is often the case with apparent paradoxes in relativity, from loose use of language. As usual, to reconcile these two points of view we need to think more carefully about the question we're asking. What does it mean to 'fit a ladder inside a barn'?

In this case, it is the use of the word 'fit'; what does it mean to say the ladder 'fits' exactly into the barn? Clearly, we mean that the two events:

- front end of ladder hits back of barn;
- back end of ladder goes through the door.

are simultaneous. Any observer will agree that we've achieved this if the back end gets in the door before the front end hits the far wall. But we know that simultaneity of events is not fixed, as observers in different frames do not agree on simultaneity, so 'fit into' is a frame-dependent concept: we should not expect observers in different frames to agree so there is no paradox to account for. The two statements are true and compatible and that is really the end of the story. However, we can investigate further.

The spacetime diagram (see Fig. 2.8) in the frame of the barn is drawn in the figure with $\gamma > 2$. We see that, from the barn's perspective, both back and front ends of the ladder are happily inside the barn at the same time. We've also drawn the line of simultaneity for the ladder's frame. This shows that when the front of the ladder hits the far wall, the back end of the ladder has not yet got in the door. Is the ladder in the barn? Well, it all depends who you ask.

2.3.4 The Twins Non-Paradox

Twins Alice and Bob synchronise watches in an inertial frame and then Bob sets off at speed $\sqrt{3}c/2$, which corresponds to $\gamma = 2$. When Bob has been travelling for a time T according to Alice, he reaches Proxima Centauri[4] and turns round by means of accelerations that are very large in his frame and goes back to Alice at the same speed. Since Bob is in a moving frame, relative to Alice, his time runs slower by a factor of γ than Alice's, so he will only have aged by $2T \times \frac{1}{2}$ on the two legs of the journey. Thus when they meet up again, Alice has aged by $2T$ but Bob has aged only by T. *This is not the paradox: it is just a fact of life.*[5]

The difficulty some people have with Alice and Bob is the apparent symmetry: surely exactly the same argument could be made, from Bob's point of view, to show that Alice would be the younger when they met again? But the same argument *cannot* be made for Bob because the situation is not symmetric: Alice's frame is inertial, whereas Bob has to accelerate to turn round: while he is accelerating, his frame is not inertial.

BUT, some people might say, suppose we just consider the event of Bob's arrival at Proxima Centauri, so as not to worry about acceleration. Now the situation is symmetric. Surely from Alice's point of view, when Bob arrives he will have aged half as much as Alice, and from Bob's point of view, when he arrives, Alice will have aged half as much as Bob? The answer to this is a simple 'yes'. Surely, they would then say, this doesn't make sense? But it does, as long as you are careful about the word 'when'.

In the diagram in Fig. 2.9, Alice's world line is the ct (containing points A, B and C) axis and Bob's world line is the line containing A and P. P represents the event 'Bob arrives at Proxima Centauri'.

The line CP is a line of simultaneity in Alice's frame and C is the event 'Alice is at this point in space-time *when* – according to Alice – Bob arrives at Proxima Centauri'; the first use of the word 'when'.

The line BP is a line of simultaneity in Bob's frame and B is the event 'Alice is at this point in space-time when – according to Bob – he arrives Proxima Centauri'; the second use of the word 'when'. The two 'whens' don't mean the same thing, since one is a 'when' in Alice's frame the other is a 'when' in Bob's frame.

We can do the calculation. Let us assume for simplicity that Bob sets off the moment he is born. The event C has coordinates $(cT, 0)$ in Alice's frame, and the event P has coordinates (cT, vT). In Bob's frame, the elapsed time T' is given by the Lorentz transformation:

$$T' = \gamma(T - v^2T/c^2) = T/\gamma = \tfrac{1}{2}T. \tag{2.38}$$

[4] The closest star to the Sun: about 4.2 light years away

[5] In 1971, Hafele and Keating packed four atomic (caesium) clocks into suitcases and went round the Earth, in different directions, on commercial flights. When they returned, they found that the clocks were slightly behind a clock remaining at the first airport. The result was somewhat inconclusive. The calculations are complicated by the fact that the rate of the clocks is also affected by the gravitational field: clocks run slower in stronger fields, and in fact the two affects balance at $3R/2$ (where R is the radius of the Earth). Thus the heights of the aircraft had to be taken into account as well as their speeds, and it turns out that the two effects are of comparable magnitude, namely of the order of 100 nanoseconds.

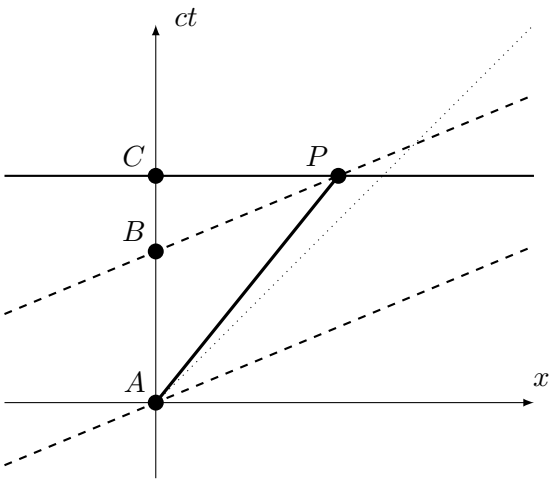

Fig. 2.9: The Twins Non-Paradox: Alice's world line is the ct (containing points A, B and C) axis and Bob's world line is the line containing A and P. P represents the event 'Bob arrives at Proxima Centauri'.

This is just the usual time dilation calculation. Thus Bob and Alice agree that Bob's age at Proxima Centauri is $\frac{1}{2}T$. In Alice's frame, Bob has aged half as much as Alice.

We now work out the coordinates of the event B, sticking with Alice's frame. The line of simultaneity, BP has equation $t' = \frac{1}{2}T$, i.e. (using a Lorentz transformation)

$$\gamma(t + vx/c^2) = \tfrac{1}{2}T, \tag{2.39}$$

so the point B, for which $x = 0$, has coordinates $(\frac{1}{2}cT/\gamma, 0)$, i.e. $(\frac{1}{4}cT, 0)$. Alice's age when, according to Bob, he arrives at Proxima Centauri is therefore $\frac{1}{4}T$, which is indeed half of Bob's age. So no paradox there either.

BUT, some other people might say, suppose Bob does not turn round but just synchronises his watch at Proxima Centauri with that of another astronaut, Bob$'$, who is going at speed v in the opposite direction (like two trains passing at a station). Each leg of the journey is then symmetric, so why should Alice age faster or slower Bob and Bob$'$ during their legs of the journey? There's no mystery here, either: the situation is indeed symmetric and Alice does indeed age by the same amount as Bob+Bob$'$. But at the synchronisation event, Bob and Bob$'$ do not agree on Alice's age, because in their different frames the synchronisation event is simultaneous with different times in Alice's life.

Let us see how this looks in a space-time diagram with figures 2.9–2.10.

As before, Bob ages by $\frac{1}{2}T$ on the outward journey to Proxima Centauri. By symmetry Bob' ages by $\frac{1}{2}T$ on the inward journey from Proxima Centauri.

However, according to Bob's idea of time, the clock synchronisation occurs when Alice is at B, and according to Bob$'$'s it occurs when Alice is at D. Thus Bob$'$'s clock will read time T when he meets Alice and Alice's clock will read $2T$. But the time Alice spends between B and D is accounted for by Bob in his journey *after* Proxima Centauri and by Bob$'$ in his journey *before* reaching Proxima Centauri, so the two Bobs would say that, while they were travelling between Earth and Proxima Centauri, Alice travelled from A

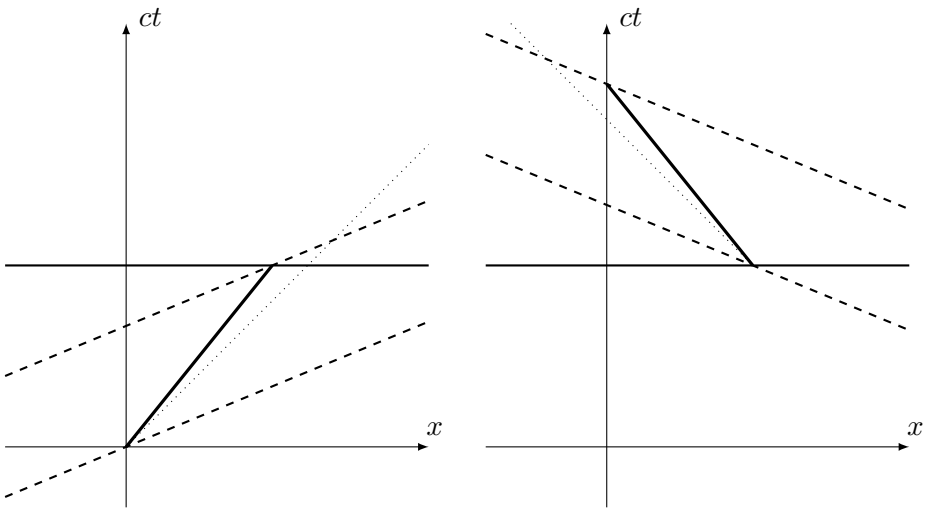

Fig. 2.10: Left: The outward journey. The heavy line is Bob's world line. The dotted line through the origin is the light cone. The dashed lines are the lines of simultaneity in Bob's frame. Right: The return journey. The heavy line is the world line of Bob′. The dotted line through the turn-round event is the light cone. The dashed lines are the lines of simultaneity in the frame of Bob′.

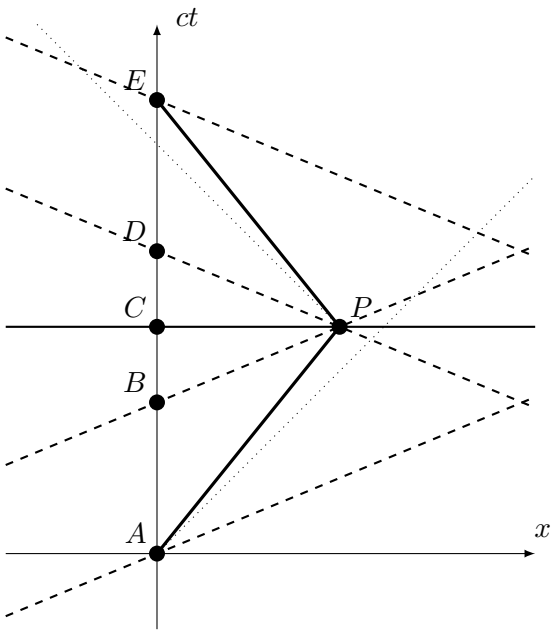

Fig. 2.11: The superposition of the previous two space-time diagrams in Fig. 2.10, representing together both the outward journey of Bob and the return journey of Bob′.

to B and then from D to E, taking on her clock a total time T – the same as the journey time of the two Bobs.

Finally, we see that if, instead of meeting Bob′, Bob turns round at Proxima Centauri, Alice ages rapidly (according to Bob) from B to D while he is changing direction.

2.4 Paths in spacetime

2.4.1 Minkowski Spacetime Line Element

As we first introduced in Eq. (2.16), the invariant interval $\Delta s^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$ corresponds to the "distance" in spacetime between two events A and B measured along the straight line connecting them.

For a general, arbitrary path through spacetime, we must express the intrinsic geometry of Minkowski spacetime in infinitesimal form using the invariant Minkowski line element for infinitesimally-separated events:

$$\boxed{\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2\,.} \tag{2.40}$$

For a path connecting events A and B, the invariant "distance" along the path is given by the line integral

$$\Delta s = \int_A^B \mathrm{d}s\,. \tag{2.41}$$

The advantage of the invariant interval is that it is something all observers agree upon. The interval is of course path-*dependent*, but a Lorentz invariant.

2.4.2 Particle Worldlines and Proper Time

A particle describes a *worldline* in spacetime. For a massive particle passing through an event A, the particle's worldline must be inside the lightcone through A and each infinitesimal step must lie *within* the lightcone at each point. For a photon or other massless particle, the worldline will be *tangent* to the lightcone.

The fact that the concept of time is frame dependent can be rather unsettling. It would be good to have some quantity that corresponds to time but does not vary at the whim of the observer. Such a quantity exists and is called *proper time*.

We can write the spacetime path as $x(t)$, $y(t)$, $z(t)$ or, parametrically, as $t(\lambda)$, $x(\lambda)$, $y(\lambda)$, $z(\lambda)$ for some parameter λ. The most natural parameter for a massive particle is *proper time* – the time measured by an ideal clock carried by the observer comoving with the particle (i.e. particle is at rest in the observer's frame).

The increment in proper time, $\mathrm{d}\tau$, is just the increment in time in the *instantaneous rest frame* of the particle, where $\mathrm{d}x' = \mathrm{d}y' = \mathrm{d}z' = 0$. It follows that $c^2\,\mathrm{d}\tau^2 = \mathrm{d}s^2$ and so,

for two infinitesimally close events on the particle's worldline separated by $\mathrm{d}t$, $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}z$ in some inertial frame,

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \tag{2.42}$$

$$\implies \mathrm{d}\tau = \mathrm{d}t\,/\gamma, \tag{2.43}$$

i.e.

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \gamma. \tag{2.44}$$

Note that here in (2.44), γ is not linked to the velocity between two frames explicitly, though of course it is implicitly related to the velocity between the rest frame of the observer and the lab frame.

The total time that elapses on the world-line of an observer moving with (not necessarily constant) velocity in a frame S is given by

$$\Delta\tau = \int \mathrm{d}\tau = \int \gamma^{-1}\,\mathrm{d}t\,; \tag{2.45}$$

this is the observer's actual time (clock or biological).

We can use proper time to derive the velocity addition formula (2.57) for an observer moving with non-constant velocity. We parameterise the observer's world line by τ:

$$x = x(\tau), \quad t = t(\tau), \qquad \text{in } S \tag{2.46}$$

$$x' = x'(\tau), \quad t' = t'(\tau), \qquad \text{in } S' \tag{2.47}$$

and

$$u = \frac{\mathrm{d}x}{\mathrm{d}\tau}\Big/\frac{\mathrm{d}t}{\mathrm{d}\tau}, \qquad u' = \frac{\mathrm{d}x'}{\mathrm{d}\tau}\Big/\frac{\mathrm{d}t'}{\mathrm{d}\tau}. \tag{2.48}$$

We can differentiate the Lorentz transformation (2.12) and (2.5) to obtain

$$\frac{\mathrm{d}x'}{\mathrm{d}\tau} = \gamma\left(\frac{\mathrm{d}x}{\mathrm{d}\tau} - v\frac{\mathrm{d}t}{\mathrm{d}\tau}\right) = \gamma(u-v)\frac{\mathrm{d}t}{\mathrm{d}\tau} \tag{2.49}$$

$$\frac{\mathrm{d}t'}{\mathrm{d}\tau} = \gamma\left(\frac{\mathrm{d}t}{\mathrm{d}\tau} - \frac{v}{c^2}\frac{\mathrm{d}x}{\mathrm{d}\tau}\right) = \gamma(1-\frac{uv}{c^2})\frac{\mathrm{d}t}{\mathrm{d}\tau} \tag{2.50}$$

and dividing these expressions gives

$$u' = \frac{u-v}{1-uv/c^2}. \tag{2.51}$$

2.4.3 Doppler Effect

Consider an observer $\mathcal{E}$ who moves at speed v along the x-axis of an intertial frame S in which an observer $\mathcal{O}$ is at rest at position x_o (see Fig. 2.12). Let successive wavecrests be emitted by $\mathcal{E}$ at events A and B, which are separated by proper time $\Delta\tau_{AB}$; this is the *proper period* of the source.

The relation between $\Delta\tau_{AB}$ and the time Δt_e between the emission events in S is

$$\Delta\tau_{AB} = \Delta t_e/\gamma. \tag{2.52}$$

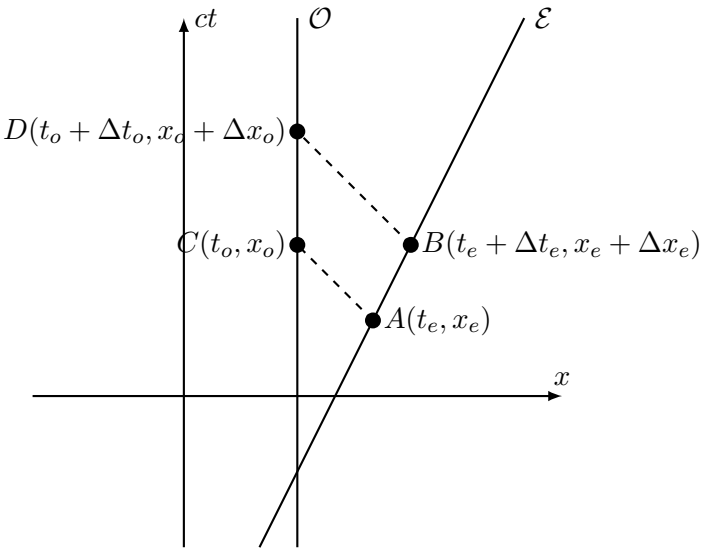

Fig. 2.12: Spacetime diagram of the Doppler effect. An observer $\mathcal{E}$ moves at speed v along the x-axis of an inertial frame S in which an observer $\mathcal{O}$ is at rest at position x_o. A wavecrest is emitted by $\mathcal{E}$ at the event A with coordinates (t_e, x_e) in S and is received by $\mathcal{O}$ at the event C with coordinates (t_o, x_o). A second crest is emitted by $\mathcal{E}$ at the event B, which occurs at a time Δt_e later than A in S, and is received by $\mathcal{O}$ at the event D a time Δt_o later than C.

The wavecrests are received by $\mathcal{O}$ at the events C and D, which are separated by time Δt_o in S; since $\mathcal{O}$ is at rest in S, the proper time between C and D is $\Delta\tau_{CD} = \Delta\tau_o$. In time Δt_e, the source $\mathcal{E}$ moves a distance $\Delta x_e = v\Delta t_e$ along the x-axis in S, and the second wavecrest has to travel Δx_e further than the first to be received by $\mathcal{O}$ at x_o. It follows that

$$\Delta t_o = \left(1 + \frac{v}{c}\right)\Delta t_e, \tag{2.53}$$

so that the ratio $\Delta\tau_{AB}/\Delta\tau_{CD}$ of proper times is

$$\frac{\Delta\tau_{AB}}{\Delta\tau_{CD}} = \frac{\sqrt{1 - \frac{v^2}{c^2}}\Delta t_e}{(1 + \frac{v}{c})\Delta t_e} = \sqrt{\frac{1 - \frac{v}{c}}{1 + \frac{v}{c}}}. \tag{2.54}$$

This ratio is also the ratio of the received frequency, as measured by $\mathcal{O}$, to the proper frequency (i.e., the frequency in the rest-frame of the source $\mathcal{E}$).

2.4.4 Addition of Velocities

A particle moves with constant velocity u' in frame S' which, in turn, moves with constant velocity v with respect to frame S. What is the velocity u of the particle as seen in S?

The Newtonian answer is just $u = u' + v$. But we know that this can't be correct because it doesn't give the right answer when $u' = c$. So what is the right answer?

The worldline of the particle in S' is

$$x' = ut'. \tag{2.55}$$

So the velocity of the particle in frame S is given by

$$u = \frac{x}{t} = \frac{\gamma(x' + vt')}{\gamma(t' + vx'/c^2)}, \tag{2.56}$$

which follows from the Lorentz transformations (2.14). (Actually, we've used the inverse Lorentz transformations since we want S coordinates in terms of S' coordinates, but these differ only changing v to $-v$). Substituting (2.55) into the expression above, and performing a little algebra, gives us the result we want:

$$u = \frac{u' + v}{1 + u'v/c^2}. \tag{2.57}$$

Note that when $u' = c$, this gives us $u = c$ as expected. We can also show that if $|u'| < c$ and $|v| < c$ then we necessarily have $-c < u < c$. The proof is simple algebra, if a little fiddly

$$c - u = c - \frac{u' + v}{1 + u'v/c^2} = \frac{c(c - u')(c - v)}{c^2 + u'v} > 0, \tag{2.58}$$

where the last equality follows because, by our initial assumptions, each factor in the final expression is positive. An identical calculation will show you that $-c < u$ as well. We learn that if a particle is travelling slower than the speed of light in one inertial frame, it will also be travelling slower than light in all others.

It follows that the velocity components in S are given by

$$\begin{aligned} u_x &= \frac{\mathrm{d}x'}{\mathrm{d}t'} = \frac{u'_x + v}{1 + u'_x v/c^2}, \\ u_y &= \frac{\mathrm{d}y'}{\mathrm{d}t'} = \frac{u'_y}{\gamma(1 + u'_x v/c^2)}, \\ u_z &= \frac{\mathrm{d}z'}{\mathrm{d}t'} = \frac{u'_z}{\gamma(1 + u'_x v/c^2)}. \end{aligned} \tag{2.59}$$

The appropriate velocity transformations from frame S to frame S' are obtained by replacing v with $-v$ (and switching u'_i and u_i).

These results replace the "common-sense" addition of velocities in Newtonian mechanics; they reduce to the Newtonian results in the limit $v/c \to 0$ (or equivalently as $c \to \infty$).

Now consider three inertial frames S, S' and S'', where S' and S are related by a standard boost along the x-direction with speed v, and S'' and S' are related by a standard boost along the x'-direction with speed u'.

If we write the velocities u, u', and v in terms of rapidities:

$$\frac{u}{c} = \tanh\psi_u, \qquad \frac{u'}{c} = \tanh\psi'_u, \qquad \frac{v}{c} = \tanh\psi_v. \tag{2.60}$$

We can then find the composition of the two Lorentz transforms in terms of the rapidities by substitution into (2.51) to give

$$\tanh\psi'_u = \frac{\tanh\psi_u - \tanh\psi_v}{1 - \tanh\psi_v \tanh\psi_u} = \tanh(\psi_u - \psi_v). \tag{2.61}$$

so an alternative form of the transformation law is

$$\psi'_u = \psi_u - \psi_v \tag{2.62}$$

i.e.

$$\tanh^{-1}\left(\frac{u'}{c}\right) = \tanh^{-1}\left(\frac{u}{c}\right) + \tanh^{-1}\left(\frac{v}{c}\right). \tag{2.63}$$

We see that the composition of two colinear boosts is another boost along the same direction and the rapidities add (like adding angles for rotations about a common axis).

2.4.4.1 3D Velocity transformation

In Subsection 2.2.3.1, we derived a general Lorentz transform (2.31), for an inertial frame S' relating to the frame S by a boost of $\vec{v}$ whose components in S are (v_x, v_y, v_z).

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma\beta_x & -\gamma\beta_y & -\gamma\beta_z \\ -\gamma\beta_x & 1+\alpha\beta_x^2 & \alpha\beta_x\beta_y & \alpha\beta_x\beta_z \\ -\gamma\beta_y & \alpha\beta_y\beta_x & 1+\alpha\beta_y^2 & \alpha\beta_y\beta_z \\ -\gamma\beta_z & \alpha\beta_z\beta_x & \alpha\beta_z\beta_y & 1+\alpha\beta_z^2 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}. \tag{2.64}$$

where $\vec{\beta} = \vec{v}/c$, $\gamma = (1 - |\vec{\beta}|^2)^{-1/2}$ and $\alpha = (\gamma - 1)/|\vec{\beta}|^2$.

From this, we can get a general velocity transform between frames by considering the *differential* of the above,

$$\begin{pmatrix} c\,\mathrm{d}t' \\ \mathrm{d}x' \\ \mathrm{d}y' \\ \mathrm{d}z' \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma\beta_x & -\gamma\beta_y & -\gamma\beta_z \\ -\gamma\beta_x & 1+\alpha\beta_x^2 & \alpha\beta_x\beta_y & \alpha\beta_x\beta_z \\ -\gamma\beta_y & \alpha\beta_y\beta_x & 1+\alpha\beta_y^2 & \alpha\beta_y\beta_z \\ -\gamma\beta_z & \alpha\beta_z\beta_x & \alpha\beta_z\beta_y & 1+\alpha\beta_z^2 \end{pmatrix} \begin{pmatrix} c\,\mathrm{d}t \\ \mathrm{d}x \\ \mathrm{d}y \\ \mathrm{d}z \end{pmatrix}. \tag{2.65}$$

Note, that Lorentz transformations link the coordinates in two reference frames. Velocity in S is $\mathrm{d}\vec{u}/\mathrm{d}t$, while that in S' will be $\mathrm{d}\vec{u}'/\mathrm{d}t'$ The derivatives like $\mathrm{d}\vec{u}'/\mathrm{d}t$ or $\mathrm{d}\vec{u}/\mathrm{d}t'$ do not make any sense because they consist of values from different references frames. So to derive the velocity formula, you have to find first the expressions for $\mathrm{d}x$ and $\mathrm{d}t$, and then divide them one on the other. In practise, the resulting equations will simplify considerably when some velocity components vanish, for example.

2.4.5 Aberration and the Headlight Effect

The change in direction of travel of waves (especially light waves) when the same wave is observed in one of two different inertial frames is called *aberration*. The new name should not be taken to imply there is anything new here, however, beyond what we have already discussed. It is just an example of the change in direction of a 4-vector. The name arose historically because changes in the direction of rays in optics were referred to as "aberration".

Consider a light source moving with velocity v in frame S. In frame S', which is moving with velocity v away from S, then light source is thus at rest.

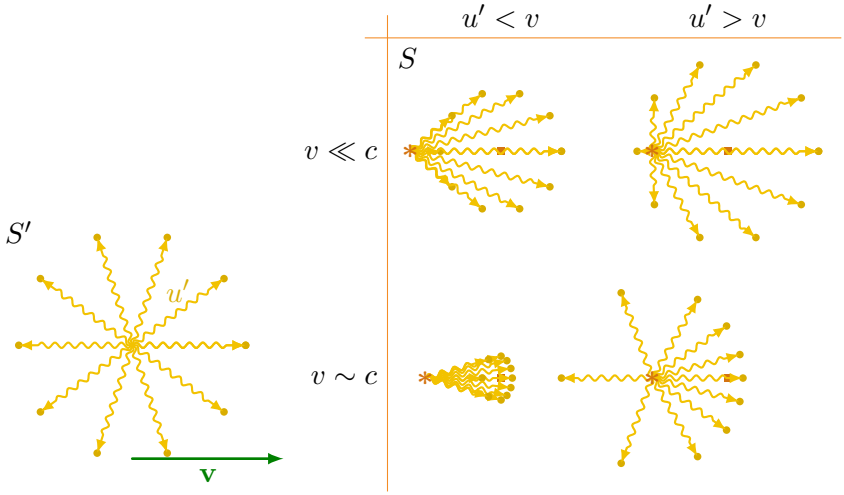

Fig. 2.13: Transformation of velocities and the headlight effect. An isotropic explosion in frame S' produces particles all moving at speed u' in S', and a fragment is left at the centre of the explosion (left diagram). The fragment and frame S' move to the right at speed v relative to frame S. The right four diagrams show the situation in frame S. The $*$ shows the location of the explosion event. The square shows shows the present position of the central fragment; the circles show positions of the particles; the arrows show the velocities of the particles. The left diagrams show examples with $u' < v$, the right with $u' > v$. The top two diagrams show the case $u', v \ll c$. Here the particles lie on a circle centred at the fragment, as in classical physics. The bottom diagrams show examples with $v \sim c$, thus bringing out the difference between the relativistic and the classical predictions. The lower right shows $u' = c$: headlight effect for photons. The photons lie on a circle centred at the position of the explosion (not the fragment) but more of them move forward than backward.

Consider a photon travelling at angle θ' as measured in S'. Clearly the x-component of the velocity of the photon is $u'_x = c\cos\theta'$. In S, the x-component is $u_x = c\cos\theta$.

From the velocity transformations (2.59) it follows

$$\begin{aligned} u_x &= \frac{u'_x + v}{1 + u'_x v/c^2} \\ \Longrightarrow \quad c\cos\theta &= \frac{c\cos\theta' + v}{1 + vc\cos\theta'/c^2} \\ \Longrightarrow \quad \cos\theta &= \frac{\cos\theta' + v/c}{1 + \cos\theta' v/c} \\ \cos\theta &= \frac{\cos\theta' + \beta}{1 + \beta\cos\theta'}, \end{aligned} \tag{2.66}$$

the *angle transformation*.

A consequence of this result is the *headlight effect*: light from a moving source is observed to not be emitted in all directions. Rather, it appears to be emitted in a beam (like from a car headlight).

An important quantity is a measure of how much light is emitted into any given small range of directions. This is done by imagining a sphere around the light source, and asking how much light falls onto a given region of the sphere.

Suppose N photons are emitted isotropically in frame S'. Then the number emitted

into a ring at angle θ' with angular width $\mathrm{d}\theta'$ is equal to N multiplied by the surface area of the ring divided by the surface area of a sphere:

$$\mathrm{d}N = N\frac{(2\pi r\sin\theta')(r\,\mathrm{d}\theta')}{4\pi r^2}. \tag{2.67}$$

Here r is the radius of the sphere, so $r\sin\theta'$ is the radius of the ring, and we used the fact that the surface area of such a narrow ring is simply its circumference multiplied by its width $r\,\mathrm{d}\theta'$. Hence

$$\frac{\mathrm{d}N}{\mathrm{d}\theta'} = \frac{1}{2}\sin\theta'. \tag{2.68}$$

We would like to find the corresponding quantity $\mathrm{d}N/\mathrm{d}\theta$ representing the number of photon velocities per unit range of angle in the other reference frame. This is obtained from $\mathrm{d}N/\mathrm{d}\theta$. We invert (2.66) to obtain an expression for $\cos\theta'$ in terms of $\cos\theta$, and then differentiate, which gives

$$\sin\theta'\frac{\mathrm{d}\theta'}{\mathrm{d}\theta} = \sin\theta\frac{1}{\gamma^2(1-\beta\cos\theta)^2}, \tag{2.69}$$

and therefore

$$\frac{\mathrm{d}N}{\mathrm{d}\theta} = \frac{\mathrm{d}N}{\mathrm{d}\theta'}\frac{\mathrm{d}\theta'}{\mathrm{d}\theta} = \frac{1}{2}\sin\theta\frac{1}{\gamma^2(1-\beta\cos\theta)^2}. \tag{2.70}$$

The solid angle subtended by the ring is $\mathrm{d}\Omega = 2\pi\sin\theta\,\mathrm{d}\theta$ in S and $\mathrm{d}\Omega' = 2\pi\sin\theta'\,\mathrm{d}\theta'$ in S'. The conclusion for emission per unit range of solid angle is

$$\frac{\mathrm{d}N}{\mathrm{d}\Omega'} = \frac{N}{4\pi}, \quad \frac{\mathrm{d}N}{\mathrm{d}\Omega} = \frac{N}{4\pi}\frac{1}{\gamma^2(1-\beta\cos\theta)^2}. \tag{2.71}$$

Note that N, the total number of emitted particles, must be the same in both reference frames. The equation for $\frac{\mathrm{d}N}{\mathrm{d}\Omega}$ gives the enhancement (or reduction) factor for emission in forward (or backward) directions. For example, the enhancement factor for emission into a small solid angle in the directly forward direction (at $\theta = \theta' = 0$) is $(1-\beta^2)/(1-\beta)^2 = (1+\beta)/(1-\beta)$.

2.5 Acceleration in Special Relativity

It is often said, erroneously, that Special Relativity cannot deal with acceleration because it deals only with inertial frames, and that therefore acceleration must be the preserve of General Relativity. We must, of course, only allow transformations between inertial frames; the frames must not accelerate, but the observers in the frame can move as the please. Special Relativity can deal with anything kinematic but General Relativity is required when gravitational forces are present. Acceleration is not Lorentz invariant, but *absolute*, in that observers in different reference frames will not agree on the components of the acceleration, but all will agree that an object is accelerating.

As an example of non-uniform motion, we consider an observer who is moving with constant acceleration.

The first step is to define what we mean by "constant acceleration" which is certainly a frame-dependent concept. The most common situation is that of an observer in a

rocket experiencing a constant 'G-force' due to the rocket thrust. This corresponds to the acceleration measured in the instantaneous (inertial) rest frame of the rocket being constant (acceleration having the usual definition of $\mathrm{d}v\,/\,\mathrm{d}t$), so we take this to be our definition.

We must determine the way that acceleration transforms under Lorentz transformations. We can do this in a number of ways. We will here start with the velocity transformation law (2.57) for an observer with world line given in S by $(ct(\tau), x(\tau))$ and in S' by $(ct'(\tau), x'(\tau))$. Forgetting the acceleration problem for the moment, we assume that these frames have a constant relative velocity v.

The velocities u and u' in the two frames are related by

$$u' = \frac{u - v}{1 - uv/c^2} \equiv \frac{(c^2/v)(1 - v^2/c^2)}{1 - uv/c^2} - \frac{c^2}{v}. \tag{2.72}$$

(the equivalent form is just a bit of algebra to obtain a useful expression). Differentiating this with respect to τ gives

$$\frac{\mathrm{d}u'}{\mathrm{d}\tau} = \frac{1 - v^2/c^2}{(1 - uv/c^2)^2}\frac{\mathrm{d}u}{\mathrm{d}\tau}. \tag{2.73}$$

The acceleration, a, in S is by definition $\mathrm{d}u\,/\,\mathrm{d}t$ and similarly for S' so

$$\begin{aligned}
a' &= \frac{\mathrm{d}u'}{\mathrm{d}t'} \\
&= \frac{\mathrm{d}u'}{\mathrm{d}\tau}\Big/\frac{\mathrm{d}t'}{\mathrm{d}\tau} \\
&= \frac{1 - v^2/c^2}{(1 - uv/c^2)^2}\frac{\mathrm{d}u}{\mathrm{d}\tau}\Big/\frac{\mathrm{d}t'}{\mathrm{d}\tau} && \text{(using (2.73))} \\
&= \frac{1 - v^2/c^2}{(1 - uv/c^2)^2}\frac{\mathrm{d}u}{\mathrm{d}\tau}\Big/\gamma(1 - uv/c^2)\frac{\mathrm{d}t}{\mathrm{d}\tau} && \text{(using (2.49))} \\
&= \frac{(1 - v^2/c^2)^{3/2}}{(1 - uv/c^2)^3}a \\
&= \frac{1}{\gamma^3(1 - uv/c^2)^3}a.
\end{aligned} \tag{2.74}$$

As mentioned above there are other ways of obtaining this result; for example, more elegantly using four-vectors.

To find all the corresponding components in S', connected to S by a standard Lorentz boost, we consider the differentials of the velocity transformations (2.59)

$$\begin{aligned}
\mathrm{d}u'_x &= \frac{\mathrm{d}u_x}{\gamma_v^2(1 - u_x v/c^2)^2} \\
\mathrm{d}u'_y &= \frac{\mathrm{d}u_y}{\gamma_v(1 - u_x v/c^2)} + \frac{u_y v\,\mathrm{d}u_x}{c^2\gamma_v(1 - u_x v/c^2)^2} \\
\mathrm{d}u'_z &= \frac{\mathrm{d}u_z}{\gamma_v(1 - u_x v/c^2)} + \frac{u_z v\,\mathrm{d}u_x}{c^2\gamma_v(1 - u_x v/c^2)^2},
\end{aligned} \tag{2.75}$$

and

$$\mathrm{d}t' = \gamma(\mathrm{d}t - v\,\mathrm{d}x\,/c^2) = \gamma_v(1 - u_x v/c^2)\,\mathrm{d}t\,. \tag{2.76}$$

The acceleration thus transforms as

$$\begin{aligned} a'_x &= \frac{\mathrm{d}u'_x}{\mathrm{d}t'} = \frac{1}{\gamma_v^3(1-u_x v/c^2)^3}a_x \\ a'_y &= \frac{\mathrm{d}u'_y}{\mathrm{d}t'} = \frac{1}{\gamma_v^2(1-u_x v/c^2)^2}a_y + \frac{u_y v}{c^2\gamma_v^2(1-u_x v/c^2)^3}a_x \\ a'_z &= \frac{\mathrm{d}u'_z}{\mathrm{d}t'} = \frac{1}{\gamma_v^2(1-u_x v/c^2)^2}a_z + \frac{u_z v}{c^2\gamma_v^2(1-u_x v/c^2)^3}a_x. \end{aligned} \tag{2.77}$$

We see that acceleration is not invariant in special relativity but is, however, an absolute quantity in that all observers agree whether a particle is accelerating or not.

In the situation we have in mind, S' is the instantaneous rest frame of the accelerating observer, so that $u' = 0$ and $u = v$, and the acceleration a' in this frame is constant (i.e. independent of v). Thus (2.74) becomes

$$a = a'/\gamma^3. \tag{2.78}$$

Now

$$a = \frac{\mathrm{d}u}{\mathrm{d}\tau}\Big/\frac{\mathrm{d}t}{\mathrm{d}\tau}, \tag{2.79}$$

and using (2.44), so we can find the parameterised equation of the world line by integrating

$$\frac{\mathrm{d}u}{\mathrm{d}\tau} = a\frac{\mathrm{d}t}{\mathrm{d}\tau} = a'/\gamma^2. \tag{2.80}$$

This gives

$$u = c\tanh(a'\tau/c), \tag{2.81}$$

choosing the origin of τ so that $u = 0$ when $\tau = 0$, and hence

$$\gamma = \cosh(a'\tau/c). \tag{2.82}$$

Then from $\mathrm{d}t\,/\,\mathrm{d}\tau = \gamma$, we find that

$$t = c/a'\sinh(a'\tau/c), \tag{2.83}$$

choosing the origin of t such that $t = 0$ when $\tau = 0$. Finally,

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau} \tag{2.84}$$

$$= u\gamma \tag{2.85}$$

$$= c\sinh(a'\tau/c), \tag{2.86}$$

so, choosing the origin of x such that $x = c^2/a'$ when $t = 0$,

$$x = c^2/a'\cosh(a'\tau/c). \tag{2.87}$$

Uniformly accelerated particles therefore move on rectangular hyperbolas of the form

$$x^2 - (ct)^2 = (c^2/a')^2. \tag{2.88}$$

The diagram in Fig 2.14 shows the trajectory. The dotted lines are the light cones. An event taking place within the dashed lines can influence an accelerated observer at the

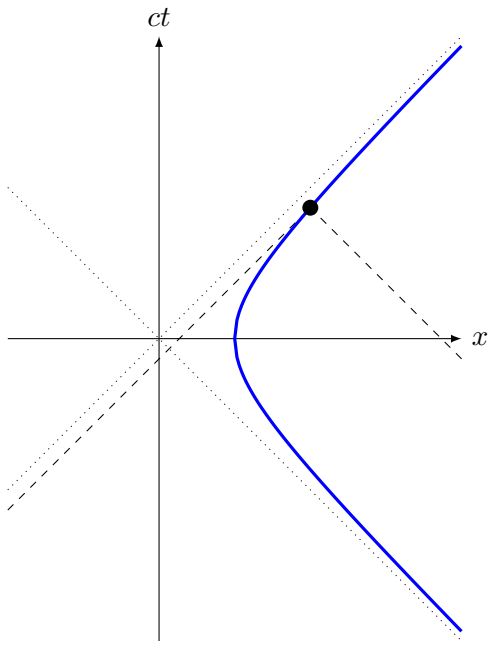

Fig. 2.14: The space-time diagram for an accelerated observer. The thick hyperbola is the observer's world line. An observer 'below' the dashed lines could in principle send a message to the observer marked as a heavy dot; other observers could not.

position shown, but events taking place outside the dashed lines would have to move faster than the speed of light to do so. As $\tau \to \infty$, the whole of the space-time to the left of the dotted line $x = ct$ would be inaccessible to the observer. This line is called the *Rindler event horizon* for the accelerated observer. In some ways, it performs the same function as the event horizon of a black hole. In particular, the observer has to accelerate to avoid falling through it and anything happening on the other side would be hidden to the observer. Of course, the accelerating observer could just stop accelerating whereas the observer in a black hole space-time can do nothing to affect the event horizon.

Moreover, the accelerated observer sees the emitted light Doppler shifted to longer and longer wavelengths as the object approaches the event horizon and is observed as $\tau \to \infty$.

2.5.1 Rectilinear Acceleration

Consider a particle moving at a variable speed $u(\tau)$ along the x-axis in the inertial frame S, where τ is the particle's proper time. Let the particle carry an accelerometer that reads $f(\tau)$, this is the *proper acceleration*, the acceleration in the *instantaneous rest frame* of the particle at τ.

In the instantaneous rest frame at τ, $u'(\tau) = 0$ and $\mathrm{d}u'/\mathrm{d}t' = f(\tau)$; transforming back to the frame S using (the inverse of) Eq. (2.77), we have

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \left(1 - \frac{u^2}{c^2}\right)^{3/2} f(\tau). \tag{2.89}$$

We can express things with respect to proper time using $\mathrm{d}\tau = (1 - u^2/c^2)^{1/2}\,\mathrm{d}t$ so that

$$\frac{\mathrm{d}u}{\mathrm{d}\tau} = \left(1 - \frac{u^2}{c^2}\right) f(\tau); \tag{2.90}$$

in terms of the rapidity $\psi(\tau)$, with $u(\tau) = c\tanh\psi(\tau)$, this is $c\frac{\mathrm{d}\psi}{\mathrm{d}\tau} = f(\tau)$ so

$$c\psi(\tau) = \int_0^\tau f(\tau')\,\mathrm{d}\tau'\,, \tag{2.91}$$

taking $u = 0$ at $\tau = 0$.

To parameterise the worldline of the particle in S, we can use

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \left(1 - \frac{u^2}{c^2}\right)^{-1/2} = \cosh\psi(\tau) \tag{2.92}$$

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = u\left(1 - \frac{u^2}{c^2}\right)^{-1/2} = c\sinh\psi(\tau). \tag{2.93}$$

Integrating these equations gives the coordinates in S of the wordline, $t(\tau)$ and $x(\tau)$.

Consider now the simple case of uniform or constant proper acceleration. This does *not* mean that $\mathrm{d}u/\mathrm{d}t$ = const., since u cannot exceed c. Rather, for f = const. the rapidity rises linearly with τ, $\psi(\tau) = f\tau/c$, and the worldline is

$$t = t_0 + \frac{c}{f}\sinh\left(\frac{f\tau}{c}\right) \tag{2.94}$$

$$x = x_0 + \frac{c^2}{f}\left(\cosh\left(\frac{f\tau}{c}\right) - 1\right), \tag{2.95}$$

where t_0 and x_0 are integration constants.

Setting $ct_0 = x_0 = 0$, we have a hyperbolic trajectory through the origin, as shown to the right **??**, with an oblique asymptote in the future of $ct = c^2/f + x$.

This means that there are regions of spacetime containing events that can never influence (i.e., communicate causally with) the accelerated particle (events to the left of the dotted line). The boundary of this region defines an *event horizon* of the accelerated observer.

As an example, light emitted from an object at rest at $x = 0$ in S will only reach the accelerated observer if it is emitted before $t = c/f$. Moreover, the accelerated observer sees the emitted light Doppler shifted to longer and longer wavelengths as the object approaches the event horizon and is observed as $\tau \to \infty$.

Chapter 3

Introducing Differential Geometry

Gravity is geometry. To fully understand this statement, we will need more sophisticated tools and language to describe curved space and, ultimately, curved spacetime. This is the mathematical subject of differential geometry and will be introduced in this chapter and the next.

Our discussion of differential geometry is not particularly rigorous. We will not prove many big theorems. Furthermore, a number of the statements that we make can be checked straightforwardly but we will often omit this. We will, however, be careful about building up the mathematical structure of curved spaces in the right logical order. As we proceed, we will come across a number of mathematical objects that can live on curved spaces. Many of these are familiar – like vectors, or differential operators – but we'll see them appear in somewhat unfamiliar guises. The main purpose of this chapter is to understand what kind of objects can live on curved spaces, and the relationships between them. This will prove useful for both general relativity and other areas of physics.

Moreover, there is a wonderful rigidity to the language of differential geometry. It sometimes feels that any equation that you're allowed to write down within this rigid structure is more likely than not to be true! This rigidity is going to be of enormous help when we return to discuss theories of gravity.

3.1 Concept of a Manifold

The stage on which our story will play out is a mathematical object called a *manifold*. We will give a precise definition below, but for now you should think of a manifold as a curved, n-dimensional space. If you zoom in to any patch, the manifold looks like $\mathbb{R}^n$. But, viewed more globally, the manifold may have interesting curvature or topology.

To begin with, our manifold will have very little structure. For example, initially there will be no way to measure distances between points. But as we proceed, we will describe the various kinds of mathematical objects that can be associated to a manifold, and each one will allow us to do more and more things. It will be a surprisingly long time before we can measure distances between points!

You have met many manifolds in your education to date, even if you didn't call them by name. Some simple examples in mathematics include Euclidean space $\mathbb{R}^n$, the sphere $\mathbb{S}^n$, and the torus $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$. Some simple examples in physics include the configuration space and phase space that we use in classical mechanics and the state space of thermodynamics. As we progress, we will see how familiar ideas in these subjects can be expressed in a more formal language. Ultimately our goal is to explain how spacetime is a manifold and to understand the structures that live on it.

We now come to our main character: an n-dimensional manifold is a space which, locally, looks like $\mathbb{R}^n$. Globally, the manifold may be more interesting than $\mathbb{R}^n$, but the idea is that we can patch together these local descriptions to get an understanding for the entire space.

Informally, an N-dimensional manifold is a set of objects that locally resembles ND Euclidean space $\mathbb{R}^n$. In relativity, the objects are events and the set of events is spacetime. What "locally resembles" means is that there exists a map ϕ from the ND manifold $\mathcal{M}$ to an *open subset* of $\mathbb{R}^n$ that is one-to-one and onto.[1]

Under the map ϕ, a point $P \in \mathcal{M}$ maps to a point in the open subset U of $\mathbb{R}^n$ with *coordinates* x_a , $a = 1, \cdots, N$. Generally, we cannot cover the entire manifold with a single map ϕ (or, equivalently, set of coordinates), but it is sufficient if we can subdivide $\mathcal{M}$ and map each piece separately onto open subsets of $\mathbb{R}^n$.

The manifold is *differentiable* if these subdivisions join up smoothly so that we can define scalar fields on the manifold that are differentiable everywhere.

We can generally think of manifolds as surfaces embedded in some higher-dimensional Euclidean space, and we shall often do so, but it is important to appreciate that a given manifold exists independent of any embedding. A non-trivial example of a manifold is the set of rotations in 3D; these can be parameterised by three Euler angles, which form a coordinate system for the 3D manifold.

3.2 Coordinates

As we have just seen, points in an ND manifold can be labelled by N real-valued coordinates $(x_1, x_2, \cdots, x_N)$. We shall denote these collectively by x_a with $a = 1, \cdots, N$. The coordinates are not unique: think of them as labels of points in the manifold that can change under a coordinate transformation (i.e., a change of map ϕ) while the point itself does not.

We have also noted that, generally, it will not be possible to cover a manifold with a single *non-degenerate* coordinate system, i.e., one where the correspondence between points and coordinate labels is one-to-one. In such cases, multiple coordinate systems are required to cover the whole manifold.

Here are a few simple examples of differentiable manifolds:

- Coordinates (ρ, ϕ) in the plane $\mathbb{R}^2$: The Euclidean plane $\mathbb{R}^2$ is a 2D manifold that can be covered globally with the usual Cartesian coordinates. However, we could instead use plane-polar coordinates, (ρ, ϕ) with $0 \leq \rho \leq \infty$ and $0 \leq \phi < 2\pi$. Plane-polar coordinates are degenerate at $\rho = 0$ since ϕ is indeterminate there.

[1] An open subset U of $\mathbb{R}^n$ is such that for any point one can construct a sphere centred on the point whose interior lies entirely inside U. A map from $\mathcal{M}$ to U is one-to-one and onto if every element of U is mapped to by exactly one element of $\mathcal{M}$.

- Coordinates (θ, ϕ) on the 2-sphere $\mathbb{S}^2$: The 2-sphere is the set of points in $\mathbb{R}^3$ with $x^2+y^2+z^2=1$. It is an example of a 2D manifold. The spherical polar coordinates (θ, ϕ), with $0 \leq \theta \leq \pi$ and $0 \leq \phi < 2\pi$, are degenerate at the poles $\theta = 0$ and $\theta = \pi$, where ϕ is indeterminate. For $\mathbb{S}^2$, there is no single coordinate system that covers the whole manifold without degeneracy: at least two coordinate patches are required.

3.2.1 Curves and Surfaces

Subsets of points in a manifold define *curves* and *surfaces*. These are usually defined parametrically for some coordinate system, e.g., for a curve with parameter u:

$$x^a = x^a(u) \qquad (a = 1, 2, \cdots, N). \tag{3.1}$$

For a *submanifold* (or surface) of M ($M < N$) dimensions, we need M parameters for a total of N defining equations:

$$x^a = x^a(u^1, u^2, \cdots, u^M) \qquad (a = 1, 2, \cdots, N). \tag{3.2}$$

The special case $M = N - 1$ is called a *hypersurface*. In this case, we can eliminate the $N-1$ parameters from the N equations (3.2) to give

$$f(x^1, x^2, \cdots, x^N) = 0, \tag{3.3}$$

for some function f. E.g. $(x^1)^2 + (x^2)^2 + (x^3)^1 - 1 = 0$.

Similarly, points in an M-dimensional surface can be specified by $N-M$ (independent) constraints

$$f_1(x^1, x^2, \cdots, x^N) = 0, \cdots, f_{N-M}(x^1, x^2, \cdots, x^M) = 0, \tag{3.4}$$

i.e., by the intersection of $N - M$ hypersurfaces, as an alternative to the parametric representation of Eq. (3.2).

3.2.2 Coordinate Transformations

Coordinates are used to label points in a manifold, but the labelling is arbitrary. Later, we shall learn how to construct geometric objects that are independent of the way we assign coordinates, and that express the true physical content of the theory (think vectors in $\mathbb{R}^n$).

We can relabel points by performing a coordinate transformation given by N equations

$$x'^a = x'^a(x^1, x^2, \cdots, x^N) \qquad (a = 1, 2, \cdots, N). \tag{3.5}$$

We shall view coordinate transformations as *passive*, i.e., assigning new coordinates x'^a to a given point in terms of the original coordinates x^a. We shall further assume that the functions $x'^a(x^1, x^2, \cdots, x^N)$ are single-valued, continuous and differentiable.

Consider two neighbouring points P and Q with coordinates x^a and $x^a + \mathrm{d}x^a$. In the new (primed) coordinates,

$$\mathrm{d}x'^a = \sum_{b=1}^{N} \underbrace{\frac{\partial x'^a}{\partial x^b}}_{J^a{}_b} \mathrm{d}x^b \,, \tag{3.6}$$

where the partial derivatives are evaluated at the point P. This defines an $N \times N$ transformation matrix at the point P with elements

$$J^a{}_b = \frac{\partial x'^a}{\partial x^b} = \begin{pmatrix} \frac{\partial x'^1}{\partial x^1} & \cdots & \frac{\partial x'^1}{\partial x^N} \\ \vdots & & \vdots \\ \frac{\partial x'^N}{\partial x^1} & \cdots & \frac{\partial x'^N}{\partial x^N} \end{pmatrix}, \tag{3.7}$$

where the numerator (index a) labels the rows and the denominator (index b) the columns. The determinant of $J \equiv \det(J^a{}_b)$ is the *Jacobian* of the transformation. If $J \neq 0$ for some range of the coordinates, the coordinate transformation can be inverted locally to give x^a as a function of the x'^a.

The transformation matrix for the inverse

$$x^a = x^a(x'^1, x'^2, \cdots, x'^N) \tag{3.8}$$

is the inverse of $J^a{}_b$; this follows from the chain rule for partial derivatives,

$$\sum_{b=1}^{N} \frac{\partial x'^a}{\partial x^b} \frac{\partial x^b}{\partial x'^c} = \frac{\partial x'^a}{\partial x'^c} = \delta_{ac}. \tag{3.9}$$

It also follows that the determinant of the inverse transformation is $1/J$.

3.2.3 Einstein Summation Convention

It will rapidly get cumbersome to include the summation over indices explicitly, as in Eq. (3.6). We therefore introduce the Einstein summation convention: Whenever an index occurs twice in an expression, once as a subscript and once as a superscript, summation over the index from 1 to N is implied.

For example, for an infinitesimal displacement

$$\mathrm{d}x'^a = \frac{\partial x'^a}{\partial x^b} \mathrm{d}x^b \,. \tag{3.10}$$

Here, the index a is a free index and may take any value from 1 to N, while the index b is summed over 1 to N. Note the following points about the summation convention.

- A superscript in the denominator of a partial derivative is considered a subscript, which is why the index b in Eq. (3.10) is summed over.
- Indices that are summed over are called dummy indices because can be replaced by any other index not already in use, e.g.,

$$\frac{\partial x'^a}{\partial x^b} \mathrm{d}x^b = \frac{\partial x'^a}{\partial x^c} \mathrm{d}x^c \,. \tag{3.11}$$

- In any term, an index should not occur more than twice, and any repeated index must occur once as a subscript and once as a superscript (and is summed over).

3.3 Local Geometry of Riemannian Manifolds

The general definition of a differentiable manifold does not define its *geometry*. To do so requires introducing additional structure to the manifold. Consider two neighbouring points P and Q in a manifold, i.e., points with coordinates x^a and $x^a + \mathrm{d}x^a$, in some coordinate system, which differ infinitesimally. The *local geometry* near P is specified by giving the invariant "distance" or "interval" between the points. In a *Riemannian manifold*, the interval takes the form (summation convention!)

$$\boxed{\mathrm{d}s^2 = g_{ab}(x)\,\mathrm{d}x^a\,\mathrm{d}x^b\,,} \tag{3.12}$$

i.e., the interval is quadratic in the coordinate differentials. The coefficients $g_{ab}(x)$ contain information about the local geometry but also depend on the particular coordinate system. Strictly, the geometry is *Riemannian* if $\mathrm{d}s^2 > 0$ and *pseudo-Riemannian* otherwise (the latter being the relevant case for spacetime). It is also possible to consider more general intervals, but these are not relevant for general relativity because of the equivalence principle.

3.3.1 The Metric

The metric functions relate infinitesimal changes in the coordinates to invariantly-defined "distances" in the manifold. In general relativity, these will be proper distances and times. The metric functions $g_{ab}(x)$ can always be chosen symmetric, $g_{ab}(x) = g_{ba}(x)$. To see this, note that we can write a general gab as the sum of a symmetric and antisymmetric part:

$$g_{ab}(x) = \tfrac{1}{2}[g_{ab}(x) + g_{ba}(x)] + \tfrac{1}{2}[g_{ab}(x) - g_{ba}(x)]. \tag{3.13}$$

The contribution of the antisymmetric part to $\mathrm{d}s^2$ vanishes since

$$\begin{aligned}(g_{ab} - g_{ba})\,\mathrm{d}x^a\,\mathrm{d}x^b &= g_{ab}\,\mathrm{d}x^a\,\mathrm{d}x^b - g_{ba}\,\mathrm{d}x^b\,\mathrm{d}x^a \\ &= (g_{ab} - g_{ab})\,\mathrm{d}x^a\,\mathrm{d}x^b \\ &= 0\end{aligned} \tag{3.14}$$

where we have relabelled the dummy indices $a \leftrightarrow b$ in the first line on the right.

It follows that in an N-dimensional Riemannian manifold there are $N(N+1)/2$ independent metric functions at each point. Given two neighbouring points, the interval between them is independent of the coordinate system used. Since the coordinate differentials change under a change of coordinates, so must the metric functions, i.e.,

$$\begin{aligned}\mathrm{d}s^2 &= g_{ab}(x)\,\mathrm{d}x^a\,\mathrm{d}x^b \\ &= \underbrace{g_{ab}(x)\frac{\partial x^a}{\partial x'^c}\frac{\partial x^b}{\partial x'^d}}_{g'_{cd}(x')}\,\mathrm{d}x'^c\,\mathrm{d}x'^d \\ &= g'_{cd}(x')\,\mathrm{d}x'^c\,\mathrm{d}x'^d\,,\end{aligned} \tag{3.15}$$

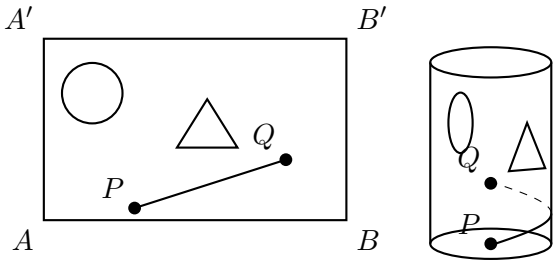

Fig. 3.1: The Euclidean plane $\mathbb{R}^2$ can be rolled up into a cylindrical surface without distortion. The intrinsic geometry of the cylindrical surface is therefore the same as the plane. In particular, a bug confined to the surface would measure the sum of the angles of a triangle to be 180° and the circumference of a circle to be 2π times its radius.

where the metric functions in the new coordinates at the same physical point are $g'_{cd}(x')$.

We can read off from Eq. (3.15) that the metric functions must transform as

$$\boxed{g'_{cd}(x') = g_{ab}(x(x'))\frac{\partial x^a}{\partial x'^c}\frac{\partial x^b}{\partial x'^d}.} \tag{3.16}$$

Since there are N arbitrary coordinate transformations that we can make, there are really only $N(N-1)/2$ independent functional degrees of freedom associated with $g_{ab}(x)$.

3.3.2 Intrinsic and Extrinsic Geometry

The interval (or *line element*) $\mathrm{d}s^2$ characterises the local geometry (or curvature), which is an *intrinsic* property of the manifold independent of any possible embedding in some higher-dimensional space.

Intrinsic properties are those that can be determined by a "bug" *confined* to the manifold – the bug can set up a coordinate system, measure physical distances and hence determine the metric functions.

As an example of the distinction between intrinsic and extrinsic geometry, consider the surface of a cylinder of radius a embedded in $\mathbb{R}^3$ (see Fig. 3.1).

In a cylindrical polar coordinate system, (z, ϕ), the interval is

$$\mathrm{d}s^2 = \mathrm{d}z^2 + a^2\,\mathrm{d}\phi^2\,. \tag{3.17}$$

The intrinsic geometry is locally identical to the 2D Euclidean plane $\mathbb{R}^2$ since the coordinate transformation $'0 = a\phi$ and $z' = z$ gives $\mathrm{d}s^2 = \mathrm{d}y'^2 + \mathrm{d}z'^2$ everywhere. This makes physical sense since the cylinder can be unrolled to give the plane without buckling, tearing or otherwise distorting.

However, the *extrinsic geometry* as seen within the embedding space $\mathbb{R}^3$ is clearly curved (non-Euclidean). We can contrast the cylinder to a 2-sphere of radius a embedded in $\mathbb{R}^3$. The intrinsic geometry, based on measurements made within the surface, is now

not identical to the Euclidean plane since the surface of a sphere cannot be formed from the flat plane without deformation (this is why gift-wrapping a ball is hard!).

If we use polar coordinates (θ, ϕ), the interval on the 2-sphere is

$$\mathrm{d}s^2 = a^2(\mathrm{d}\theta^2 + \sin\theta\,\mathrm{d}\phi^2). \tag{3.18}$$

This cannot be transformed to Euclidean form $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2$ over the *entire* surface by any coordinate transformation, which shows that the intrinsic geometry is non-Euclidean – the space is intrinsically curved. Note that at any point A, we can find coordinates (see example below) such that $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2$ in the local neighbourhood of A, but *not* over the entire surface.

General relativity is a theory involving the (local) intrinsic geometry of the spacetime manifold – no embedding in some higher-dimensional space is required.

- The 2-sphere in $\mathbb{R}^3$: For a surface embedded in a higher-dimensional space, the induced line element in the surface is determined by the line element in the embedding space and the "shape" of the surface. Consider the 2-sphere embedded in $\mathbb{R}^3$; the embedding space has the Euclidean line element $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$ in Cartesian coordinates. If the sphere has radius a, points on its surface satisfy $x^2+y^2+z^2 = a^2$, so that

$$\begin{aligned} 0 &= 2x\,\mathrm{d}x + 2y\,\mathrm{d}y + 2z\,\mathrm{d}z \\ \mathrm{d}z &= -\frac{(x\,\mathrm{d}x + y\,\mathrm{d}y)}{z} = -\frac{(x\,\mathrm{d}x + y\,\mathrm{d}y)}{\sqrt{a^2 - x^2 - y^2}}. \end{aligned} \tag{3.19}$$

  This is the constraint on $\mathrm{d}z$ that keeps us on the spherical surface for a displacement $\mathrm{d}x$ and $\mathrm{d}y$ in the x and y coordinates. We obtain the induced line element by substituting $\mathrm{d}z$ in the line element of the embedding space ($\mathbb{R}^3$ here) to find

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \frac{(x\,\mathrm{d}x + y\,\mathrm{d}y)^2}{a^2 - (x^2 + y^2)}. \tag{3.20}$$

  Near the north or south poles, where $x^2 + y^2 \ll a^2$, the induced line element is approximately the Euclidean form, $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2$.

  The induced metric looks neater if we use plane polar coordinates $x = \rho\cos\phi$ and $y = \rho\sin\theta$; then

$$\begin{aligned} \mathrm{d}x &= \cos\phi\,\mathrm{d}\rho - \rho\sin\phi\,\mathrm{d}\phi \\ \mathrm{d}y &= \sin\phi\,\mathrm{d}\rho + \rho\cos\phi\,\mathrm{d}\phi\,, \end{aligned} \tag{3.21}$$

  and so $x\,\mathrm{d}x + y\,\mathrm{d}y = \rho\,\mathrm{d}\rho$ and $\mathrm{d}x^2 + \mathrm{d}y^2 = \mathrm{d}\rho^2 + \rho^2\,\mathrm{d}\phi^2$. Putting these pieces together gives

$$\mathrm{d}s^2 = \frac{a^2\,\mathrm{d}\rho^2}{(a^2 - \rho^2)} + \rho^2\,\mathrm{d}\phi^2\,. \tag{3.22}$$

- The 3-sphere in $\mathbb{R}^4$: Now consider the 3-sphere, defined by $x^2 + y^2 + z^2 + w^2 = a^2$, embedded in 4D Euclidean space $\mathbb{R}^4$ with line element

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \mathrm{d}w^2\,. \tag{3.23}$$

Differentiating gives

$$\begin{aligned} 0 &= 2x\,\mathrm{d}x + 2y\,\mathrm{d}y + 2z\,\mathrm{d}z + 2w\,\mathrm{d}w \\ \implies \mathrm{d}w &= -\frac{(x\,\mathrm{d}x + y\,\mathrm{d}y + z\,\mathrm{d}z)}{w} \\ &= -\frac{(x\,\mathrm{d}x + y\,\mathrm{d}y + z\,\mathrm{d}z)}{\sqrt{a^2 - (x^2+y^2+z^2)}}, \end{aligned} \tag{3.24}$$

and so the induced line element is

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \frac{(x\,\mathrm{d}x + y\,\mathrm{d}y + z\,\mathrm{d}z)^2}{a^2 - (x^2+y^2+z^2)}. \tag{3.25}$$

As for the 2-sphere, the line element looks neater in polar coordinates; this time we use spherical-polar coordinates

$$\begin{aligned} x &= r\sin\theta\cos\phi, \\ y &= r\sin\theta\sin\phi, \\ z &= r\cos\theta. \end{aligned} \tag{3.26}$$

We find $x\,\mathrm{d}x + y\,\mathrm{d}y + z\,\mathrm{d}z = r\,\mathrm{d}r$ so that

$$\mathrm{d}s^2 = \frac{a^2}{(a^2 - r^2)}\,\mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2\,. \tag{3.27}$$

This line element describes a non-Euclidean 3D space. We shall meet this space again towards the end of the course, where we shall see that it describes the spatial part of a cosmological model with compact spatial sections (a closed universe). In the limit $a \to \infty$ we recover 3D Euclidean space in spherical-polar coordinates. More generally, for $r \ll a$ we recover $\mathbb{R}^3$ locally.

3.4 Lengths and Volumes

The metric functions determine an invariant distance measure on the manifold, and so also determine invariant "lengths" of curves and "volumes" of subregions.

3.4.1 Lengths along Curves

Consider a curve $x^a(u)$ between points A and B on some manifold. Since $\mathrm{d}s^2 = g_{ab}(x)\,\mathrm{d}x^a\,\mathrm{d}x^b$ is the invariant distance between neighbouring points with coordinates separated by $\mathrm{d}x^a$, the invariant length along the curve, as the sum of the infinitesimal invariant displacements $\int \mathrm{d}s$, is

$$\boxed{L_{AB} = \int_{u_B}^{u_A} \left| g_{ab}\frac{\mathrm{d}x^a}{\mathrm{d}u}\frac{\mathrm{d}x^b}{\mathrm{d}u}\right|^2 \mathrm{d}u\,.} \tag{3.28}$$

(The modulus sign is not required for a Riemannian manifold, where $\mathrm{d}s^2 > 0$, but is generally required for application to spacetime.)

3.4.2 Volumes of Regions

To calculate the volume of some region, we shall initially consider the simple case where the metric is diagonal, i.e., $g_{ab}(x) = 0$ for $a \neq b$. In this case,

$$\mathrm{d}s^2 = g_{11}(\mathrm{d}x^1)^2 + g_2(\mathrm{d}x^2)^2 + \cdots + g_{NN}(\mathrm{d}x^N)^2. \tag{3.29}$$

A coordinate system with a diagonal metric is called orthogonal since, as we shall discuss later when considering tangent vectors to curves, the coordinate curves (i.e., the curves obtained by allowing a single coordinate to vary in turn) are orthogonal to each other.

To be concrete, consider the 2D manifold $\mathcal{M}$ illustrated by the curved surface, in which the coordinates x^1 and x^2 form an orthogonal coordinate system in $\mathcal{M}$. The volume element (infinitesimal rectangle for an orthogonal coordinate system in 2D) defined by coordinate increments $\mathrm{d}x^1$ and $\mathrm{d}x^2$ has sides of invariant length $\sqrt{g_{11}}\,\mathrm{d}x^1$ and $\sqrt{g_{22}}\,\mathrm{d}x^2$. It follows that the invariant volume element is

$$\mathrm{d}V = \sqrt{|g_{11}g_{22}|}\,\mathrm{d}x^1\,\mathrm{d}x^2\,. \tag{3.30}$$

This generalises to the volume element of an ND manifold,

$$\boxed{\mathrm{d}V = \sqrt{|g_{11}g_{22}\cdots g_{NN}|}\,\mathrm{d}x^1\,\mathrm{d}x^2\cdots\mathrm{d}x^N\,.} \tag{3.31}$$

Similarly, one can define "area"-like elements on surfaces within manifolds by using the induced line element on the surface.

3.4.2.1 Invariance of the Volume Element

The result (3.31) for the volume element involves the determinant of the metric, since for a diagonal metric $g \equiv \det(g_{ab}) = g_{11}g_{22}\cdots g_{NN}$. This suggests that the generalisation to an arbitrary coordinate system is

$$\mathrm{d}V = \sqrt{|g|}\,\mathrm{d}x^1\,\mathrm{d}x^2\cdots\mathrm{d}x^N\,. \tag{3.32}$$

Let us check that this is indeed an invariant volume element.

Consider a coordinate transformation $x^a \to x'^a$; under this, $\mathrm{d}x^1\,\mathrm{d}x^2\cdots\mathrm{d}x^N$ transforms with the Jacobian of the transformation matrix:

$$\mathrm{d}x'^1\,\mathrm{d}x'^2\cdots\mathrm{d}x'^N = \mathrm{d}x^1\,\mathrm{d}x^2\cdots\mathrm{d}x^N\,. \tag{3.33}$$

where, recall from Eq. (3.7) that $J = \det\left(\partial x'^a/\partial x^b\right)$. Since the metric transforms as

$$g'_{ab} = \frac{\partial x^c}{\partial x'^a}\frac{\partial x^d}{\partial x'^b}g_{cd}, \tag{3.34}$$

the determinant of the metric transforms as

$$g' = g/J^2. \tag{3.35}$$

Here, we have used that $1/J = \det\left(\partial x^a/\partial x'^b\right)$, which follows since $\partial x^a/\partial x'^b$ is the inverse of the transformation matrix. It follows that

$$\sqrt{|g'|}\,\mathrm{d}x'^1\,\mathrm{d}x'^2\cdots\mathrm{d}x'^N = \frac{\sqrt{|g|}}{J}J\,\mathrm{d}x^1\,\mathrm{d}x^2\cdots\mathrm{d}x^N\,, \tag{3.36}$$

and so $\mathrm{d}V = \sqrt{|g|}\,\mathrm{d}x^1\,\mathrm{d}x^2\cdots\mathrm{d}x^N$ is indeed invariant.

3.4.2.2 Example: Surface of the 2-sphere in $\mathbb{R}^3$

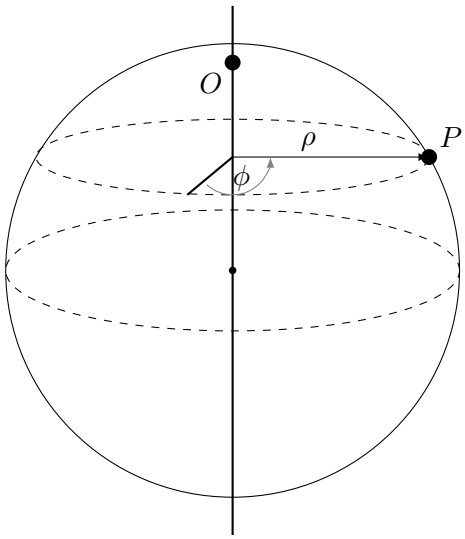

Fig. 3.2: Surface of the 2-sphere in $\mathbb{R}^3$, with centre O.

Consider again the 2-sphere of radius a embedded in $\mathbb{R}^3$. We write the line element as

$$\mathrm{d}s^2 = \frac{a^2\,\mathrm{d}\rho^2}{(a^2-\rho^2)} + \rho^2\,\mathrm{d}\phi^2\,, \tag{3.37}$$

so the metric is diagonal with components

$$g_{11} = \frac{a^2}{(a^2-\rho^2)}, \qquad \text{and,} \qquad g_{22} = \rho^2. \tag{3.38}$$

Consider the circle $\rho = R$ (upper dashed circle in Fig. 3.2); we shall compute its length, the distance from its centre O to its perimeter, and the area enclosed. The distance from the centre O to the perimeter along the curve $\phi = \text{const.}$ is given by

$$D = \int_0^R \frac{a^2}{(a^2-\rho^2)^{1/2}}\,\mathrm{d}\rho = a\sin^{-1}\left(\frac{R}{a}\right). \tag{3.39}$$

For the circumference of the circle, we have

$$C = \int_0^{2\pi} R\,\mathrm{d}\phi = 2\pi R. \tag{3.40}$$

For the area enclosed, we use Eq. (3.31) noting that in 2D the enclosed area is the "volume":

$$\begin{aligned} A &= \int_0^{2\pi} \int_0^R \frac{a^2}{(a^2-\rho^2)^{1/2}} \rho \, \mathrm{d}\rho \, \mathrm{d}\phi \\ &= 2\pi a^2 \left[1 - \left(1 - \frac{R^2}{a^2} \right)^{1/2} \right]. \end{aligned} \tag{3.41}$$

We can rewrite these results for C and A in terms of the (radius) distance D as follows:

$$\begin{aligned} C &= 2\pi a \sin\left(\frac{D}{a}\right) \\ A &= 2\pi a^2 \left[1 - \cos\left(\frac{D}{a}\right) \right] \end{aligned} \tag{3.42}$$

We note the following points about these results:

- For $D \ll a$, we recover the Euclidean results $C = 2\pi D$ and $A = \pi D^2$.
- As D increases, both C and A increase until $D = \pi a/2$, after which C decreases.
- The coordinates (ρ, ϕ) are degenerate beyond the equator (the metric coefficient g_{11} makes it clear that the coordinates are poorly behaved at $\rho = a$).

However, if we switch to coordinates (D, ϕ), this system is well defined beyond the equator, becoming degenerate only at $D = \pi a$ (the south pole). The metric in these coordinates is

$$\mathrm{d}s^2 = \mathrm{d}D^2 + a^2 \sin^2\left(\frac{D}{a}\right) \mathrm{d}\phi^2 \, . \tag{3.43}$$

3.5 Local Cartesian Coordinates

On a Riemannian manifold (assume $\mathrm{d}s^2 > 0$ for now) it is generally *not* possible to choose coordinates such the line element takes the Euclidean form at every point. This follows since $g_{ab}(x)$ has $N(N+1)/2$ independent functions, but there are only N functions involved in coordinate transformations. However, it is always possible to adopt coordinates such that in the neighbourhood of some point P, the line element takes the Euclidean form. More precisely, we can always find coordinates such that at P,

$$g_{ab}(P) = \delta_{ab}, \qquad \text{and,} \qquad \left.\frac{\partial g_{ab}}{\partial x^c}\right|_P = 0. \tag{3.44}$$

This means that, in the neighbourhood of P, we have

$$g_{ab}(x) = \delta_{ab} + \mathcal{O}\Big((x - x_P)^2\Big) \tag{3.45}$$

in these special coordinates. Such coordinates are called local Cartesian coordinates at P. In general relativity, we shall see that the generalisation of such coordinates to spacetime corresponds to coordinates defined by locally-inertial (i.e., free-falling) observers.

3.5.1 Proof of Existence of Local Cartesian Coordinates

We shall prove the existence of local Cartesian coordinates by showing that a coordinate transformation $x^a \to x'^a$ has enough degrees of freedom to bring the metric to the form in Eq. (3.45). Under the coordinate transformation, the metric and its derivatives transform as

$$g'_{ab} = \frac{\partial x^c}{\partial x'^a}\frac{\partial x^d}{\partial x'^b} g_{cd}, \tag{3.46}$$

$$\frac{\partial g'_{ab}}{\partial x'^e} = \frac{\partial}{\partial x'^e}\left(\frac{\partial x^c}{\partial x'^a}\frac{\partial x^d}{\partial x'^b}\right) + \frac{\partial x^c}{\partial x'^a}\frac{\partial x^d}{\partial x'^b}\frac{\partial x^f}{\partial x'^e}\frac{\partial g_{cd}}{\partial x^f}. \tag{3.47}$$

We now try to construct the (as-yet) unknown relation $x^a(x')$ such that in the primed coordinates $g'_{ab} = \delta_{ab}$ and $\partial g'_{ab}/\partial x'^c = 0$ at P.

Consider the transformation matrices and their derivatives that appear in Eq.s (3.46) and (3.47); at the point P, the numbers of independent degrees of freedom in these are

$$\left.\frac{\partial x^a}{\partial x'^b}\right|_P \qquad N^2 \text{ values}, \tag{3.48}$$

$$\left.\frac{\partial^2 x^a}{\partial x'^b \partial x'^c}\right|_P \qquad N^2(N+1)/2 \text{ values}. \tag{3.49}$$

Contrast these with the number of degrees of freedom in the metric and its derivatives at P:

$$g'_{ab}(P) \qquad N(N+1)/2 \text{ values}, \tag{3.50}$$

$$\left.\frac{\partial g'_{ab}}{\partial x'^c}\right|_P \qquad N^2(N+1)/2 \text{ values}. \tag{3.51}$$

If we try and set $g'_{ab}(P) = \delta_{ab}$ (which are $N(N+1)/2$ equations), we have more than enough degrees of freedom in the $\partial x^a \big/ \partial x'^b$ to do so. Indeed, we are left with $N(N-1)/2$ 'unused' degrees of freedom in the $\partial x^a \big/ \partial x'^b$. For $N = 4$ in spacetime, these correspond to the six degrees of freedom (three boosts, three rotations) associated with homogeneous Lorentz transformations that preserve the Minkowski form of the metric.

Now consider trying to enforce further that $\partial g'_{ab}/\partial x'^c = 0$ at P. These are $N^2(N+1)/2$ equations, which consume all of the second derivatives $\partial^2 x^a \big/ \partial x'^b \partial x'^c$. This proves that it is always possible to construct local Cartesian coordinates at a point. Can we go further, i.e., can we also set the second derivatives of the metric to zero? The answer is no: $\partial^2 g'_{ab} \big/ \partial x'^c \partial x'^d = 0$ gives $N^2(N+1)^2/4$ equations, but the number of degrees of freedom in the third derivatives of the coordinates, $\partial^3 x^a \; \partial x'^b \partial x'^c \partial x'^d$ is only $N^2(N+1)(N+2)/6$.

We see that there are generally $N^2(N+1)^2/4 - N^2(N+1)(N+2)/6 = N^2(N^2-1)/12$ independent degrees of freedom in the second derivatives of the metric that cannot be eliminated by coordinate transformations. It is these (20 for $N = 4$) that describe the *curvature* of the manifold and, in general relativity, the physical degrees of freedom associated with gravity.

3.6 Pseudo-Riemannian Manifolds

In a Riemannian manifold, $\mathrm{d}s^2 = g_{ab}\,\mathrm{d}x^a\,\mathrm{d}x^b$ is always positive for all $\mathrm{d}x^a$. Considered as a matrix, g_{ab} has to be positive definite at every point and so have all eigenvalues positive. In a *pseudo-Riemannian* manifold, $\mathrm{d}s^2$ can be positive, negative or zero depending on $\mathrm{d}x^a$, which implies that some of the eigenvalues of g_{ab} are negative. In a pseudo-Riemannian manifold one can always find coordinates such that at a point P,

$$g_{ab}(P) = \eta_{ab}, \tag{3.52}$$

and the first derivatives of the metric vanish at P. Here,

$$\eta_{ab} = \mathrm{diag}(\pm 1, \pm 1, \cdots, \pm 1), \tag{3.53}$$

where the number of positive entries in η_{ab} minus the number of negative is the *signature* of the manifold.(We shall always assume that the metric is sufficiently regular that the signature is the same at all points in the manifold.) In the Minkowski spacetime of special relativity, we have the line element

$$\mathrm{d}s^2 = \mathrm{d}(ct)^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2\,. \tag{3.54}$$

This is an example of a pseudo-Riemannian manifold with $eta_{ab} = \mathrm{diag}(+1, -1, -1, -1)$ taking the coordinates to be (ct, x, y, z).

3.7 Topology of Manifolds

So far, we have discussed only the *local* geometry of manifolds, defined at any point by the line element. In addition, a manifold also has a *global* geometry or *topology*, defined (crudely) by identification of points with different coordinates as being coincident. For example, the surface of cylinder in $\mathbb{R}^3$ has same local intrinsic geometry as the Euclidean plane $\mathbb{R}^2$, but a different topology. Indeed, the compact dimension on the surface of the cylinder could be detected by a "bug" confined to the surface since by continuing in a straight line (we shall define what we mean by a "straight line" in a general manifold later in the course) in a certain direction the bug would return to the same physical point. Topology is an *intrinsic*, but non-local, property of a manifold. General relativity is a local theory, in which the local intrinsic geometry is determined by energy density of matter/radiation at that point. The field equations of general relativity do *not* constrain the global topology of the spacetime manifold.

Chapter 4

Vector and Tensor Algebra

In general relativity, spacetime is described by a nontrivial (pseudo-)Riemannian manifold and this is the arena on which the rest of physics is enacted. The equivalence principle tells us that, locally, the laws of physics reduce to those of special relativity when expressed in terms of locally-inertial coordinates defined by free-falling observers.

Our goal is therefore to formulate physical laws in such a way that they reduce to special relativity in locally-inertial coordinates. The most efficient way to do this is to write down equations that are true in a general coordinate system (i.e., their form is the same in all coordinate systems) and then demand that they reduce to the usual form in special relativity when expressed in locally-inertial coordinates. Such a coordinate-independent, or geometric, approach, naturally gives rise to vector-valued fields – at any point these are geometric objects that are independent of the choice of coordinate system (while their components are coordinate dependent).

In previous courses (e.g., electromagnetism) you have studied the calculus of vector fields in the Euclidean spaces $\mathbb{R}^2$ and $\mathbb{R}^3$, and considered the components of vectors in simple coordinate systems such as Cartesian and spherical polar coordinates. You have also met the notion of *tensors*, for example, the moment of inertia tensor that relates the angular velocity of a solid body to its angular momentum, and these are also essential geometric objects in general relativity. In this chapter, we shall see how to generalise familiar Euclidean ideas to define vectors and tensors in general (pseudo-)Riemannian manifolds and *arbitrary* coordinate systems.

4.1 Scalar and Vector Fields on Manifolds

4.1.1 Scalar Fields

A real (or complex) scalar field defined on (some subset of) a manifold $\mathcal{M}$ assigns a real (or complex) number to each *point* P in (the subset of) $\mathcal{M}$. If we label the points in $\mathcal{M}$ using some coordinate system x^a, we can express the scalar field as a function $\phi(x^a)$ of the coordinates.

The value of a scalar field at a given point P is independent of the chosen coordinate system. This means that if we change coordinates to x'^a, the scalar field is expressed as some different function of the new coordinates *at the same point*, $\phi'(x'^a)$, such that

$$\boxed{\phi'(x'^a) = \phi(x^a).} \tag{4.1}$$

4.1.2 Vector Fields and Tangent Spaces

When dealing with vectors in Euclidean space, you will have met two types of vector:

- *displacement vectors* connecting two points in the space;
- *local vectors* that are measured at a given observation point and refer solely to that point (e.g., the electric field).

Note that displacement vectors between infinitesimally-separated points are really local vectors, as are derivatives of displacement vectors (e.g., the velocity of a particle).

On a general manifold, we can only define local vectors – vectors defined at any given point P and that can be measured by a "bug" making local measurements in a small region around P. In particular, we must abandon the idea of displacement vectors as these generally have no intrinsic meaning except in the infinitesimal limit. Displacement vectors do make sense if we specify an embedding of $\mathcal{M}$ in some higher-dimensional Euclidean space, but we are interested only in intrinsic geometry here.

However, to gain some intuition, let us first consider the case where M is embedded in a Euclidean space but restrict attention to local vectors, such as the velocity of a particle confined to $\mathcal{M}$. The usual velocity vector, defined by the derivative of the displacement in the Euclidean space, then lies tangent to the manifold at P. For an ND manifold $\mathcal{M}$, the set of all possible local vectors at any point P lie in an ND subspace of the Euclidean embedding space.

This subspace is an ND vector space[1] $T_P(\mathcal{M})$, called the tangent space at P. The tangent spaces at different points are distinct so we cannot add local vectors at different points, only at the same point. These ideas can be generalised to remove any reference to embedding: at each point P of a general ND manifold $\mathcal{M}$, we can construct a ND vector space – the tangent space $T_P(\mathcal{M})$ – whose elements are (local) vectors.

4.1.3 Vectors as Differential Operators

We have not yet specified what we mean by a vector on a general manifold. In older texts, one will often see vectors introduced as N-tuples, say $v^a = (v^1, v^1, \cdots, v^N)$, that transform in a specific way under changes of coordinates. The v^a are the *coordinate components* of the vector and the operations of addition of vectors and multiplication by a scalar are defined by the corresponding operations on the components. This approach is fine, but rather hides the geometric nature of vectors.

[1] Recall that a vector space is, generally, a non-empty set of objects, called vectors, together with an associative and commutative operation of addition and an operation of scalar multiplication, which is distributive over addition. Moreover, the set must be closed under these operations, and must contain an additive zero vector, which leaves any vector unchanged under addition, and an additive inverse, which returns the zero vector when added to any vector.

An alternative approach is to think of a vector at a point P as a differential operator there, which maps scalar fields on $\mathcal{M}$ to a number. By extension, a vector *field* is associated with a differential operator at every point and maps scalar fields to scalar fields. Intuitively, it is the directionality of the differential operator that captures the idea of vectors as having an associated direction. Consider the operator

$$\mathbf{v} = v^a \frac{\partial}{\partial x^a} \tag{4.2}$$

evaluated at P, where x^a is some coordinate chart. v^a are the coordinate components of $\mathbf{v}$, and $\partial/\partial x^a \,|_P$ are the coordinate basis vectors at P for $T_P(\mathcal{M})$.

The sum of two such operators is also a differential operator, as is the result of multiplying by a scalar, so the space of all such operators at P is closed and forms a vector space. In this way, we have explicitly constructed the tangent space $T_P(\mathcal{M})$. Eq. (4.2) expresses the vector $\mathbf{v}$ as a linear combination of the real-valued N-tuple v^a and the partial derivatives along the coordinate directions.

The N partial derivative operators $\left\{ \partial/\partial x^1 \, , \cdots , \, \partial \Big/ \partial x^N \right\}$ at P can therefore be considered a set of *basis vectors* for $T_P(\mathcal{M})$, and the v^a are the associated components. If we change the coordinates to x'^a, the basis vectors will change since (by the chain rule)

$$\frac{\partial}{\partial x'^a} = \frac{\partial x^b}{\partial x'^a} \frac{\partial}{\partial x^b}, \tag{4.3}$$

where $\partial x^b \Big/ \partial x'^a$ holds the role of a transformation matrix.[2] If the vector $\mathbf{v}$ is to remain invariant, its components must transform as

$$\boxed{v'^a = \frac{\partial x'^a}{\partial x^b} v^b,} \tag{4.4}$$

since then

$$\begin{aligned}
\mathbf{v} &\to v'^a \frac{\partial}{\partial x'^a} && (4.5)\\
&= \underbrace{\frac{\partial x'^a}{\partial x^b} \frac{\partial x^c}{\partial x'^a}}_{\delta^b_c} v^b \frac{\partial}{\partial x^c} && (4.6)\\
&= v^b \frac{\partial}{\partial x^b} && (4.7)\\
&= \mathbf{v} && (4.8)
\end{aligned}$$

Note that the components v^a and the basis vectors transform inversely under changes of coordinates, and thus we see that $\mathbf{v}$ transforms as the identity $\mathbb{I}$. Any N-tuple that transforms according to Eq. (4.4) forms the components of a vector. For example, the coordinate differentials $\mathrm{d}x^a$ between two neighbouring points transform with the chain rule as

$$\mathrm{d}x'^a = \frac{\partial x'^a}{\partial x^b} \mathrm{d}x^b \, , \tag{4.9}$$

[2] Where we have $\partial x^b \big/ \partial x'^a$ as the inverse of the Jacobian $J^a{}_b$ as defined in Eq. (3.7), from which $J^a{}_b$ transforms from a coordinate system x'^a to x^b.

and so are the components of a vector (the infinitesimal "displacement" vector).

An important example of a vector is the *tangent vector* to a curve $x^a(u)$, which has components $\mathrm{d}x^a/\mathrm{d}u$; the associated vector (i.e., differential operator) is

$$\frac{\mathrm{d}x^a}{\mathrm{d}u}\frac{\partial}{\partial x^a} = \frac{\mathrm{d}}{\mathrm{d}u}. \tag{4.10}$$

Here it is useful to think of $\partial/\partial x^a$ as the basis vectors since we have defined vectors in Eq. (4.2) as the directions in which you can differentiate functions. Then it follows that $\mathrm{d}x^a/\mathrm{d}u$ as the coordinate components of the tangent vector along those basis vectors, which together combine to give $\mathrm{d}/\mathrm{d}u$, the derivative along the curve.

Finally, a word about notation: as most operations with vectors involve working with the components in some coordinate system, we shall often (rather sloppily!) write things like "the vector v^a" rather than the more correct "the vector with components v^a".

4.1.4 Dual Vector Fields

Another class of vector-like objects arises when we consider the gradient of a scalar field, i.e., N-tuples such as

$$X_a = \frac{\partial \phi}{\partial x^a}. \tag{4.11}$$

Under a change of coordinates, we have

$$X'_a = \frac{\partial \phi'}{\partial x'^a} = \frac{\partial x^b}{\partial x'^a}\frac{\partial \phi}{\partial x^b} = \frac{\partial x^b}{\partial x'^a}X_a. \tag{4.12}$$

The X^a do not transform as the components of a vector (c.f. Eq. (4.4)); rather, they transform, by construction, in the same way as the basis vectors $\partial/\partial x^a$. Objects that transform as

$$\boxed{X'_a = \frac{\partial x^b}{\partial x'^a}X_b} \tag{4.13}$$

under a coordinate transformation are called the components of a *dual vector*. (Again, this should be understood as the transformation law *at the point* P.) Given the linearity of the transformation (4.13), it is clear that objects like X'_a, with addition and multiplication by a scalar defined element-wise, form a vector space at P.

We shall see below that dual vectors at P should be considered as inhabiting a different vector space than $T_P(\mathcal{M})$, called the *dual vector space* $T^*_P(\mathcal{M})$. Dual vectors are dual to vectors in the sense that the contraction of a dual vector X_a and vector v^a, defined by the summation $X_a v^a$, is invariant under coordinate transformations:

$$X'_a v^a = \underbrace{\frac{\partial x^b}{\partial x'^a}\frac{\partial x'^a}{\partial x^c}}_{\delta^b_c} X_b v^c \tag{4.14}$$

$$= X_b v^b. \tag{4.15}$$

We have so far defined dual vectors via the transformation law of their components. However, as with vectors, we should think of dual vectors (as opposed to their components) as geometric objects that are invariant under changes of coordinates.

The way to formalise this is to regard dual vectors as linear maps that take vectors to real (or, more generally, complex) numbers. Indeed, you may already be familiar (from courses in linear algebra) with the idea of a dual vector space to a vector space, defined as the set of linear maps of vectors to real (or, more generally, complex) numbers. When expressed in terms of components, the result of the linear map between a dual vector X_a and vector v^a is just the contraction $X_a v^a$. If we introduce a basis for the dual vector space, we can write down coordinate-independent expressions for dual vectors as linear maps on $T_P(\mathcal{M})$, but we shall not need such an approach here.

If all this seems unfamiliar and opaque, it might help to recall the bra-ket notation of quantum mechanics. There, state vectors are written as $|\psi\rangle$ and are elements of a vector space. The objects $\langle\phi|$ are elements of the dual vector space and are really linear maps of state vectors $|\psi\rangle$ to (complex) numbers as $\langle\phi|\psi\rangle$.

Finally, we note that there is, in general, no invariant way to relate vectors and dual vectors, i.e., given a vector v^a we cannot construct a dual vector. An important exception is for (pseudo-)Riemannian manifolds, which are equipped with a metric. We shall see shortly that the metric naturally associates vectors and dual vectors.

4.2 Tensor Fields

Tensors are an extension of local vectors and dual vectors. At a given point P, a tensor there can be formally introduced as a multi-linear map on tensor products of $T_P(\mathcal{M})$ and $T_P^*(\mathcal{M})$ that take k dual vectors and l vectors at P as input and returns a number. Such a tensor is said to be of *type* (k, l) and to have *rank* $k + l$.

Here, we shall take the less formal route and define tensors via the transformation laws of their components. The components of a tensor of type (k, l) has k "upstairs" (sometimes called *contravariant*) indices and l "downstairs" (*covariant*) indices, e.g., T_{ab} is type $(0, 2)$ and T^{ab} is type $(2, 0)$. Note that we can also have tensors with a mix of upstairs and downstairs indices, such as $T_a{}^b$. (The reason for offsetting the indices, thus defining an order, will become clear later when we consider how the metric may be used to change the type of a tensor.) The components of a type-(k, l) tensor transform under changes of coordinates like

$$\boxed{T'^{a\cdots b}{}_{c\cdots d} = \frac{\partial x'^a}{\partial x^p}\cdots\frac{\partial x'^b}{\partial x^q}\frac{\partial x^r}{\partial x'^c}\cdots\frac{\partial x^s}{\partial x'^d}T^{p\cdots q}{}_{r\cdots s}\,.} \tag{4.16}$$

We see that rank-0 tensors are scalar fields, while type-$(1, 0)$ tensors are vectors v^a and type-$(0, 1)$ tensors are dual vectors X_a. As with vectors and dual vectors, we should think of tensors as geometric objects that are invariant under changes of coordinates (although the coordinate components do change, of course). A tensor field assigns a tensor *of the same type* to every point in the manifold. Finally, we shall sometimes want to write the

tensor itself rather than its components; generally, we shall use the same bold symbol, for example, the tensor $\mathbf{T}$ with components T_{ab}.

4.2.1 Tensor Equations

The reason that we are interested in working with tensor-valued objects is that they allow us to write down equations that are independent of any coordinate system. In particular, suppose in some coordinate system one finds the components of two tensors, T_{ab} and S_{ab}, to be equal. The tensor transformation law implies that their components are the same in *any* coordinate system, i.e., they are the same tensor. In components, the *form* of the equation $T_{ab} = S_{ab}$ is the same in all coordinate systems. Moreover, if the components of a tensor vanish in some coordinate system they vanish in all (the tensor itself vanishes).

4.2.2 Elementary Operations with Tensors

4.2.2.1 Addition and Multiplication by a Scalar

Tensors of the same type at the same point P can be added (subtracted) to give a tensor of the same type. Addition is defined in the usual way for components, and the result is denoted by, e.g., $T_{ab} + S_{ab}$. It is straightforward to check that the object with components $T_{ab} + S_{ab}$ is a tensor since under a coordinate transformation,

$$\begin{aligned} T'_{ab} + S'_{ab} &= \frac{\partial x^c}{\partial x'^a}\frac{\partial x^d}{\partial x'^b}T_{cd} + \frac{\partial x^c}{\partial x'^a}\frac{\partial x^d}{\partial x'^b}S_{cd} \\ &= \frac{\partial x^c}{\partial x'^a}\frac{\partial x^d}{\partial x'^b}(T_{cd} + S_{cd}). \end{aligned} \tag{4.17}$$

Tensors can also be multiplied by a real number c, which just multiplies each component by c, to return a tensor of the same type.

4.2.2.2 Outer (or Tensor) Product

The outer product of a type-(p, q) tensor $S^{a_1\cdots a_p}{}_{b_1\cdots b_q}$ and a type-(r, s) tensor $T^{c_1\cdots c_r}{}_{d_1\cdots d_s}$ is a type-$(p + r, q + s)$ tensor with components $S^{a_1\cdots a_p}{}_{b_1\cdots b_q}T^{c_1\cdots c_r}{}_{d_1\cdots d_s}$. If we denote the tensors themselves (the coordinate-independent objects) as $\mathbf{S}$ and $\mathbf{T}$, the outer product is denoted by $\mathbf{S} \otimes \mathbf{T}$, where

$$(\mathbf{S} \otimes \mathbf{T})^{a_1\cdots a_p}{}_{b_1\cdots b_q}{}^{c_1\cdots c_r}{}_{d_1\cdots d_s} = S^{a_1\cdots a_p}{}_{b_1\cdots b_q}T^{c_1\cdots c_r}{}_{d_1\cdots d_s}\,. \tag{4.18}$$

As an example, consider two vectors u^a and v^a and denote the outer product by T^{ab}, so that $T^{ab} = u^a v^b$. Under a change of coordinates

$$\begin{aligned} T'^{ab} &= \frac{\partial x'^a}{\partial x^c}u^c\frac{\partial x'^b}{\partial x^d}v^d \\ &= \frac{\partial x'^a}{\partial x^c}\frac{\partial x'^b}{\partial x^d}T^{cd}, \end{aligned} \tag{4.19}$$

which shows that T^{ab} is indeed a type-$(2,0)$ tensor. Note that, in general, the outer product is *not commutative*, $\mathbf{S} \otimes \mathbf{T} \neq \mathbf{T} \otimes \mathbf{S}$; for example, $u^a v^b$ does not equal $v^a u^b$ generally.

4.2.2.3 Contraction

In terms of components, the operation of contraction consists of setting an upstairs and downstairs index equal and summing. For a type-(k,l) tensor, contraction returns a type-$(k-1, l-1)$ tensor.

For example, consider $T^{ab}{}_{c}$; contracting on the second and third indices gives a new object with just one upstairs index, say $S^a \equiv T^{ab}{}_{b}$ (summation convention!). To show that S^a is indeed a vector, let us first transform $T^{ab}{}_{c}$,

$$T'^{ab}{}_{c} = \frac{\partial x'^a}{\partial x^p}\frac{\partial x'^b}{\partial x^q}\frac{\partial x^r}{\partial x'^c} T^{pq}{}_{r}\,, \tag{4.20}$$

and then take the contraction in the new coordinates to find $S'^a \equiv T'^{ab}{}_{b}$ as

$$\begin{aligned} S'^a &= \frac{\partial x'^a}{\partial x^p}\underbrace{\frac{\partial x'^b}{\partial x^q}\frac{\partial x^r}{\partial x'^c}}_{\delta^q_r} T^{pq}{}_{r} \\ &= \frac{\partial x'^a}{\partial x^p} T^{pq}{}_{q} \\ &= \frac{\partial x'^a}{\partial x^p} S^p. \end{aligned} \tag{4.21}$$

This is just the expected transformation law for a vector showing that contraction does indeed return a tensor of appropriate type. Note that the order of the indices matters when contracting – the vectors $T^{ab}{}_{b}$ and $T^{ba}{}_{b}$ are different in general.

We can combine the outer product and contraction to define a type of inner product. For example, for tensors T^{ab} and S_{ab}, if we take the outer product to form $T^{ab}S_{cd}$ and then contract on, say, the second index of $\mathbf{T}$ and the first of $\mathbf{S}$, we have the type-$(1,1)$ tensor $T^{ab}S_{bc}$. For the specific case of a vector v^a and a dual vector X_a, this composition reduces to what we previously called their contraction, i.e., the scalar $v^a X_a$.

4.2.2.4 Symmetrisation

A type-$(0,2)$ tensor S_{ab} is *symmetric* if $S_{ab} = S_{ba}$ and *antisymmetric* if $S_{ab} = -S_{ba}$. Similarly, a type-$(2,0)$ tensor T^{ab} is symmetric if $T^{ab} = T^{ba}$ and antisymmetric if $T^{ab} = -T^{ba}$.

We can always decompose a type-$(0,2)$, or type-$(2,0)$, tensor into a sum of symmetric and antisymmetric parts as

$$S_{ab} = \frac{1}{2}(S_{ab} + S_{ba}) + \frac{1}{2}(S_{ab} - S_{ba}). \tag{4.22}$$

The operation of symmetrising is usually denoted by putting round brackets around the enclosed indices:

$$S_{(ab)} \equiv \frac{1}{2}(S_{ab} + S_{ba}). \tag{4.23}$$

Antisymmetrisation is usually denoted by square brackets:

$$S_{[ab]} \equiv \frac{1}{2}(S_{ab} - S_{ba}). \tag{4.24}$$

These ideas extend to arbitrary numbers of indices; for $S_{ab\cdots c}$ we can construct totally-symmetric and totally-antisymmetric tensors as

$$\begin{aligned} S_{(ab\cdots c)} &= \frac{1}{n!}(\text{sum over all perms of } a, b, \cdots, c), \\ S_{[ab\cdots c]} &= \frac{1}{n!}(\text{alternating sum over all perms}), \end{aligned} \tag{4.25}$$

where n is the number of indices.

Here, the alternating sum denotes that a term enters with a positive sign if the permutation is even and a negative sign if it is odd. For example, for S_{abc} we have

$$S_{[abc]} = \frac{1}{6}(S_{abc} - S_{acb} + S_{cab} - S_{abc} + S_{bca} - S_{bac}). \tag{4.26}$$

We could also contruct this using the totally antisymmetric property of the Levi-Civita tensor,

$$S_{[abc]} = \frac{1}{6}\varepsilon_{ijk}S_{ijk}, \tag{4.27}$$

where it is clear that the indices $\{i, j, k\}$ take on all values from $\{a, b, c\}$.

The normalisation $1/n!$ ensures that $S_{(ab\cdots c)} = S_{ab\cdots c}$ for a totally-symmetric tensor, and similarly for a totally-antisymmetric tensor. We can also consider (anti)symmetrising on subsets of indices; for example

$$S_{(ab)c} = \frac{1}{2}(S_{abc} - S_{bac}). \tag{4.28}$$

It is straightforward to check that (anti)symmetry is a coordinate-independent notion, e.g., if the components of a tensor are symmetric in some coordinate system, they are symmetric in all. Finally, we note that it only makes sense to discuss symmetry of pairs of upstairs or downstairs indices, but not a mix of up and downstairs.

4.2.3 Quotient Theorem

Not all objects with indices are components of tensors, i.e., they may not transform correctly under changes of coordinates. A useful way to test whether a set of quantities are the components of a tensor is provided by the *quotient theorem*:

> If a set of quantities when contracted with an arbitrary tensor produces another tensor, the original set of quantities form the components of a tensor.

To illustrate the proof of the quotient theorem, suppose v^a are the components of an arbitrary vector, and we have a set of quantities $T^a{}_{bc}$ that transform under a general change of coordinates in such a way that $T^a{}_{bc}\,v^c$ transforms as the components of a type-$(1,1)$ tensor.

This means that, however $T^a{}_{bc}$ transform (to $T'^a{}_{bc}$), they do so such that

$$T'^a{}_{bc}\,v^c = \frac{\partial x'^a}{\partial x^d}\frac{\partial x^e}{\partial x'^b}T^d{}_{ef}\,v^f. \tag{4.29}$$

Since v^c is a vector, $v'^c = (\partial x'^c \big/ \partial x^f)v^f$, and, since it is arbitrary, we must have

$$\begin{aligned} T'^a{}_{bc}\frac{\partial x'^c}{\partial x^f} &= \frac{\partial x'^a}{\partial x^d}\frac{\partial x^e}{\partial x'^b}T^d{}_{ef} \\ \implies T'^a{}_{bc} &= \frac{\partial x'^a}{\partial x^d}\frac{\partial x^e}{\partial x'^b}\frac{\partial x^f}{\partial x'^c}T^d{}_{ef}\,. \end{aligned} \tag{4.30}$$

It follows that the transformation law for the quantities $T^a{}_{bc}$ must be the same as for the components of a type-$(1,2)$ tensor, and so $T^a{}_{bc}$ must be the components of such a tensor.

4.3 Metric Tensor

We previously introduced the metric functions g_{ab} on a (pseudo-)Riemannian manifold via the line element

$$\mathrm{d}s^2 = g_{ab}\,\mathrm{d}x^a\,\mathrm{d}x^b\,. \tag{4.31}$$

We argued that, at a given point, the metric functions must transform as

$$g'_{ab} = \frac{\partial x^c}{\partial x'^a}\frac{\partial x^d}{\partial x'^b}g_{cd} \tag{4.32}$$

to preserve $\mathrm{d}s^2$. This transformation law shows us that g_{ab} must be the coordinate components of a type-$(0,2)$ tensor, which we call the *metric tensor*.

In the geometric language of tensors, the metric defines a symmetric, bilinear map from pairs of vectors to real numbers. It therefore defines a natural scalar (or inner) product between vectors, $\mathbf{g}(\mathbf{u},\mathbf{v})$, where

$$\mathbf{g}(\mathbf{u},\mathbf{v}) = g_{ab}u^a v^b. \tag{4.33}$$

Note that this has the property of *linearity* in *both* arguments, it does *not* have positive-definiteness in a Pseudo-Riemannian manifold. For any orthonormal coordinate basis, the scalar product evaluates as $\mathbf{g}(\mathbf{e}_i,\mathbf{e}_j) = \delta_{ij}$. The metric provides a map between vectors and dual vectors at a point, i.e., between the tangent space $T_P(\mathcal{M})$ and its dual $T_P^*(\mathcal{M})$.

To see this, consider the object $g_{ab}v^b$, where v^a is a vector. This is contracting the outer product of the type-$(0,2)$ metric tensor and a type-$(1,0)$ tensor, which necessarily returns a type-$(0,1)$ tensor, i.e., a dual vector. It is conventional to denote the dual vector $g_{ab}v^b$ with the same kernel symbol (v) as the vector from which it is derived, so we write

$$\boxed{v_a \equiv g_{ab}v^b.} \tag{4.34}$$

The operation of mapping vectors to dual vectors by the metric tensor is often referred to as "lowering an index".

The quantities v^a and v_a are the components of distinct mathematical objects (a vector and a dual vector, respectively) but, since we shall always be working with a manifold equipped with a metric, they should be regarded as just two ways of representing the same *physical* object. Physics usually picks out the most convenient representation, e.g., a vector for the 4-velocity of a particle and a dual vector for the gradient of a scalar field, but the metric allows us to map between these freely. More generally, we can change the type of tensors (lower their indices) by contracting with the metric; for example, given a type-$(1,1)$ tensor $T^a{}_b$,

$$T_{ab} \equiv g_{ac} T^c{}_b \tag{4.35}$$

is the associated type-$(0,2)$ tensor. We can lower multiple indices with repeated application of the metric, e.g.,

$$T_{abc} \equiv g_{ap} g_{bq} T^{pq}{}_c \,. \tag{4.36}$$

4.3.1 Inverse Metric

The matrix inverse of the metric functions transforms as a type-$(2,0)$ tensor under a change of coordinates. To see this, let us denote the array formed from the inverse of the metric functions by $\left(g^{-1}\right)^{ab}$, so that

$$\left(g^{-1}\right)^{ab} g_{bc} = \delta^a_c. \tag{4.37}$$

If we transform g_{ab} and compute the inverse of these transformed components, we get a new matrix $\left(g'^{-1}\right)^{ab}$ with

$$\left(g'^{-1}\right)^{ab} = \frac{\partial x'^a}{\partial x^c} \frac{\partial x'^b}{\partial x^d} \left(g^{-1}\right)^{cd}, \tag{4.38}$$

since then

$$\begin{aligned}
\left(g'^{-1}\right)^{ab} g'_{bc} &= \frac{\partial x'^a}{\partial x^p} \left(g^{-1}\right)^{pq} \underbrace{\frac{\partial x'^b}{\partial x^q} \frac{\partial x^r}{\partial x'^b}}_{\delta^r_q} \frac{\partial x^s}{\partial x'^c} g_{rs} \\
&= \frac{\partial x'^a}{\partial x^p} \frac{\partial x^s}{\partial x'^c} \underbrace{\left(g^{-1}\right)^{pq} g_{qs}}_{\delta^p_s} \\
&= \frac{\partial x'^a}{\partial x^p} \frac{\partial x^p}{\partial x'^c} \\
&= \delta^a_c,
\end{aligned} \tag{4.39}$$

as required.

However, Eq. (4.38) is just the transformation law for a type-$(2,0)$ tensor, showing that $\left(g^{-1}\right)^{ab}$ are indeed the components of a type-$(2,0)$ tensor. It is cumbersome to write $\left(g^{-1}\right)^{ab}$ for the inverse metric; instead it is usual to write it simply as g^{ab} so that

$g^{ab}g_{bc} = \delta^a_c$. Indeed, this is consistent with our earlier idea of lowering indices with the metric tensor since lowering those on g^{ab} gives

$$g_{ac}g_{bd}g^{cd} = g_{ac}\delta^b_c = g_{ab}. \tag{4.40}$$

The inverse metric provides a map ("raising the index") from dual vectors to vectors, e.g.,

$$X^a \equiv g^{ab}X_b, \tag{4.41}$$

given a dual vector X_a. This is just the inverse of the map from vectors to dual vectors provided by the metric since lowering and then raising an index returns the original object[3]:

$$v^a \xrightarrow{\mathbf{g}} g_{ab}v^b \xrightarrow{\mathbf{g}^{-1}} g^{ac}g_{cb}v^b = v^a. \tag{4.42}$$

We can now use the metric and its inverse to lower and raise indices on general tensors, e.g., given $T^{ab}{}_c$, we define

$$T_a{}^{bc} \equiv g_{ad}g^{ce}T^{db}{}_e. \tag{4.43}$$

Note the careful positioning of the indices here: we raise and lower vertically with no horizontal shift of indices to keep track of which index was raised/lowered.[4] This is necessary to distinguish, for example, $g^{ac}T_{cb}$ from $g_{ac}T^{bc}$ – these are generally different (unless T_{ab} is symmetric).

Finally, if we raise only one index on the metric we get the components of a type-$(1,1)$ tensor and these components are the kronecker delta: $g^a{}_b = g_b{}^a = \delta^a_b$. This follows since g_{ab} and g^{ab} are inverses:

$$g^{ab}g_{bc} = \delta^a_c. \tag{4.44}$$

The tensor $g^a{}_b$ is a particularly special tensor as it is the only rank-2 tensor whose components are the same in all coordinate systems; indeed,

$$\begin{aligned} g'^a{}_b &= \frac{\partial x'^a}{\partial x^c}\frac{\partial x^d}{\partial x'^b}g^c{}_d \\ &= \frac{\partial x'^a}{\partial x^c}\frac{\partial x^c}{\partial x'^b} \\ &= \delta^a_b = g^a{}_b, \end{aligned} \tag{4.45}$$

under a change of coordinates.

4.4 Scalar Products of Vectors Revisited

We can now write the scalar product between two vectors, $\mathbf{u}$ and $\mathbf{v}$, in terms of components in the equivalent forms:

$$g_{ab}u^a v^b = g^{ab}u_a v_b = u^a v_a = u_a v^a. \tag{4.46}$$

On a strictly Riemannian manifold, $g_{ab}v^a v^b \geq 0$ for any vector $\mathbf{v}$, with $g_{ab}v^a v^b = 0$ only if $\mathbf{v} = \mathbf{0}$. On a pseudo-Riemannian manifold, these conditions are relaxed – we can have

[3] This is why we can consistently use the same kernel letter after raising and lowering indices.

[4] Note that for symmetric tensors, such as δ^a_b, this does not matter. It thus follows that $T^a{}_b = T_b{}^a = T^a_b$.

non-zero vectors (*null vectors*) v^a with $g_{ab}v^a v^b = 0$ (See the Minkowski spacetime line element in Subsection 2.4.1).

Generally, we can define the "length" of a vector $|\mathbf{v}|$ by

$$|\mathbf{v}| \equiv \left| g_{ab} v^a v^b \right|^{1/2}; \tag{4.47}$$

on a pseudo-Riemannian manifold the length of a nonzero vector can be zero.

We can also define a generalised "angle" θ between two non-null vectors $\mathbf{u}$ and $\mathbf{v}$, with

$$\cos\theta \equiv \frac{u_a v^a}{|u_b u^b|^{1/2} |v_c v^c|^{1/2}}. \tag{4.48}$$

One should be aware that on a pseudo-Riemannian manifold it is possible to have $|\cos\theta > 1|$.

We say that two vectors are *orthogonal* if their scalar product vanishes.

Chapter 5

Vector and Tensor Calculus on Manifolds

The laws of physics are differential equations involving (mostly) tensor-valued objects. We therefore need to understand how to take derivatives of vectors and tensors on a general manifold, i.e., to develop vector and tensor calculus. The issue we face is that, on a general manifold, tensors at different points inhabit separate (tangent) vector spaces and there is no unique way to compare tensors at different points.

In this topic we shall see how to construct tensor-valued *covariant derivatives* of tensors, and in so doing connect together tangent spaces at different points. We shall also look at *geodesic curves* as an important application.

5.1 Covariant Derivatives

5.1.1 Derivatives of Scalar Fields

Consider a scalar field $\phi(x)$ which is differentiable function of the coordinates x^a. We saw in the last handout that the partial derivatives $\partial\phi/\partial x^a$ form the components of a dual vector, which we call the *gradient* of ϕ, since, under a change of coordinates, $\phi'(x') = \phi(x)$ and

$$\frac{\partial\phi'}{\partial x'^a} = \frac{\partial x^b}{\partial x'^a}\frac{\partial\phi}{\partial x^b}. \tag{5.1}$$

We are used to thinking of the gradient as a vector, and we can always associate a vector by forming $g^{ab}\,\partial\phi\big/\partial x^b$.

However, the gradient is more naturally thought of as a dual vector, i.e., a linear map from vectors to real numbers. This is because the gradient maps an infinitesimal displacement – a vector with components δx^a – into the change in the function between points with coordinate separation δx^a as.

$$\delta\phi = \frac{\partial\phi}{\partial x^a}\delta x^a. \tag{5.2}$$

5.1.2 Covariant Derivatives of Tensor Fields

We want to work with derivatives that preserve the tensorial nature of the object being differentiated. In Euclidean space, this is straightforward: we work in global Cartesian coordinates and take the partial derivatives of the Cartesian components of tensors. The resulting object transforms as a Cartesian tensor under orthogonal coordinate transformations.

However, on a general manifold we cannot do this as there are no global Cartesian coordinates. Even in Euclidean space, if we want to work in a general coordinate system the partial derivatives of the components of a tensor do not transform as a tensor.

To see the problem, consider a vector field $v^a(x)$ and construct the derivative $\partial v^b \big/ \partial x^a$. Now transform to some other coordinates, x^a, in which case the vector field has components $v'^a(x')$, and take the derivative with respect to the new coordinates; we have

$$\begin{aligned}\frac{\partial v'^b}{\partial x'^a} &= \frac{\partial}{\partial x'^a}\left(\frac{\partial x'^b}{\partial x^c}v^c\right)\\ &= \frac{\partial x^d}{\partial x'^a}\frac{\partial}{\partial x^d}\left(\frac{\partial x'^b}{\partial x^c}v^c\right)\\ &= \frac{\partial x^d}{\partial x'^a}\frac{\partial x'^b}{\partial x^c}\frac{\partial v^c}{\partial x^d} + \frac{\partial x^d}{\partial x'^a}\frac{\partial^2 x'^b}{\partial x^d \partial x^c}v^c. \end{aligned} \tag{5.3}$$

The first term on the right is the usual transformation law for a type-$(1,1)$ tensor, but the second term means that $\partial v^b \big/ \partial x^a$ do *not* form the components of a tensor. To fix this problem requires the introduction of a more complicated derivative construction, called the *covariant derivative*.

The covariant derivative of a type-(k,l) tensor $T^{a_1\cdots a_k}{}_{b_1\cdots b_l}$ is a type-$(k,l+1)$ tensor, denoted by $\nabla_c T^{a_1\cdots a_k}{}_{b_1\cdots b_l}$, which satisfies the following usual properties of a derivative.

- Action on scalar fields: acting on a scalar field ϕ, the covariant derivative is simply the gradient of the scalar field, i.e.,

$$\nabla_a \phi = \frac{\partial \phi}{\partial x^a}. \tag{5.4}$$

- Linearity: for tensors $T^{a_1\cdots a_k}{}_{b_1\cdots b_l}$ and $S^{a_1\cdots a_k}{}_{b_1\cdots b_l}$ of the same type, and for constant scalars α and β, the covariant derivative of a linear combination is the linear combination of the covariant derivatives, i.e.,

$$\nabla_c\left(\alpha T^{a_1\cdots a_k}{}_{b_1\cdots b_l} + \beta S^{a_1\cdots a_k}{}_{b_1\cdots b_l}\right) = \alpha \nabla_c T^{a_1\cdots a_k}{}_{b_1\cdots b_l} + \beta \nabla_c S^{a_1\cdots a_k}{}_{b_1\cdots b_l}. \tag{5.5}$$

- Leibnitz rule: for arbitrary tensors $T^{a_1\cdots a_k}{}_{b_1\cdots b_l}$ and $S^{c_1\cdots c_m}{}_{d_1\cdots d_n}$, the covariant derivative of the outer product satisfies the product rule

$$\begin{aligned}\nabla_f\left(T^{a_1\cdots a_k}{}_{b_1\cdots b_l} S^{c_1\cdots c_m}{}_{d_1\cdots d_n}\right) &= \left(\nabla_f T^{a_1\cdots a_k}{}_{b_1\cdots b_l}\right) S^{c_1\cdots c_m}{}_{d_1\cdots d_n}\\ &\quad + T^{a_1\cdots a_k}{}_{b_1\cdots b_l}\left(\nabla_f S^{c_1\cdots c_m}{}_{d_1\cdots d_n}\right). \end{aligned} \tag{5.6}$$

5.1.3 The Connection

We shall now try and construct an appropriate covariant derivative, starting with a vector field $v^a(x)$. Recalling Eq. (5.3), our strategy is to combine $\partial v^b \big/ \partial x^a$ with an additional

piece, linear in v^a (and with v^a undifferentiated), designed to cancel the unwanted final term on the right. We write

$$\boxed{\nabla_a v^b = \frac{\partial v^b}{\partial x^a} + \Gamma^b_{ac} v^c.} \tag{5.7}$$

where the Γ^b_{ac} are called *connection coefficients* or sometimes simply *the connection*.

Note how in the final term of Eq. (5.7), the b index has moved onto the connection coefficient from the vector $\mathbf{v}$, and a new (dummy) index c is summed over. Although the connection coefficients have indices, they are *not* the components of a tensor. Instead, they must transform under a change of coordinates (to Γ'^b_{ac}) in such a way that $\nabla_a v^b$ transforms as a type-$(1,1)$ tensor.

Forming the covariant derivative in the new coordinates, we have

$$\begin{aligned}
\nabla'_a v'^b &= \frac{\partial v'^b}{\partial x'^a} + \Gamma'^b_{ac} v'^c \\
&= \frac{\partial x^d}{\partial x'^a}\frac{\partial x'^b}{\partial x^c}\frac{\partial v^c}{\partial x^d} + \frac{\partial x^d}{\partial x'^a}\frac{\partial^2 x'^b}{\partial x^d \partial x^c} v^c + \Gamma'^b_{ac}\frac{\partial x'^c}{\partial x^d} v^d \\
&= \frac{\partial x^d}{\partial x'^a}\frac{\partial x'^b}{\partial x^c}\nabla_d v^c - \frac{\partial x^d}{\partial x'^a}\frac{\partial x'^b}{\partial x^c}\Gamma^c_{de} v^e + \frac{\partial x^d}{\partial x'^a}\frac{\partial^2 x'^b}{\partial x^d \partial x^c} v^c + \Gamma'^b_{ac}\frac{\partial x'^c}{\partial x^d} v^d.
\end{aligned} \tag{5.8}$$

If $\nabla_a v^b$ are the components of a tensor, the final three terms on the right here must vanish for arbitrary $\mathbf{v}$, which requires

$$\Gamma'^b_{ac} = \frac{\partial x'^a}{\partial x^d}\frac{\partial x^e}{\partial x'^b}\frac{\partial x^f}{\partial x'^c}\Gamma^d_{ef} - \frac{\partial x^d}{\partial x'^b}\frac{\partial x^e}{\partial x'^c}\frac{\partial^2 x'^a}{\partial x^d \partial x^e}. \tag{5.9}$$

Note that the presence of the final (inhomogeneous) term on the right means that the connection coefficients do not transform as the components of a tensor.

The connection is not unique: any coefficients that satisfy Eq. (5.9) will give a valid covariant derivative. However, we shall see shortly how on a manifold with a metric, the metric naturally picks out a unique connection. Given two connections, Γ and $\tilde{\Gamma}$, which satisfy Eq. (5.9), their difference does transform as a type-$(1,2)$ tensor since the last term on the right of Eq. (5.9) cancels. This means that, generally, the connection is unique up to a type-$(1,2)$ tensor.

5.1.3.1 Extension to Other Tensor Fields

We now construct the covariant derivative for more general tensor fields. First, consider a type-$(2,0)$ tensor T^{ab}, which we can always decompose into a sum of outer products of vectors. We can consider these terms separately because of linearity of the covariant derivative, so consider $T^{ab} = u^a v^b$ for some vectors u^a and v^b. The Leibnitz rule gives

$$\begin{aligned}
\nabla_a\Big(u^a v^b\Big) &= \Big(\nabla_a u^b\Big)v^c + u^b(\nabla_a v^c) \\
&= \left(\frac{\partial u^b}{\partial x^a} + \Gamma^b_{ad} u^d\right)v^c + u^b\left(\frac{\partial v^c}{\partial x^a} + \Gamma^c_{ad} v^d\right) \\
&= \frac{\partial}{\partial x^a}\Big(u^b v^c\Big) + \Gamma^b_{ad} u^d v^c + \Gamma^c_{ad} u^b v^d,
\end{aligned} \tag{5.10}$$

so that, generally,

$$\nabla_a T^{bc} = \frac{\partial T^{bc}}{\partial x^a} + \Gamma^b_{ad} T^{dc} + \Gamma^c_{ad} T^{bd}. \tag{5.11}$$

For dual vector fields, $X_a(x)$, the covariant derivative is inherited from that for vector fields if we impose the further requirement that *the covariant derivative commutes with contraction.*

Let us think about what this means for the scalar formed from the contraction of a dual vector X_a and a vector v^a. Using, in addition, the Leibnitz rule, we have

$$\nabla_a \Big(X_b v^b \Big) = (\nabla_a X_b) v^b + X_b \Big(\nabla_a v^b \Big). \tag{5.12}$$

However, we have already specified that the covariant derivative of a scalar is the gradient, so

$$\nabla_a \Big(X_b v^b \Big) = \frac{\partial X_b}{\partial x^a} v^b + X_b \frac{\partial v^b}{\partial x^a}. \tag{5.13}$$

Comparing with Eq. (5.12), and using the expansion of the covariant derivative of a vector in terms of the connection, we are left with

$$\boxed{\nabla_a X_b = \frac{\partial X_b}{\partial x^a} - \Gamma^c_{ab} X_c.} \tag{5.14}$$

Note, in particular, the minus sign and the placement of indices on the connection term.

We can build up the covariant derivative of more general tensors now as outer products of vectors and dual vectors as needed. For example, for rank-2 tensors we have

$$\boxed{\begin{array}{rcl} \nabla_c T^{ab} &=& \partial_c T^{ab} + \Gamma^a_{cd} T^{db} + \Gamma^b_{cd} T^{ad}, \\ \nabla_c T^a{}_b &=& \partial_c T^a{}_b + \Gamma^a_{cd} T^d{}_b - \Gamma^d_{cb} T^a{}_d, \\ \nabla_c T_{ab} &=& \partial_c T_{ab} - \Gamma^d_{ca} T_{db} - \Gamma^d_{cb} T_{ad}. \end{array}} \tag{5.15}$$

Here, we have introduced a very convenient shorthand notation writing $\partial/\partial x^a$ as ∂_a; we shall use this extensively from now on.

Finally, we note that the covariant derivative of the mixed metric tensor $g^a{}_b$ vanishes since

$$\begin{aligned} \nabla_c g^a{}_b &= \partial_c \delta^a_b + \Gamma^a_{cd} \delta^d_b - \Gamma^d_{cb} \delta^a_d \\ &= \Gamma^a_{cb} - \Gamma^a_{cb} = 0, \end{aligned} \tag{5.16}$$

where we used $g^a{}_b = \delta^a_b$. This is equivalent to requiring that the covariant derivative commutes with contraction.

5.1.4 The Metric Connection

On a manifold equipped with a metric, such as the spacetime of general relativity, there is a natural connection that is singled out by the following two further conditions.

- Metric compatibility, where we enforce that the covariant derivative of the metric vanishes:

$$\boxed{\nabla_a g_{bc} = 0.} \tag{5.17}$$

- Commutative action on scalar fields, so that

$$\boxed{\nabla_a \nabla_b \phi = \nabla_b \nabla_a \phi.} \tag{5.18}$$

We shall see shortly why it is reasonable to impose these conditions. However, for the moment let us just explore their consequences.

We begin with the commutative action on scalar fields; this implies that the connection must be symmetric in its lower indices,

$$\Gamma^a_{bc} = \Gamma^a_{cb}. \tag{5.19}$$

To see this, we expand $\nabla_a \nabla_b \phi$ as

$$\nabla_a \nabla_b \phi = \partial_a \partial_b \phi - \Gamma^c_{ab} \partial_c \phi. \tag{5.20}$$

The first term on the right is symmetric in a and b, so if $\nabla_{[a} \nabla_{b]} \phi = 0$ for all ϕ, we must have $\Gamma^c_{[ab]} = 0$. More generally, the antisymmetric part of the connection transforms as a tensor (this follows from Eq. (5.9)), which is called the *torsion tensor*.

However, in general relativity we shall only be concerned with a symmetric, or torsion-free, connection so that $\nabla_{[a} \nabla_{b]} \phi = 0$. We now turn to metric compatibility:

$$0 = \nabla_c g_{ab} = \partial_c g_{ab} - \Gamma^d_{ca} g_{db} - \Gamma^d_{cb} g_{ad}. \tag{5.21}$$

If we write down the other two cyclic permutations of the indices a, b and c, we have

$$0 = \partial_b g_{ca} - \Gamma^d_{bc} g_{da} - \Gamma^d_{ba} g_{cd} \tag{5.22}$$

$$0 = \partial_a g_{bc} - \Gamma^d_{ab} g_{dc} - \Gamma^d_{ac} g_{bd} \tag{5.23}$$

Adding Eq. (5.21) and (5.23) and subtracting Eq. (5.22), and using the symmetry of the connection gives

$$2\Gamma^d_{ca} g_{db} = \partial_c g_{ab} + \partial_a g_{bc} - \partial_b g_{ca}. \tag{5.24}$$

Solving for Γ by contracting with the inverse metric, we find an explicit and unique expression for the connection coefficients[1]:

$$\boxed{\Gamma^a_{bc} = \frac{1}{2} g^{ad} (\partial_b g_{dc} + \partial_c g_{db} - \partial_d g_{bc}).} \tag{5.25}$$

This expression allows computation of the connection coefficients in an arbitrary coordinate system.

The covariant derivative of the inverse metric also has vanishing covariant derivative,

$$\nabla_a g^{bc} = 0, \tag{5.26}$$

which follows from taking the covariant derivative of $g_{ab} g^{bc} = \delta^c_a$.

[1] The coefficients of the metric connection are sometimes called *Christoffel symbols*

5.1.4.1 Other Useful Properties of the Metric Connection

Since $\nabla_a g_{bc} = 0$, we can interchange the order of raising/lowering indices and covariant differentiation, e.g.,

$$\begin{aligned}\nabla_c t^{ab} &= \nabla_c\Big(g^{bd}T^a{}_d\Big)\\ &= \Big(\nabla_c g^{bd}\Big)T^a{}_d + g^{bd}(\nabla_c T^a{}_d)\\ &= g^{bd}(\nabla_c T^a{}_d).\end{aligned} \tag{5.27}$$

Note that the (downstairs) index associated with the covariant derivative is a genuine tensor index and so can be raised with the inverse metric in the usual way, e.g.,

$$\nabla^a v^b = g^{ac}\nabla_c v^b. \tag{5.28}$$

Finally, we sometimes require the connection coefficients summed over the upper and a lower index, which we denote by Γ^a_{ab}.

We can relate this to the derivative of the (coordinate-dependent) determinant of the metric functions as follows. Since $\nabla_c g_{ab} = 0$, we have

$$\partial_c g_{ab} = \Gamma^d_{ca} g_{db} + \Gamma^d_{cb} g_{ad}, \tag{5.29}$$

which implies

$$\begin{aligned}g^{ab}\partial_c g_{ab} &= g^{ab}\Big(\Gamma^d_{ca} g_{db} + \Gamma^d_{cb} g_{ad}\Big)\\ &= 2g^{ab}g_{db}\Gamma^d_{ca}\\ &= 2\Gamma^a_{ac}.\end{aligned} \tag{5.30}$$

The contraction on the left can be written as $g^{-1}\partial_c g$, where g is the determinant of the matrix with elements given by the metric functions.

This follows from the general result (known as Jacobi's formula) for an invertible matrix $\mathbf{M}$:

$$(\det \mathbf{M})^{-1}\partial_c \det \mathbf{M} = \mathrm{Tr}\left(\mathbf{M}^{-1}\partial_c \mathbf{M}\right). \tag{5.31}$$

(A simple proof for the case of a symmetric matrix follows from taking the derivative of the result[2]

$$\ln(\det \mathbf{M}) = \mathrm{Tr}(\ln \mathbf{M}), \tag{5.32}$$

but Eq. (5.31) holds generally.)

Putting these pieces together, we get the useful result

$$\Gamma^a_{ac} = \frac{1}{2}g^{-1}\partial_c g = |g|^{-1/2}\partial_c |g|^{1/2}. \tag{5.33}$$

[2]The log of a symmetric matrix is defined by symmetrising with an orthogonal matrix $\mathbf{O}$, taking the log of the diagonal elements of the resultant, and rotating back with $\mathbf{O}^T$

5.1.5 Relation to Local Cartesian Coordinates

The covariant derivative constructed with the metric connection has the nice property that it reduces to partial differentiation in local Cartesian coordinate. Recall that at any point P, we can find local Cartesian coordinates such that

$$g_{ab}(P) = \mathrm{diag}(\pm 1, \pm 1, \cdots, \pm 1), \qquad \left.\frac{\partial g_{ab}}{\partial x^c}\right|_P = 0. \tag{5.34}$$

Since the derivative of the metric vanishes at P, the metric connection also vanishes there and, in these coordinates, the components of the covariant derivative of a tensor reduce at P to the partial derivatives of the components of the tensor. This is very important for enforcing the equivalence principle as it is straightforward to check that some law of physics, written as a tensor equation, reduces to its usual special-relativistic form in local Cartesian coordinates.

Moreover, in Euclidean space, we see that the metric-compatible covariant derivative is equivalent *everywhere* to the usual derivative employed in Euclidean tensor calculus. Indeed, in this case, one can *define* the covariant derivative of a tensor by specifying that its form in global Cartesian coordinates is simply the partial derivatives of the Cartesian components; the form in some general coordinate system then follows from the appropriate coordinate transformation of these components. This is exactly what we do when constructing expressions for derivative operations on tensors in curvilinear coordinates in Euclidean space.

5.1.5.1 Covariant Derivative in Euclidean Space

Let x^a be a global Cartesian coordinate system in Euclidean space, and x'^a some other general coordinate system. Given a vector field $\mathbf{v}$, with Cartesian components v^a, let us define the covariant derivative of $\mathbf{v}$ to be that tensor whose Cartesian components are $\partial_a v^b$. The components in the x'^a coordinates are then given by the usual transformation law for the components of a type-$(1,1)$ tensor, so

$$\nabla'_a v'^b = \frac{\partial x^c}{\partial x'^a}\frac{\partial x'^b}{\partial x^d}\frac{\partial v^d}{\partial x^c}. \tag{5.35}$$

Let us express this in terms of derivatives of the components $\mathbf{v}$ in the x'^a coordinates, using

$$\begin{aligned}\frac{\partial v^d}{\partial x^c} &= \frac{\partial}{\partial x^c}\left(\frac{\partial x^d}{\partial x'^e}v'^e\right)\\ &= \frac{\partial}{\partial x^c}\left(\frac{\partial x^d}{\partial x'^e}\right)v'^e + \frac{\partial x^d}{\partial x'^e}\frac{\partial v'^e}{\partial x^c},\end{aligned} \tag{5.36}$$

to give

$$\nabla'_a v'^b = \frac{\partial v'^b}{\partial x^a} + \underbrace{\frac{\partial^2 x^d}{\partial x'^a \partial x'^e}\frac{\partial x'^b}{\partial x^d}}_{\Gamma'^b_{ac}} v'^e. \tag{5.37}$$

We see that a connection-like term (with the connection being symmetric) naturally arises as a consequence of a non-linear coordinate transformation or, equivalently, as the basis vectors $\partial/\partial x'^a$ of the primed coordinate system having Cartesian components $\partial x^b \big/ \partial x'^a$ that depend on position.

Indeed, we can verify that the connection coefficients that appear in Eq. (5.37) are exactly the metric connection by noting that the metric in the primed coordinates is

$$g'_{ab} = \frac{\partial x^c}{\partial x'^a}\frac{\partial x^d}{\partial x'^b}\delta_{cd}, \tag{5.38}$$

and for the inverse,

$$g'^{ab} = \frac{\partial x'^a}{\partial x^c}\frac{\partial x'^b}{\partial x^d}\delta^{cd}. \tag{5.39}$$

Forming the connection coefficients from

$$\Gamma'^b_{ae} = \frac{1}{2}g'^{bf}\left(\frac{\partial g'_{ef}}{\partial x'^a} + \frac{\partial g'_{af}}{\partial x'^e} - \frac{\partial g'_{ae}}{\partial x'^f}\right) \tag{5.40}$$

gives

$$\Gamma'^b_{ae} = \frac{\partial^2 x^d}{\partial x'^a \partial x'^e}\frac{\partial x'^b}{\partial x^d}, \tag{5.41}$$

consistent with Eq. (5.37).

5.1.6 Divergence, Curl and the Laplacian

The familiar operations of taking the divergence and curl of a vector field, and the Laplacian, generalise to tensor calculus on manifolds.

The *divergence* of a vector field $\mathbf{v}$ is the scalar field $\nabla_a v^a$. It follows from Eq. (4.38) that

$$\nabla_a v^a = \partial_a v^a + \Gamma^a_{ab} v^b = |g|^{-1/2}\partial_a\Big(|g|^{1/2} v^a\Big). \tag{5.42}$$

which is often convenient.

The *curl* of a dual-vector field $\mathbf{X}$ is defined to be the antisymmetric part of its covariant derivative; it is the type-$(0,2)$ tensor

$$(\mathrm{curl}\mathbf{X})_{ab} \equiv \nabla_a X_b - \nabla_b X_a. \tag{5.43}$$

The curl is actually independent of the connection (for a symmetric connection) since

$$\begin{aligned}\nabla_a X_b - \nabla_b X_a &= \partial_a X_b - \Gamma^c_{ab} X_c - \partial_b X_a + \Gamma^c_{ba} X_c \\ &= \partial_a X_b - \partial_b X_a.\end{aligned} \tag{5.44}$$

The curl of a gradient vanishes by construction for a symmetric connection: $\nabla_{[a}\nabla_{b]}\phi = 0$.

You are used to thinking of the curl as a vector, obtained by contracting $(\text{curl }\mathbf{X})_{ab}$ with the Levi-Civita (alternating) symbol, but this does not generalise to beyond three dimensions.[3]

Finally, we generalise the Laplacian operator. Acting on a scalar field ϕ, we have

$$\nabla^2\phi \equiv \nabla_a\Big(g^{ab}\nabla_b\phi\Big) = |g|^{-1/2}\partial_a\Big(|g|^{1/2}g^{ab}\partial_b\phi\Big). \tag{5.45}$$

The Laplacian generalises to tensor fields, e.g.

$$\nabla^2 T^{ab} = g^{cd}\nabla_c\nabla_d T^{ab}. \tag{5.46}$$

5.2 Intrinsic Derivative of Vectors Along a Curve

We often need to take derivatives of tensors defined along a curve, for example, the derivative of some tenso-rvalued property of a particle with respect to proper time for the particle.

Consider a vector $\mathbf{v}(u)$ defined along a curve $x^a(u)$. The *intrinsic derivative* of $\mathbf{v}$ along the curve $x^a(u)$ is the vector [defined along $x^a(u)$] obtained by contracting the tangent vector to the curve, $\mathrm{d}x^a/\mathrm{d}u$, with the covariant derivative of $\mathbf{v}$; we write

$$\frac{\mathrm{D}v^a}{\mathrm{D}u} \equiv \frac{\mathrm{d}x^b}{\mathrm{d}u}\nabla_b v^a = \frac{\mathrm{d}x^b}{\mathrm{d}u}(\partial_b v^a + \Gamma^a_{bc}v^c). \tag{5.47}$$

Note that, since

$$\frac{\mathrm{d}x^b}{\mathrm{d}u}\frac{\partial v^a}{\partial x^b} = \frac{\mathrm{d}v^a}{\mathrm{d}u}, \tag{5.48}$$

we only require knowledge of $\mathbf{v}$ along the curve $x^a(u)$ to compute the intrinsic derivative:

$$\frac{\mathrm{D}v^a}{\mathrm{D}u} = \frac{\mathrm{d}v^a}{\mathrm{d}u} + \frac{\mathrm{d}x^b}{\mathrm{d}u}\Gamma^a_{bc}v^c. \tag{5.49}$$

Note carefully the distinction between $\mathrm{d}v^a/\mathrm{d}u$ and $\mathrm{D}v^a/\mathrm{D}u$:

- $\mathrm{d}v^a/\mathrm{d}u$ are the usual ordinary derivatives of the components of $\mathbf{v}$ with respect to u, and do not form the components of a vector;
- Dv^a/Du include the connection term and do form the components of a vector.

The intrinsic derivative can be extended to other tensor-valued objects. For example, for a type-$(1,1)$ tensor $T^a{}_b(u)$, we define the intrinsic derivative as

$$\frac{\mathrm{D}T^a{}_b}{\mathrm{D}u} = \frac{\mathrm{d}x^c}{\mathrm{d}u}\nabla_c T^a{}_b = \frac{\mathrm{d}T^a{}_b}{\mathrm{d}u} + \frac{\mathrm{d}x^c}{\mathrm{d}u}\Big(\Gamma^a_{cd}T^d{}_b - \Gamma^d_{cb}T^a{}_d\Big). \tag{5.50}$$

[3] In mathematics, the seven-dimensional cross product exists and is a bilinear operation on vectors in seven-dimensional Euclidean space. It assigns to any two vectors $\mathbf{a}$, $\mathbf{b}$ in $\mathbb{R}^7$ a vector $\mathbf{a}\times\mathbf{b}$ also in $\mathbb{R}^7$. Like the cross product in three dimensions, the seven-dimensional product is anticommutative and $\mathbf{a}\times\mathbf{b}$ is orthogonal both to $\mathbf{a}$ and to $\mathbf{b}$. Unlike in three dimensions, it does not satisfy the Jacobi identity, and while the three-dimensional cross product is unique up to a sign, there are many seven-dimensional cross products.

The seven-dimensional cross product is one way of generalising the cross product to dimensions other than three, and it is the only other bilinear product of two vectors that is vector-valued, orthogonal, and has the same magnitude as in the 3D case.

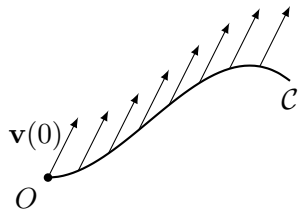

Fig. 5.1: The vector field $\mathbf{v}(u)$ defined by the parallel transport of a vector $\mathbf{v}(0)$ along a curve $\mathcal{C}$ defined in 2D Euclidean space in Cartesian coordinates by $x^a(u)$.

5.3 Parallel Transport

Consider a curve $\mathcal{C}$ defined in 2D Euclidean space in Cartesian coordinates by $x^a(u)$. At some initial point O, where $u = 0$, take a vector $\mathbf{v}(0)$ and transport it along $\mathcal{C}$ keeping its Cartesian components constant, so preserving its length and direction (see Fig. 5.1). The resulting vector field $\mathbf{v}(u)$, defined along $x^a(u)$, is said to be *parallel transported* along $x^a(u)$. In this Euclidean example, in Cartesian coordinates we have $\mathrm{d}v^a/\mathrm{d}u = 0$.

This is equivalent to the tensor equation, $\mathrm{D}v^a/\mathrm{D}u = 0$, when written in Cartesian coordinates, but the tensor equation now gives a coordinate-independent notion of parallel transport in Euclidean space. More generally, we define parallel transport on a Riemannian manifold by

$$\boxed{\frac{\mathrm{D}v^a}{\mathrm{D}u} = 0.} \tag{5.51}$$

This definition easily extends to parallel transport of other tensors, e.g., $\mathrm{D}T^{ab}/\mathrm{D}u = 0$.

5.3.1 Properties of Parallel Transport

Note the following properties of parallel transport:

- The equation $\mathrm{D}v^a/\mathrm{D}u = 0$ is an ordinary differential equation for the components v^a, and it has a unique solution if the v^a are specified at some initial point A.
- The vector obtained by parallel transporting from A to a second point B on the curve $x^a(u)$ is independent of the parameterisation used since, for an infinitesimal step, the change in the components are

$$\delta v^a = \delta u \frac{\mathrm{d}v^a}{\mathrm{d}u} = -\delta u \Gamma^a_{bc} \frac{\mathrm{d}x^b}{\mathrm{d}u} v^c = -\Gamma^a_{bc} \delta x^b v^c. \tag{5.52}$$

- The length of a vector is preserved under parallel transport since[4]

$$\frac{\mathrm{d}|\mathbf{v}|^2}{\mathrm{d}u} = \frac{\mathrm{D}}{\mathrm{D}u}\left(g_{ab} v^a v^b\right) = 2 g_{ab} v^a \frac{\mathrm{D}v^b}{\mathrm{D}u} = 0. \tag{5.53}$$

[4]The intrinsic derivative inherits the properties of the covariant derivative, such as commutativity with contraction and the Leibnitz property.

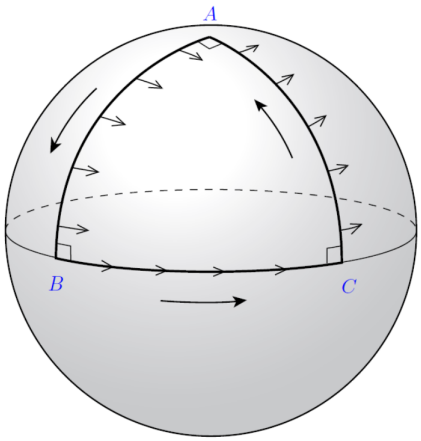

Fig. 5.2: Parallel transport around a closed path on the surface of the 2-sphere. The path consists of a great circle through the north pole (A) down to the equator at B, a length of the equator from B to C, and the great circle through C and A. The vector indicated by the small arrows is parallel transported around this path and ends up back at A rotated by $\pi/2$.

- More generally, if two vectors are parallel transported along a curve, their scalar product is constant.

Note from Eq. (5.52) how the connection, through the operation of parallel transport, allows us to *connect* vector at neighbouring points separated by coordinate increments δx^a. If we make such a step in local Cartesian coordinates, we keep the components of the vector constant.

However, we generally cannot find a global system of such coordinates and this leads to a major difference between parallel transport in Euclidean and non-Euclidean space: the latter is generally path dependent, and so the vector obtained by parallel transporting around a closed loop differs from the original vector (see Fig. 5.2). This path dependence is a measure of the intrinsic curvature of the manifold (which we shall discuss in detail later in the course).

Finally, we note that on a surface embedded in Euclidean space, parallel transport from a point A to an infinitesimally-separated point B corresponds to parallel transport in the embedding space followed by projection into the surface at B (see *General Theory of Relativity* by Dirac).

5.4 Geodesic Curves

Geodesic curves on a manifold are the generalisation of straight lines in Euclidean space. They can be defined as curves of extremal distance between two points (except in the special case of null curves; see following). Geodesics can equivalently be defined as curves $x^a(u)$ that parallel transport their tangent vector $t^a = \mathrm{d}x^a/\mathrm{d}u$, generalising the usual notion of "straight" in Euclidean space.

Geodesics are important in general relativity because, as we shall argue later, free test particles[5], including massless particles, follow geodesic curves in spacetime.

5.4.1 Tangent Vectors

We have already mentioned the idea of a tangent vector to a curve. For a curve $x^a(u)$, the tangent vector is a vector $\mathbf{t}$ with coordinate components

$$t^a = \frac{\mathrm{d}x^a}{\mathrm{d}u}. \tag{5.54}$$

Note that the tangent vector depends on the choice of parameterisation (although the tangent vectors in all parameterisations are parallel, of course).

In a pseudo-Riemannian manifold, the square of a vector, defined by $\mathbf{g}(\mathbf{t},\mathbf{t})$, is said to be timelike, spacelike or null according to

$\mathbf{g}(\mathbf{t},\mathbf{t}) > 0$	timelike;
$\mathbf{g}(\mathbf{t},\mathbf{t}) < 0$	spacelike;
$\mathbf{g}(\mathbf{t},\mathbf{t}) = 0$	null.

(5.55)

At a point, a curve is timelike, spacelike or null according to the character of its tangent vector there.

For a non-null curve, the length of the tangent vector is the derivative of the proper path length s along the curve with respect to the parameter u:

$$|\mathbf{t}| = \left| g_{ab} \frac{\mathrm{d}x^a}{\mathrm{d}u} \frac{\mathrm{d}x^b}{\mathrm{d}u} \right|^{1/2} = \left| \frac{\mathrm{d}s}{\mathrm{d}u} \right|. \tag{5.56}$$

5.4.2 Stationary Property of Non-Null Geodesics

Consider a non-null curve ${}^x a(u)$ between points A and B, with $u = 0$ at A and $u = 1$ at B. The length from A to B along the curve is

$$L = \int_A^B \mathrm{d}s = \int_0^1 \underbrace{\left| g_{ab} \dot{x}^a \dot{x}^b \right|^{1/2}}_{F} \mathrm{d}u\,, \tag{5.57}$$

where $\dot{x}^a = \mathrm{d}x^a/\mathrm{d}u$.

The form of the integrand F is invariant under reparameterisation: if we switch to some other parameter $\kappa(u)$, where $\kappa(u)$ is monotonic in the interval $0 \le u \le 1$, the length becomes

$$L = \int_{\kappa(0)}^{\kappa(1)} \left| g_{ab} \frac{\mathrm{d}x^a}{\mathrm{d}\kappa} \frac{\mathrm{d}x^b}{\mathrm{d}\kappa} \right|^{1/2} \mathrm{d}\kappa\,. \tag{5.58}$$

[5] In this context, a test particle is supposed to have sufficiently small mass that its motion does not affect the spacetime geometry

If the curve is extremal, the length is unchanged to first order for arbitrary changes in the path, $x^a(u) \to x^a(u) + \delta x^a(u)$, which have fixed endpoints.

This is a standard problem in the calculus of variations, and extremal curves satisfy the Euler–Lagrange equations

$$\frac{\partial F}{\partial x^a} = \frac{\mathrm{d}}{\mathrm{d}u}\left(\frac{\partial F}{\partial \dot{x}^a}\right). \tag{5.59}$$

For completeness, the proof of Eq. (5.59) is provided in the Appendix A. The derivatives here are

$$\begin{aligned} \frac{\partial F}{\partial x^c} &= \pm\frac{1}{2F}\partial_c g_{ab}\dot{x}^a\dot{x}^b, \\ \frac{\partial F}{\partial \dot{x}^c} &= \pm\frac{1}{F}g_{ac}\dot{x}^a, \end{aligned} \tag{5.60}$$

with the + sign for timelike curves and the − sign for spacelike. Thus, the Euler–Lagrange equations become

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(\frac{1}{F}g_{ac}\dot{x}^a\right) = \frac{1}{2F}\partial_c g_{ab}\dot{x}^a\dot{x}^b. \tag{5.61}$$

The left-hand side is

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(\frac{1}{F}g_{ac}\dot{x}^a\right) = -\frac{1}{F}\frac{\mathrm{d}F}{\mathrm{d}u}g_{ac}\dot{x}^a + \frac{1}{F}g_{ac}\ddot{x}^a + \frac{1}{F}\partial_b g_{ac}\dot{x}^a\dot{x}^b, \tag{5.62}$$

where we used $\mathrm{d}g_{ac}/\mathrm{d}u = \partial_b g_{ac}\dot{x}^b$. Moving terms around, we have

$$\begin{aligned} g_{ac}\ddot{x}^a &= \frac{1}{F}\frac{\mathrm{d}F}{\mathrm{d}u}g_{ac}\dot{x}^a - \frac{1}{2}[2\partial_b g_{ac} - \partial_c g_{ab}]\dot{x}^a\dot{x}^b \\ \implies \ddot{x}^d &= \frac{1}{F}\frac{\mathrm{d}F}{\mathrm{d}u}\dot{x}^d - \frac{1}{2}g^{dc}[\partial_a g_{bc} + \partial_b g_{ac} - \partial_c g_{ab}]\dot{x}^a\dot{x}^b \\ &= \frac{1}{F}\frac{\mathrm{d}F}{\mathrm{d}u}\dot{x}^d - \Gamma^d_{ab}\dot{x}^a\dot{x}^b. \end{aligned} \tag{5.63}$$

We find that a non-null geodesic satisfies

$$\ddot{x}^a + \Gamma^a_{bc}\dot{x}^b\dot{x}^c = \left(\frac{\ddot{s}}{\dot{s}}\right)\dot{x}^a, \tag{5.64}$$

where we have used $F = \mathrm{d}s/\mathrm{d}u$.

This is a tensor equation; this is more obvious if we write it in terms of the tangent vector $t^a = \mathrm{d}x^a/\mathrm{d}u$ since then it becomes

$$\frac{\mathrm{D}t^a}{\mathrm{D}u} = \left(\frac{\ddot{s}}{\dot{s}}\right)t^a. \tag{5.65}$$

There is a preferred class of parameters such that Eq. (5.64) simplifies to

$$\boxed{\ddot{x}^a + \Gamma^a_{bc}\dot{x}^b\dot{x}^c = 0.} \tag{5.66}$$

Such parameters have $\ddot{s} = 0$ and so are linearly related to the path length: $u = as + b$ for constants a and b. These are called *affine parameters*.

5.4.3 Relation to Parallel Transport

For a non-null geodesic in an affine parameterisation, the tangent vector $t^a = \mathrm{d}x^a/\mathrm{d}u$ is parallel transported

$$\boxed{\frac{\mathrm{D}t^a}{\mathrm{D}u} = 0.} \tag{5.67}$$

Indeed, an equivalent definition of a non-null geodesic with affine parameterisation is that it is a curve whose tangent vector is parallel transported.[6] This is all consistent with what we know in Euclidean space: there, a geodesic between two points is just the straight line connecting them, and the tangent vector is constant if we use a parameter linearly related to length along the line (i.e., an affine parameter).

Note that $\mathrm{D}t^a/\mathrm{D}u = 0$ means that the length of the tangent vector is constant, which makes sense as $|\mathbf{t}| = \mathrm{d}s/\mathrm{d}u$ and is constant for an affine parameter.

For null curves, we cannot use the stationary property to define geodesics since the path length vanishes. Instead, we define null geodesics as curves with null tangent vector satisfying Eq. (5.67).

In all cases, if we pick a vector at some starting point, and then solve $\mathrm{D}t^a/\mathrm{D}u = 0$ and $t^a = \mathrm{d}x^a/\mathrm{d}u$, we generate a unique geodesic curve in an affine parameterisation that is everywhere timelike, spacelike or null according to the character of the initial vector. This follows since parallel transport preserves $g_{ab}t^at^b$.

5.4.4 Alternative "Lagrangian" Procedure

There is an alternative Lagrangian procedure to generate the equations for an affinely-parameterised geodesic. Consider lowering the index on the equation of parallel transport for the tangent vector t^a of a geodesic in an affine parameterisation:

$$g_{ab}\frac{\mathrm{D}t^a}{\mathrm{D}u} = 0 \implies \frac{\mathrm{D}t_a}{\mathrm{D}u} = \frac{\mathrm{d}t_a}{\mathrm{d}u} - \Gamma^c_{ba}t^bt_c = 0. \tag{5.68}$$

Using the explicit form for the metric connection gives

$$\frac{\mathrm{d}t_a}{\mathrm{d}u} - \frac{1}{2}g^{cd}(\partial_b g_{ad} + \partial_a g_{bd} - \partial_d g_{ab})t^bt_c = 0. \tag{5.69}$$

The first and third terms in brackets cancel to leave the following useful alternative form of the geodesic equation:

$$\boxed{\frac{\mathrm{d}t_a}{\mathrm{d}u} = \frac{1}{2}\partial_a g_{bc}t^bt^c.} \tag{5.70}$$

As $t_a = g_{ab}\,\mathrm{d}x^b\big/\mathrm{d}u$, we have

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(g_{ab}\frac{\mathrm{d}x^b}{\mathrm{d}u}\right) = \frac{1}{2}\frac{\partial g_{bc}}{\partial x^a}\frac{\mathrm{d}x^b}{\mathrm{d}u}\frac{\mathrm{d}x^c}{\mathrm{d}u}. \tag{5.71}$$

[6] For connections more general than the metric connection, such *auto-parallel curves* are generally non-geodesic.

This is exactly the Euler–Lagrange equation,

$$\frac{\partial L}{\partial x^a} = \frac{\mathrm{d}}{\mathrm{d}u}\left(\frac{\partial L}{\partial \dot{x}^a}\right), \tag{5.72}$$

which would follow from the "Lagrangian"

$$L = g_{ab}\frac{\mathrm{d}x^a}{\mathrm{d}u}\frac{\mathrm{d}x^b}{\mathrm{d}u}. \tag{5.73}$$

This route through to the geodesic equations in an affine parameterisation is often very convenient as it avoids us having to compute the metric connection directly. We shall make use of this later when discussing motion around spherical masses.

5.4.5 Conserved Quantities Along Geodesics

For an affinely-parameterised geodesic, the tangent vector $\mathbf{t}$ is parallel transported so $|\mathbf{t}|$ is constant. For a non-null geodesic, we can always take $|\mathbf{t}| = 1$ by taking $u = s$, where, recall, s is path length along the curve. For a null geodesic, we have $|\mathbf{t}| = 0$.

The constraint

$$g_{ab}\frac{\mathrm{d}x^a}{\mathrm{d}u}\frac{\mathrm{d}x^b}{\mathrm{d}u} = \text{const.} \tag{5.74}$$

is a very useful first-integral of the geodesic equation. This first integral follows directly from the alternative Lagrangian approach by noting that L does not depend explicitly on u (in the same way that energy conservation arises in classical mechanics when the Lagrangian has no explicit time dependence).

Further conserved quantities arise when the manifold has special symmetries. In particular, from Eq. (5.70) we see that

$$\boxed{\partial_c g_{ab} = 0 \quad \Longrightarrow \quad t_c = \text{const.}} \tag{5.75}$$

In words, if the metric does not depend on a coordinate x^c, then the cth component of the tangent (dual) vector is conserved along an affinely-parameterised geodesic. This also follows directly from the alternative Lagrangian route as conservation of the *conjugate momentum*, $\pi_c = \partial L/\partial \dot{x}^c$, if the Lagrangian does not depend on x^c, i.e., $\partial_c g_{ab} = 0$.

Chapter 6

Minkowski Spacetime and Particle Dynamics

Now that we have the machinery of tensor algebra and calculus in place, in this topic we shall first apply this to special relativity and consider how to express this theory in a more formal manner. We shall also develop the theory of relativistic mechanics, which is best expressed in terms of 4D vectors in spacetime ("4-vectors").

The spacetime of special relativity is a pseudo-Euclidean manifold, over which we can globally define Cartesian coordinates. Most of our treatment of special relativity will make use of such coordinates, which correspond to the coordinates of inertial frames. However, by expressing our equations in tensor form, we can easily write them in arbitrary coordinates; we shall illustrate this with the specific example of a rotating frame of reference.

6.1 Minkowski Spacetime in Cartesian Coordinates

Minkowski spacetime is a 4D pseudo-Euclidean manifold. We can therefore adopt a global system of Cartesian coordinates x^μ ($\mu = 0, 1, 2, 3$) such that the line element is everywhere

$$\mathrm{d}s^2 = \eta_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu\,, \tag{6.1}$$

where $\eta_{\mu\nu} = \mathrm{diag}(+1, -1, -1, -1)$ is the *Minkowski metric*. Note that for applications to spacetime, we shall usually use Greek coordinate labels rather than the Roman a, b, c etc. that we have used so far on general manifolds, and allow them to run from $0 - 3$ rather than $1 - N$.

These Cartesian coordinates correspond to the coordinates (ct, x, y, z) as defined by some inertial frame, with

$$x^0 = ct, \quad x^1 = x, \quad x^2 = y, \quad x^3 = z. \tag{6.2}$$

The components of the inverse metric in Cartesian coordinates are denoted $\eta^{\mu\nu}$ and are simply

$$\eta^{\mu\nu} = \mathrm{diag}(+1, -1, -1, -1). \tag{6.3}$$

As the components of the metric are constant, the metric connection vanishes in Cartesian coordinates: $\Gamma^\mu_{\nu\sigma} = 0$.

6.1.1 Lorentz Transformations

Physically, Lorentz transformations relate Cartesian coordinates assigned to events (spacetime points) in different inertial frames. Mathematically, they correspond to the residual

freedom in our choice of global Cartesian coordinates in Minkowski spacetime, i.e., to coordinate transformations $x^\mu \to x'^\mu$ that leave the Minkowski metric unchanged:

$$\eta_{\mu\nu} = \frac{\partial x^\rho}{\partial x'^\mu}\frac{\partial x^\sigma}{\partial x'^\nu}\eta_{\rho\sigma}. \tag{6.4}$$

Multiplying through by the inverse of the transformation matrix twice we have the equivalent requirement for a Lorentz transformation,

$$\boxed{\eta_{\mu\nu} = \frac{\partial x'^\rho}{\partial x^\mu}\frac{\partial x'^\sigma}{\partial x^\nu}\eta_{\rho\sigma}.} \tag{6.5}$$

By differentiating this condition, it can be shown[1] that Lorentz transformations must be linear:

$$x'^\mu = \Lambda^\mu{}_\nu x^\nu + a^\mu, \tag{6.6}$$

for a suitable constant $\Lambda^\mu{}_\nu$, with

$$\eta_{\mu\nu} = \Lambda^\rho{}_\mu \Lambda^\sigma{}_\nu \eta_{\rho\sigma}, \tag{6.7}$$

and constant a^μ. Eq. (6.6) is known as an *inhomogeneous Lorentz transformation* or *Poincare transformation*. The constant a^μ just corresponds to changing the spacetime origin; dropping this term gives what are called *homogeneous Lorentz transformations*.

6.1.2 Homogeneous Lorentz Transformations

The constants $\Lambda^\mu{}_\nu$ of a homogeneous Lorentz transformation depend on the relative velocity and orientation of the two inertial frames. Their form for a standard Lorentz boost with speed $v = \beta c$ along the x-axis is

$$\Lambda^\mu{}_\nu = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{6.8}$$

where $\gamma = (1-\beta^2)^{-1/2}$.

The inverse of the transformation matrix is denoted by $(\Lambda^{-1})^\mu{}_\nu$ and is given by

$$\left(\Lambda^{-1}\right)^\mu{}_\nu = \frac{\partial x^\mu}{\partial x'^\nu}. \tag{6.9}$$

The inverse can be found from Eq. (6.7) to be

$$\left(\Lambda^{-1}\right)^\mu{}_\nu = \eta^{\mu\rho}\eta_{\nu\sigma}\Lambda^\sigma{}_\rho. \tag{6.10}$$

The notation $(\Lambda^{-1})^\mu{}_\nu$ is cumbersome so it is usual to define a new matrix $\Lambda_\mu{}^\nu$ (note the index positioning!) with

$$\Lambda_\mu{}^\nu = \left(\Lambda^{-1}\right)^\nu{}_\mu = \eta_{\mu\rho}\eta^{\nu\sigma}\Lambda^\rho{}_\sigma. \tag{6.11}$$

[1] see e.g., Chapter 2 of Weinberg's *Gravitation and Cosmology*.

Here, we are using the same kernel letter (Λ) to denote two different matrices, with these being distinguished by their index positions.

Note that Eq. (6.11) looks like raising and lowering indices on $\Lambda^{\mu}{}_{\nu}$ with the Minkowski metric and this is the motivation for the notation $\Lambda_{\mu}{}^{\nu}$. However, the transformation matrix $\Lambda^{\mu}{}_{\nu}$ does not contain the components of a tensor so the similarity with raising and lowering indices on tensors is really just a useful mnemonic.

6.1.3 Proper Lorentz Transformations

Proper Lorentz transformations form a subgroup of the full Lorentz transformations that only include transformations between inertial frames with the same spatial handedness and exclude time reversal. The defining condition (6.7) of Lorentz transformations gives

$$[\det\Lambda^{\mu}{}_{\nu}]^2 = 1. \tag{6.12}$$

Moreover, setting $\mu = \nu = 0$ in Eq. (6.7) gives

$$\left(\Lambda^0{}_0\right)^2 = 1 + + \sum_{i=1}^{3} \left(\Lambda^i{}_0\right)^2 \geq 1. \tag{6.13}$$

Mathematically, the subgroup of proper Lorentz transformations have

$$\boxed{\det(\Lambda^{\mu}{}_{\nu}) = 1, \qquad \Lambda^0{}_0 \geq 1,} \tag{6.14}$$

and these transformations are continuously connected to the identity. From now on, we shall generally only consider such proper Lorentz transformations.

6.1.4 Cartesian Basis Vectors

Recall that on a general manifold, a coordinate system x^a provides a set of basis vectors $\partial/\partial x^a$ that span the tangent space at any point. Since the basis vectors are differential operators corresponding to partial differentiation with respect to the coordinates, we often represent $\partial/\partial x^a$ in a diagram as an arrow tangent to the associated coordinate curves.

Recall also that the scalar product between two vectors, $\mathbf{u}$ and $\mathbf{v}$, is $\mathbf{g}(\mathbf{u}, \mathbf{v})$ or, in components, $g_{ab}u^a v^b$. If we take $\mathbf{u}$ and $\mathbf{v}$ to be the basis vectors $\partial/\partial x^a$ and $\partial\Big/\partial x^b$ of some coordinate system, then their scalar product is just the appropriate component of the metric in those coordinates, g_{ab}.

In Minkowski space, the global Cartesian coordinates x^μ associated with some inertial frame define a set of basis vectors $\partial/\partial x^\mu$ that we shall write as $\mathbf{e}_\mu$, i.e.,

$$\mathbf{e}_\mu \equiv \frac{\partial}{\partial x^\mu}. \tag{6.15}$$

These basis vectors are orthonormal since

$$\mathbf{g}(\mathbf{e}_\mu, \mathbf{e}_\nu) = \eta_{\mu\nu}. \tag{6.16}$$

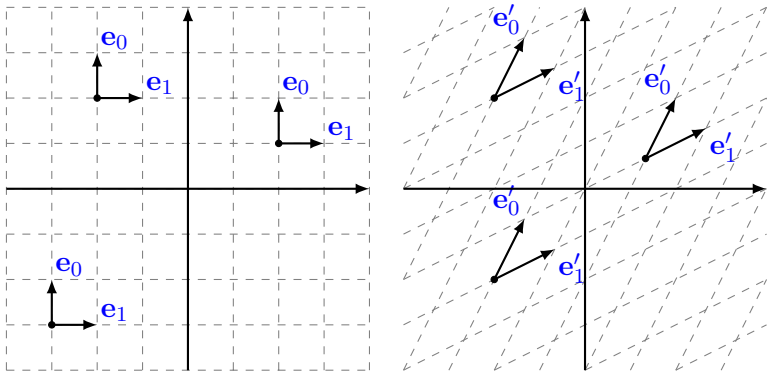

Fig. 6.1: Coordinate curves for two systems of coordinates x^μ and x'^μ, corresponding to Cartesian inertial frames S and S' in standard configuration. The coordinate basis vectors for each system are also shown, indicated as arrows tangent to the coordinate curves. The 2− and 3− directions are suppressed and null vectors would lie at 45° to the vertical.

If we change coordinates, we generate a new set of basis vectors $\partial/\partial x'^a$ related to the basis vectors in the x^a coordinates by

$$\frac{\partial}{\partial x'^a} = \frac{\partial x^b}{\partial x'^a}\frac{\partial}{\partial x^b}. \tag{6.17}$$

Applying this to a Lorentz transformation in spacetime, we have

$$\mathbf{e}'_\mu = \Lambda_\mu{}^\nu \mathbf{e}_\nu, \tag{6.18}$$

where $\mathbf{e}'_\mu \equiv \partial/\partial x'^\mu$.

Note that the basis vectors transform with the inverse transformation matrix. Since we are making a Lorentz transformation, the components of the metric in the transformed coordinates are still $\eta_{\mu\nu}$ and the new basis vectors are still orthonormal. These ideas are illustrated in Fig. 6.1.

6.1.5 4-Vectors and the Lightcone

Vectors in 4D spacetime are usually referred to as $4-$*vectors*. As usual, a vector at a point P can be decomposed into components relative to a basis there, for example,

$$\mathbf{v} = v^\mu \mathbf{e}_\mu, \tag{6.19}$$

where v^μ are the components of the vector.

Under a Lorentz transformation, the coordinate components of a vector (i.e., the components relative to the coordinate basis vectors) transform as

$$v'^\mu = \Lambda^\mu{}_\nu v^\nu. \tag{6.20}$$

A vector $\mathbf{v}$ is timelike, spacelike, or null according to the character of $\mathbf{g}(\mathbf{v},\mathbf{v})$; in Cartesian coordinates

$$\begin{array}{ll} \eta_{\mu\nu}v^\mu v^\nu > 0 & \text{timelike,} \\ \eta_{\mu\nu}v^\mu v^\nu < 0 & \text{spacelike,} \\ \eta_{\mu\nu}v^\mu v^\nu = 0 & \text{null.} \end{array} \tag{6.21}$$

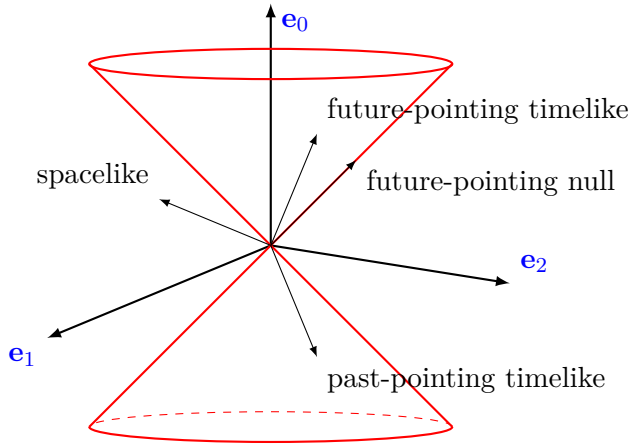

Fig. 6.2: A vector $\mathbf{v}$ is timelike, spacelike, or null according to the character of $\mathbf{g}(\mathbf{v},\mathbf{v})$; in Cartesian coordinates $\mathbf{v}$ is timelike for $\eta_{\mu\nu}v^\mu v^\nu > 0$; spacelike for $\eta_{\mu\nu}v^\mu v^\nu < 0$; and null for $\eta_{\mu\nu}v^\mu v^\nu = 0$. A timelike or null vector is future pointing if $v^0 > 0$, and past pointing if $v^0 < 0$.

For the basis vectors in an inertial frame, $\mathbf{e}_0$ is timelike, while $\mathbf{e}_i$ $(i = 1, 2, 3)$ are spacelike.

A timelike or null vector is *future pointing* if $v^0 > 0$, and *past pointing* if $v^0 < 0$. Note that the future- and past-pointing characterisations are invariant under proper Lorentz transformations (the proof is the same as the proof given in Chapter 2 that the temporal ordering of causally-connected events is Lorentz invariant).

At any point P, the set of all null vectors there define the lightcone and this separates timelike and spacelike vectors (see Fig. 6.2). To every vector we can associate a dual vector by mapping with the metric. In Cartesian coordinates, the components of the dual vector associated with the vector v^μ are

$$v_\mu = \eta_{\mu\nu}v^\nu, \tag{6.22}$$

which leaves the 0-component unchanged but reverses the spatial components. Under a Lorentz transformation, the components of a dual vector transform with the inverse transformation matrix, i.e.,

$$X'_\mu = \Lambda_\mu{}^\nu X_\nu. \tag{6.23}$$

6.2 Particle Dynamics

6.2.1 4-Velocity of a Massive Particle

A massive particle follows a trajectory through spacetime that is usually called a *wordline*. A convenient way to parameterise the wordline is with the *proper time* of the particle, τ. Recall that proper time is the time measured by an ideal clock carried by the particle, and is related to the invariant path length by $\mathrm{d}s^2 = c^2\,\mathrm{d}\tau^2$. This means that τ is an affine parameter for the worldline.

The tangent vector to the wordline is the 4-velocity of the particle, and has components

$$u^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}. \tag{6.24}$$

For a massive particle, the 4-velocity is future-pointing and timelike.

Since proper time is an affine parameter, the length of the 4-velocity is constant:

$$\eta_{\mu\nu}u^\mu u^\nu = \left(\frac{\mathrm{d}s}{\mathrm{d}\tau}\right)^2 = c^2. \tag{6.25}$$

Writing out the Cartesian components of u^μ, we have

$$\begin{aligned} u^\mu &= \left(c\frac{\mathrm{d}t}{\mathrm{d}\tau}, \frac{\mathrm{d}x}{\mathrm{d}\tau}, \frac{\mathrm{d}y}{\mathrm{d}\tau}, \frac{\mathrm{d}z}{\mathrm{d}\tau}\right) \\ &= \frac{\mathrm{d}t}{\mathrm{d}\tau}\left(c, \frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t}, \frac{\mathrm{d}z}{\mathrm{d}t}\right), \end{aligned} \tag{6.26}$$

which involves the components of the usual 3-velocity of particle, $\mathrm{d}x/\mathrm{d}t$, $\mathrm{d}y/\mathrm{d}t$ and $\mathrm{d}z/\mathrm{d}t$.

With a slight abuse of notation[2], let us write the components of the 3-velocity as $\vec{u}^i = \mathrm{d}x^i/\mathrm{d}t$ and $\vec{u} = (\vec{u}^1, \vec{u}^2, \vec{u}^3)$, so that, compactly,

$$u^\mu = \frac{\mathrm{d}t}{\mathrm{d}\tau}(c, \vec{u}). \tag{6.27}$$

The relation between coordinate and proper time is fixed by the normalisation of the 4-velocity:

$$\begin{aligned} c^2 &= \eta^\mu_{\mu\nu}u^\nu \\ &= \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2\left(c^2 - |\vec{u}|^2\right), \end{aligned} \tag{6.28}$$

so that

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \left(1 - \frac{|\vec{u}|^2}{c^2}\right)^{-1/2} = \gamma_u, \tag{6.29}$$

where we have introduced the Lorentz factor γ_u.

6.2.1.1 Velocity Transformation Laws

The transformation laws for the 3-velocity of a particle (already derived in Chapter 2 directly from the differentials of the Lorentz transformations) can now be derived simply from the transformation of the components of the 4-velocity:

$$u'^\mu = \Lambda^\mu{}_\nu u^\nu. \tag{6.30}$$

[2]This is not ideal, but at least it has the virtue of distinguishing, say, the 1-component of the 4-velocity, $u^1 = \mathrm{d}x/\mathrm{d}\tau$, from the x- or 1-component of the 3-velocity, $\vec{u}^1 = \mathrm{d}x/\mathrm{d}t$.

Let x^μ and x'^μ correspond to inertial frames S and S', respectively, related by a Lorentz boost with speed $v = \beta c$ along the x-direction, so that

$$\begin{pmatrix} \gamma_{u'} c \\ \gamma_{u'} \vec{u}'^1 \\ \gamma_{u'} \vec{u}'^2 \\ \gamma_{u'} \vec{u}'^3 \end{pmatrix} = \begin{pmatrix} \gamma_v & -\beta\gamma_v & 0 & 0 \\ -\beta\gamma_v & \gamma_v & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \gamma_u c \\ \gamma_u \vec{u}^1 \\ \gamma_u \vec{u}^2 \\ \gamma_{u'} \vec{u}^3 \end{pmatrix}. \tag{6.31}$$

The first component relates the particle's Lorentz factor in the two frames:

$$\frac{\gamma_u}{\gamma_{u'}} = \frac{1}{\gamma_v} \frac{1}{(1 - \vec{u}^1 v/c^2)}. \tag{6.32}$$

Combining this with the other components gives the usual results

$$\begin{aligned} \vec{u}'^1 &= \frac{(\vec{u}^1 - v)}{(1 - \vec{u}^1 v/c^2)}, \\ \vec{u}'^2 &= \frac{\vec{u}^2}{\gamma_v (1 - \vec{u}^1 v/c^2)}, \\ \vec{u}'^3 &= \frac{\vec{u}^3)}{\gamma_v (1 - \vec{u}^1 v/c^2)}. \end{aligned} \tag{6.33}$$

6.2.2 4-Acceleration

In an inertial frame, a free particle has $\mathrm{d}^2 x^i/\mathrm{d}t^2 = 0$, so that $\vec{u} = \text{const.}$ and $\gamma_u = \text{const.}$. It follows that the components of the 4-velocity are also constant in Cartesian coordinates so

$$\frac{\mathrm{d}u^\mu}{\mathrm{d}\tau} = 0. \tag{6.34}$$

This equation is not a tensor equation but we can easily find a tensor equation (and so one that is valid in all coordinate systems) by replacing the derivative with the intrinsic derivative $D/D\tau$ along the particle's wordline since, in global Cartesian coordinates, the metric connection vanishes:

$$\frac{\mathrm{D}u^\mu}{\mathrm{D}\tau} = 0. \tag{6.35}$$

Since u^μ is the tangent vector to the wordline in an affine parameterisation, we see that *free massive particles move on timelike geodesics in Minkowski space.*

For a particle acted on by external forces (note that we are not considering gravity yet!), the particle will accelerate and so we define the *acceleration 4-vector* by

$$a^\mu = \frac{\mathrm{D}u^\mu}{\mathrm{D}\tau}. \tag{6.36}$$

In Cartesian coordinates, this reduces to $a^\mu = \mathrm{d}u^\mu/\mathrm{d}\tau$. The acceleration 4-vector is always orthogonal to the 4-velocity: in Cartesian inertial coordinates

$$\eta_{\mu\nu} a^\mu u^\nu = \eta_{\mu\nu} \frac{\mathrm{d}u^\mu}{\mathrm{d}\tau} u^\nu = \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}\tau} (\eta_{\mu\nu} u^\mu u^\nu) = 0, \tag{6.37}$$

so, generally $\mathbf{g}(\mathbf{a}, \mathbf{u}) = 0$.

The components of **a** may be related to the usual 3-acceleration of the particle in an inertial frame as follows. Writing $u^\mu = \gamma_u(c, \vec{u})$, we have

$$a^\mu = \frac{\mathrm{d}u^\mu}{\mathrm{d}\tau} = \gamma_u \frac{\mathrm{d}}{\mathrm{d}t}(\gamma_u c, \gamma_u \vec{u}). \tag{6.38}$$

The derivative of the Lorentz factor is

$$\frac{\mathrm{d}\gamma_u}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(1 - \frac{\vec{u}\cdot\vec{u}}{c^2}\right)^{-1/2} = \frac{\gamma^3}{c^2}\vec{u}\cdot\vec{a}, \tag{6.39}$$

where $\vec{a} = \frac{\mathrm{d}\vec{u}}{\mathrm{d}t}$ is the usual 3-acceleration in the inertial frame. It follows that

$$a^\mu = \gamma_u^2\left(\frac{\gamma_u^2}{c}\vec{u}\cdot\vec{a}, \vec{a} + \frac{\gamma_u^2}{c^2}(\vec{u}\cdot\vec{a})\vec{u}\right). \tag{6.40}$$

In the instantaneous rest frame of the particle, $\vec{u} = \vec{0}$, and the components of the 4-acceleration in that frame are simply $a^\mu = (0, \vec{a}_{\mathrm{IRF}})$, where $\vec{a}_{\mathrm{IRF}}$ is the 3-acceleration in the instantaneous rest frame. Note that the magnitude of $\vec{a}_{\mathrm{IRF}}$ determines the (invariant) magnitude of the 4-acceleration:

$$|\mathbf{a}|^2 = -|\vec{a}_{\mathrm{IRF}}|^2, \tag{6.41}$$

which shows that the 4-acceleration is a spacelike vector.

6.2.3 Relativistic Mechanics of Massive Particles

The 4-momentum of a massive particle of rest mass m is the future-pointing, timelike 4-vector

$$\mathbf{p} = m\mathbf{u}. \tag{6.42}$$

At any point along the worldline of the particle, the (squared) magnitude of the 4-momentum is

$$|\mathbf{p}|^2 = m^2c^2. \tag{6.43}$$

In some inertial frame, the components of **p** are

$$p^\mu = (\gamma_u mc, \gamma_u m\vec{u}). \tag{6.44}$$

In previous courses, you will have seen that the correct relativistic generalisation of the 3-momentum of a massive point particle is

$$\vec{p} = \gamma_u m\vec{u}, \tag{6.45}$$

so the spatial components of **p** are simply the 3-momentum. Recall that this relativistic definition of the 3-momentum ensures the following are true:

1. The 3-momentum reduces to the usual non-relativistic limit $\vec{p} \approx m\vec{u}$ for $|\vec{u}| \ll c$.
2. For a free particle, $\vec{p}$ is constant since $\vec{u}$ is.

3. For a system of point particles interacting through short-range ("contact") interactions, the sum of the individual 3-momenta of all particles is conserved.
4. Newton's second law takes the form $\vec{f} = \mathrm{d}\vec{p}/\mathrm{d}t$, where $\vec{f}$ is the 3-force acting on the particle.

The time component of the 4-momentum is the total energy E of the particle (i.e., the sum of the rest-mass energy and kinetic energy):

$$E = \gamma_u m c^2. \tag{6.46}$$

To see this, consider the rate of working $\vec{f} \cdot \vec{u}$ of the force accelerating a particle:

$$\begin{aligned}
\vec{u} \cdot \vec{f} &= \vec{u} \cdot \frac{\mathrm{d}\vec{p}}{\mathrm{d}t} \\
&= \vec{u} \cdot \frac{\mathrm{d}}{\mathrm{d}t}(\gamma_u m \vec{u}) \\
&= \gamma_u m \left(\vec{u} \cdot \vec{a} + \gamma_u^2 \vec{u} \cdot \vec{a} \frac{|\vec{u}|^2}{c^2} \right) \\
&= \gamma_u^3 m \vec{u} \cdot \vec{a} \\
&= m c^2 \frac{\mathrm{d}\gamma_u}{\mathrm{d}t}.
\end{aligned} \tag{6.47}$$

With $E = \gamma_u m c^2$, the rate of working by the force is therefore $\mathrm{d}E/\mathrm{d}t$ as required.

We can now write the components of the 4-momentum in an inertial frame as

$$p^\mu = (E/c, \vec{p}). \tag{6.48}$$

Forming the invariant $|\mathbf{p}|^2$ in an inertial frame, we find the *energy–momentum invariant*

$$E^2 - |\vec{p}|^2 c^2 = m^2 c^4. \tag{6.49}$$

For a free particle, the total 4-momentum is constant, i.e., $\mathrm{d}p^\mu/\mathrm{d}\tau = 0$ in the coordinates of an inertial frame or, generally,

$$\frac{\mathrm{D}p^\mu}{\mathrm{D}\tau} = 0. \tag{6.50}$$

For an isolated system of particles undergoing collisional interactions, the total 4-momentum is the sum of the individual 4-momenta[3] and is constant; this combines *both* conservation of 3-momentum *and* energy into a Lorentz-invariant (i.e., 4-vector) law.

6.2.3.1 Force 4-Vector

For a particle acted on by a force, the 4-momentum is not constant. We can always introduce a 4-vector quantity called the 4-*force* or force 4-vector, $\mathbf{f}$, by

$$\frac{\mathrm{D}p^\mu}{\mathrm{D}\tau} = f^\mu. \tag{6.51}$$

[3] As Minkowski space is pseudo-Euclidean, we can define addition of 4-vectors at different events by addition of the components in any set of global Cartesian coordinates.

since $|\mathbf{p}|^2 = m^2c^2$ is constant, p^μ is orthogonal to $Dp^\mu/D\tau$ and so the 4-velocity and 4-force are necessarily orthogonal:

$$\mathbf{g}(\mathbf{f}, \mathbf{u}) = 0. \tag{6.52}$$

In some inertial frame,

$$f^\mu = \gamma_u \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{E}{c}, \vec{p}\right) = \gamma_u \left(\frac{\vec{f}\cdot\vec{u}}{c}, \vec{f}\right), \tag{6.53}$$

where we have used $\mathrm{d}E/\mathrm{d}t = \vec{f}\cdot\vec{u}$. Writing the components of the 4-force in the form on the right of Eq. (6.53) makes it clear that $\eta_{\mu\nu} f^\mu u^\nu = 0$. Finally, note that the 4-force can be related to the 4-acceleration via $\mathbf{f} = m\mathbf{a}$.

6.2.4 4-Momentum of a Photon

For a particle with zero rest mass, such as a photon, the energy and 3-momentum still assemble into a 4-vector with components $p^\mu = (E/c, \vec{p})$. This has to be the case if 4-momentum is to be conserved in scattering events involving photons and (charged) particles. However, for zero rest mass the limit of Eq. (6.49) gives

$$E = |\vec{p}|c, \tag{6.54}$$

and so the 4-momentum is a (future-pointing) null vector:

$$\mathbf{g}(\mathbf{p}, \mathbf{p}) = 0. \tag{6.55}$$

For a free particle, the 4-momentum is conserved as in the massive case.

If we write the photon worldline as $x^\mu(\lambda)$ for some arbitrary parameter λ, then

$$\frac{\mathrm{D}p^\mu}{\mathrm{D}\lambda} = 0. \tag{6.56}$$

The photon path is null, since photons travel at the speed of light, so we cannot use the proper time τ as a parameter ($\mathrm{d}\tau = 0$). However, we can always adopt a (dimensional) parameterisation such that

$$p^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}, \tag{6.57}$$

i.e., the tangent vector to the path is the 4-momentum.

Eq. (6.56) then tells us that $x^\mu(\lambda)$ is an affinely-parameterised null geodesic. Free massless particles move on null geodesics in Minkowski space, with $p^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$ for some affine parameterisation. To see why we can take $p^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$, we note[4] that in an inertial frame

$$\begin{aligned} p^\mu &= \frac{E}{c}\left(1, \frac{\vec{p}}{|\vec{p}|}\right) \\ &= \frac{E}{c^2}\left(c\frac{\mathrm{d}t}{\mathrm{d}t}, \frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t}, \frac{\mathrm{d}z}{\mathrm{d}t}\right) \\ &= \frac{E}{c^2}\frac{\mathrm{d}x^\mu}{\mathrm{d}t}. \end{aligned} \tag{6.58}$$

[4] This also shows that for neighbouring events separated by time $\mathrm{d}t$ on the worldline of a photon of energy E in some inertial frame, the ratio of E to $\mathrm{d}t$ is Lorentz invariant since $E\,\mathrm{d}x^\mu/\mathrm{d}t$ must be a 4-vector.

Hence, $p^m u$ is always parallel to the tangent vector $\mathrm{d}x^\mu/\mathrm{d}\lambda$ for any choice of parameterisation λ, and with a suitable choice of λ we can make $p^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$.

6.2.4.1 Doppler Effect Revisited

For photons, we can introduce the 4-*wavevector* $\mathbf{k}$ as $\mathbf{p} = \hbar\mathbf{k}$, with components in an inertial frame S of

$$k^\mu = \left(\frac{2\pi}{\lambda}, \vec{k}\right). \tag{6.59}$$

Here, λ is the wavelength in S and $\vec{k}$ is the 3D wavevector, with $\left|\vec{k}\right| = 2\pi/\lambda$.

Consider an observer at rest in inertial frame S observing light with wavelength λ propagating at an angle θ to the x-axis; the components of the 4-wavevector in S are

$$k^\mu = \frac{2\pi}{\lambda}(1, \cos\theta, \sin\theta, 0). \tag{6.60}$$

Suppose the light is emitted by a source that is moving at speed βc along the x-axis; in the rest-frame of the source (S'), the 4-wavevector has components

$$k'^\mu = \Lambda^\mu{}_\nu k^\nu, \tag{6.61}$$

where $\Lambda^\mu{}_\nu$ is the standard Lorentz boost (Eq. (6.8)). The emitted wavelength in the rest-frame, λ', follows from k'^0:

$$\begin{aligned} k'^0 &= \frac{2\pi}{\lambda'} = \frac{2\pi}{\lambda}\gamma(1 - \beta\cos\theta) \\ \implies \frac{\lambda}{\lambda'} &= \gamma(1 - \beta\cos\theta). \end{aligned} \tag{6.62}$$

For the particular case $\theta = 0$, this reduces to the result derived kinematically in Chapter 2,

$$\frac{\lambda}{\lambda'} = \sqrt{\frac{1-\beta}{1+\beta}}. \tag{6.63}$$

6.2.5 Example of Collisional Relativistic Mechanics: Compton Scattering

Compton scattering describes scattering of a photon from a charged particle. This can be considered as a collision between a photon with initial 4-momentum $\mathbf{p}$ and an electron, say, with initial 4-momentum $\mathbf{q}$. In the final state, the photon has 4-momentum $\bar{\mathbf{p}}$ and the electron has 4-momentum $\bar{\mathbf{q}}$.

We shall consider the collision in the inertial frame in which the electron is initially at rest, and the photon is propagating along the positive x-direction and has frequency ν. Suppose the photon scatters through an angle θ, and its final frequency is $\bar{\nu}$, and in the process the electron recoils (see Fig. 6.3). The components of the relevant 4-momenta are

$$\begin{aligned} p^\mu &= (h\nu/c, h\nu/c, 0, 0) \\ q^\mu &= (m_e c, 0, 0, 0) \\ \bar{p}^\mu &= (h\bar{\nu}/c, (h\bar{\nu}/c)\cos\theta, (h\bar{\nu}/c)\sin\theta, 0), \end{aligned} \tag{6.64}$$

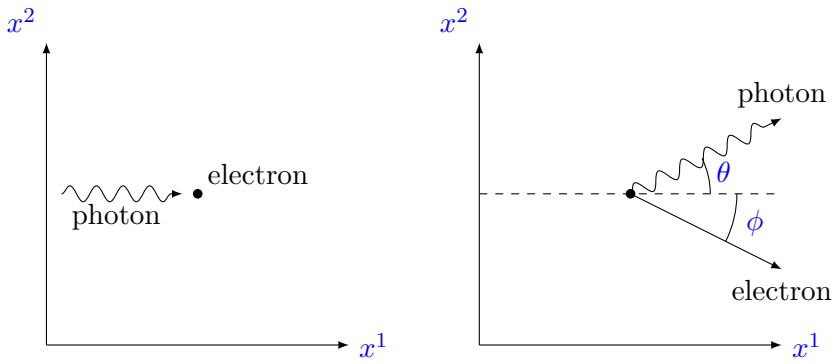

Fig. 6.3: The Compton effect showing a photon initially propagating along the x-axis scattering off an electron at rest (left). After the collision (right), the photon propagates at an angle θ to the x-axis, and the electron recoils.

where h is Planck's constant and m_e is the electron rest mass. (We shall not require the components of the final 4-momentum of the electron.)

The total 4-momentum is conserved, so

$$\mathbf{p} + \mathbf{q} - \bar{\mathbf{p}} = \bar{\mathbf{q}}. \tag{6.65}$$

We can also use the fact that the squared magnitude of the total 4-momentum is Lorentz invariant, and so equate the magnitude of the left-hand side of Eq. (6.65) evaluated in the initial rest-frame of the electron with the magnitude of the right-hand side evaluated in the final rest-frame. Using $|\mathbf{p}|^2 = 0$, and similarly for $|\bar{\mathbf{p}}|^2$, and $|\mathbf{q}|^2 = |\bar{\mathbf{q}}|^2 = m_e^2 c^2$, we have

$$\eta_{\mu\nu} p^\mu q^\nu - \eta_{\mu\nu} \bar{p}^\mu q^\nu - \eta_{\mu\nu} p^\mu \bar{p}^\nu = 0. \tag{6.66}$$

Substituting for the components from Eq. (6.64), we find

$$\begin{aligned} 0 &= h\nu m_e - h\bar{\nu} m_e - \left(\frac{h\nu}{c}\right)\left(\frac{h\bar{\nu}}{c}\right)(1 - \cos\theta) \\ \implies \bar{\nu} &= \frac{\nu}{1 + (h\nu/m_e c^2)(1 - \cos\theta)}. \end{aligned} \tag{6.67}$$

We see that, generally, the photon frequency is reduced during the collision, with energy being transferred to kinetic energy of the recoiling electron. This change in frequency follows only from a particle-like (i.e., quantum mechanical) description of light - in classical electromagnetism, the electron would be forced to oscillate at the frequency of the incident electromagnetic wave and so would also radiate at this frequency.

6.3 The Local Reference Frame of a General Observer

Consider a general observer $\mathcal{O}$ following a worldline $x^\mu(\tau)$. Their 4-velocity has components $u^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$ and the 4-acceleration is $a^\mu = Du^\mu/D\tau$. At any event on the wordline, we can define the instantaneous rest-frame of the particle as the inertial frame in which the particle is instantaneously at rest. At proper time τ, the coordinate basis vectors of

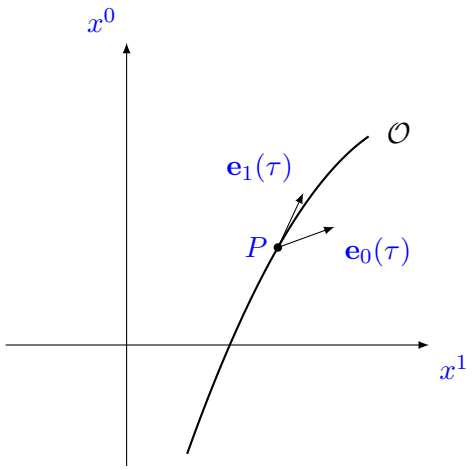

Fig. 6.4: For a general observer $\mathcal{O}$ following a worldline $x^\mu(\tau)$, we can define the instantaneous rest-frame of the particle as the inertial frame in which the particle is instantaneously at rest. At proper time τ, the coordinate basis vectors of the instantaneous rest-frame at the observer's position, P, constitute an orthonormal set of basis vectors $\mathbf{e}_\mu(\tau)$.

the instantaneous rest-frame at the observer's position constitute an orthonormal set of basis vectors $\mathbf{e}_\mu(\tau)$; see Fig. 6.4. By construction, the timelike basis vector $\mathbf{e}_0(\tau)$ is equal (up to a factor of c) to the instantaneous 4- velocity $\mathbf{u}(\tau)$. The three spacelike vectors $\mathbf{e}_i(\tau)$, $i = 1, 2, 3$, are therefore orthogonal to the observer's 4-velocity.

At some later time τ', the basis vector $\mathbf{e}_0(\tau')$ is uniquely determined by the 4-velocity $\mathbf{u}(\tau')$, but the remaining three spacelike vectors $\mathbf{e}_i(\tau')$ are only determined up to a spatial rotation. Additional information is required to specify the $\mathbf{e}_i$, such as demanding that they point along the directions specified by three orthogonal gyroscopes carried (with no torque applied) by the observer. For the special case of a non-accelerating observer carrying three such gyroscopes, the $\mathbf{e}_\mu(\tau)$ undergo parallel transport along the particle's wordline.[5] This leads us to the following idealisation of a local laboratory for an arbitrary observer: the observer (possibly accelerating) carries along four orthonormal vectors $\mathbf{e}_\mu(\tau)$ that satisfy

$$\mathbf{g}(\mathbf{e}_\mu, \mathbf{e}_\nu) = \eta_{\mu\nu} \qquad \text{and} \qquad c\mathbf{e}_0(\tau) = \mathbf{u}(\tau). \tag{6.68}$$

Such a frame of vectors is called an *orthonormal tetrad*. The results of any local measurement made by the observer at proper time τ can be represented as the components of tensor-valued quantities in this tetrad.

6.4 Minkowski Space in Other Coordinate Systems

In Minkowski space, it is usually most convenient to work in the Cartesian coordinates of an inertial frame. The advantages of working in these coordinates are the following:

[5] More generally, for an accelerated observer the $\mathbf{e}_i(\tau)$ *cannot* be parallel-transported since they have to remain orthogonal to $\mathbf{u}(\tau)$. If the orientation of the $\mathbf{e}_i(\tau)$ is determined by gyroscopes, the basis vectors at proper time $\tau + \mathrm{d}\tau$ are obtained from those at τ by first parallel-transporting to the observer's new position, then applying the additional pure Lorentz boost required to boost the parallel-transported $\mathbf{e}_0$ onto $\mathbf{u}(\tau + \mathrm{d}\tau)$. Such basis vectors are said to be Fermi–Walker transported and are the idealisation of a local *non-rotating* laboratory

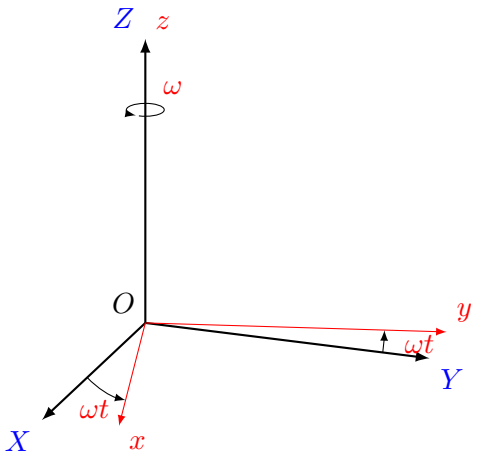

Fig. 6.5: Non-Inertial Coordinates: A coordinate system $x^\mu = (ct, x, y, z)$, where points with fixed x, y and z coordinates rotate with angular speed ω about the Z axis in an inertial frame S, defined by Cartesian coordinates $X^\mu = (cT, X, Y, Z)$.

1. the coordinates have a simple physical interpretation in terms of distances and times measured by observers in some inertial frame; and
2. covariant differentiation of tensors reduces to partial differentiation of the components.

However, for some applications other coordinate systems are more appropriate. A trivial example is to use spherical polar coordinates, say, rather than spatial Cartesian coordinates, in some inertial frame. A less trivial example is to use a rotating coordinate system, an example that we shall now discuss.

6.4.1 Non-Inertial Coordinates: a Rotating Frame

Let $X^\mu = (cT, X, Y, Z)$ be Cartesian coordinates of an inertial frame S (see Fig. 6.5). Introduce new coordinates $x^\mu = (ct, x, y, z)$ where

$$\begin{aligned} X &= x\cos\omega t - y\sin\omega t, \\ Y &= x\sin\omega t + y\cos\omega t, \\ Z &= z, \\ T &= t. \end{aligned} \tag{6.69}$$

Points with fixed x, y and z coordinates rotate with angular speed ω about the Z axis in S.

Evaluating the differentials

$$\mathrm{d}X = \mathrm{d}x\cos\omega t - \mathrm{d}y\sin\omega t - \omega\,\mathrm{d}t\,(x\sin\omega t + y\cos\omega t), \tag{6.70}$$

$$\mathrm{d}Y = \mathrm{d}x\sin\omega t + \mathrm{d}y\cos\omega t + \omega\,\mathrm{d}t\,(x\cos\omega t - y\sin\omega t), \tag{6.71}$$

we find

$$\mathrm{d}X^2 + \mathrm{d}Y^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \omega^2(x^2 + y^2)\,\mathrm{d}t^2 + 2\omega\,\mathrm{d}t\,(x\,\mathrm{d}y - y\,\mathrm{d}x). \tag{6.72}$$

The line element $\mathrm{d}s^2 = c^2\,\mathrm{d}T^2 - \mathrm{d}X^2 - \mathrm{d}Y^2 - \mathrm{d}Z^2$ becomes

$$\mathrm{d}s^2 = \left[c^2 - \omega^2(x^2+y^2)\right]\mathrm{d}t^2 + 2\omega y\,\mathrm{d}t\,\mathrm{d}x - 2\omega x\,\mathrm{d}t\,\mathrm{d}y - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \tag{6.73}$$

in terms of the x^μ coordinates.

Free particles move on timelike geodesics in Minkowski space, with equation of motion

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{\rho\sigma}\frac{\mathrm{d}x^\rho}{\mathrm{d}\tau}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\tau} = 0. \tag{6.74}$$

Rather than calculating the metric connection directly, it is often quicker to follow the "Lagrangian" route, with $L = g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu$, where overdots denote differentiation with respect to proper time τ. For the line element in the rotating coordinates, the Lagrangian is

$$L = \left[c^2 - \omega^2(x^2+y^2)\right]\dot{t}^2 + 2\omega y\dot{x}\dot{t} - 2\omega x\dot{y}\dot{t} - \dot{x}^2 - \dot{y}^2 - \dot{z}^2. \tag{6.75}$$

The Euler–Lagrange equation for $t(\tau)$ gives

$$\begin{gathered}\frac{\mathrm{d}}{\mathrm{d}\tau}\Big(\left[c^2 - \omega^2(x^2+y^2)\right]\dot{t} + \omega(y\dot{x} - x\dot{y})\Big) = 0 \\ \Longrightarrow\ \left[c^2 - \omega^2(x^2+y^2)\right]\ddot{t} - 2\omega^2(x\dot{x} + y\dot{y})\dot{t} + \omega(y\ddot{x} - x\ddot{y}) = 0.\end{gathered} \tag{6.76}$$

For $x(\tau)$ and $y(\tau)$, we have

$$\ddot{x} = \omega y\ddot{t} + \omega^2 x\dot{t}^2 + 2\omega\dot{y}\dot{t}, \tag{6.77}$$

$$\ddot{y} = -\omega x\ddot{t} + \omega^2 y\dot{t}^2 - 2\omega\dot{x}\dot{t}, \tag{6.78}$$

while for z we find $\ddot{z} = 0$. Substituting for $\dot{x}$ and $\ddot{y}$ into Eq. (6.76), a large number of cancellations take place to leave

$$\ddot{t} = 0. \tag{6.79}$$

This is not unexpected: t is also the time coordinate in the Cartesian inertial coordinates X^μ, so $\ddot{t} = 0$ must hold for a free particle.

If we now parameterise x, y and z in terms of t rather than τ, using $\mathrm{d}t/\mathrm{d}\tau = \text{const.}$, we get

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = \omega^2 x + 2\omega\frac{\mathrm{d}y}{\mathrm{d}t}, \tag{6.80}$$

$$\frac{\mathrm{d}^2y}{\mathrm{d}t^2} = \omega^2 y - 2\omega\frac{\mathrm{d}x}{\mathrm{d}t}, \tag{6.81}$$

$$\frac{\mathrm{d}^2z}{\mathrm{d}t^2} = 0. \tag{6.82}$$

These are just the usual equations of motion for a free particle in a rotating frame, involving the centrifugal and Coriolis accelerations

Chapter 7

Electromagnetism

Electromagnetism is a relativistic theory, despite being conceived before special relativity. Maxwell's equations in free space, expressed in terms of Cartesian coordinates in some inertial frame S, are

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0}, \tag{7.1}$$

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}, \tag{7.2}$$

$$\vec{\nabla} \cdot \vec{B} = 0, \tag{7.3}$$

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{J} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t}, \tag{7.4}$$

where $\vec{\nabla}$ is the usual 3D derivative, ρ and $\vec{J}$ are the charge and current density in S, and $\vec{E}$ and $\vec{B}$ are the electric and magnetic fields in S. Maxwell's equations are supplemented by the *Lorentz force law*,

$$\vec{f} = q\Big(\vec{E} + \vec{u} \times \vec{B}\Big), \tag{7.5}$$

which gives the electromagnetic 3-force on a particle of charge q and 3-velocity $\vec{u}$. Charge conservation is built into Maxwell's equations, taking the divergence of Eq. (7.4) and using Eq. (7.1) gives the continuity equation

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot \vec{J} = 0. \tag{7.6}$$

It is not obvious when written in their 3D form, but Maxwell's equations do satisfy the principle of relativity: it is possible to find transformation laws for the electric and magnetic fields such that the equations are unchanged in form under Lorentz transformations. You will already have seen hints of this in previous courses, where you have shown that Maxwell's equations can be reduced to wave equations with solutions that propagate at the speed of light. In this section, we shall develop electromagnetism as a relativistic field theory on Minkowski spacetime. In particular, we shall show how the 3D Maxwell's equations can be combined into 4D tensor equations, which, by virtue of being tensor equations, show that the theory is relativistically covariant. Moreover, by writing the equations in tensor form, we can express the theory in a form that is independent of the particular coordinate system that is used.

7.1 Lorentz Force Law

Recall from Eq. (6.53) that the 3-force $\vec{f}$ acting on a particle with 3-velocity $\vec{u}$ in some inertial frame is related to the spatial components of the force 4-vector,

$$f^\mu = \gamma_u \left(\frac{\vec{f} \cdot \vec{u}}{c}, \vec{f} \right). \tag{7.7}$$

Since the 3-force depends linearly on the 3-velocity, we might expect the force 4-vector to depend linearly on the 4-velocity, i.e.,

$$f_\mu = qF_{\mu\nu}u^\nu. \tag{7.8}$$

Here, we have lowered the index of f_μ (to form the associated dual vector), and introduced a type-$(0,2)$ tensor with components $F_{\mu\nu}$ called the *Maxwell field-strength tensor*. The charge q is a scalar quantity (i.e., all observers agree on its value), so $F_{\mu\nu}$ must be the components of a tensor since f_μ and u^μ are.

The equation of motion for the 4-velocity in the presence of this force is

$$\frac{\mathrm{D}u^\mu}{\mathrm{D}\tau} = \frac{q}{m}F^\mu{}_\nu\, u^\nu, \tag{7.9}$$

where we have raised the first index on $F_{\mu\nu}$. The 4-force has to be orthogonal to the 4-velocity, $f_\mu u^\mu = 0$, which requires the field-strength tensor to be antisymmetric:

$$f_\mu u^\mu = qF^\mu{}_\nu\, u^\mu u^\nu = 0 \quad \Longrightarrow \quad F_{\mu\nu} = -F_{\nu\mu}. \tag{7.10}$$

If we raise both indices to form $f^{\mu\nu} = g^{\mu\rho}g^{\nu\sigma}F_{\rho\sigma}$ (working in general coordinates), we preserve antisymmetry: $F^{\mu\nu} = -F^{\nu\mu}$.

We can relate the components $F_{\mu\nu}$ in Cartesian inertial coordinates to the electric and magnetic field components in that inertial frame by comparing Eq. (7.8) to the 3D Lorentz force law. We have

$$f_0 = q\sum_i F_{0i}u^i = \frac{q}{c}\gamma_u \vec{E}\cdot\vec{u}. \tag{7.11}$$

From this, we can read off that, numerically, $F_{0i} = \vec{E}^i/c$. For the spatial components, we have

$$f_i = qF_{i0}u^0 + q\sum_j F_{ij}u^j = -q\gamma_u\Big(\vec{E}^i + (\vec{u}\times\vec{B})^i\Big), \tag{7.12}$$

where we used $f_i = -\gamma\vec{f^i}$ in Cartesian inertial coordinates. The $-q\gamma_u\vec{E}^i$ term equals $qF_{i0}u^0$, so we are left with

$$q\gamma_u\sum_i F_{ij}\vec{u}^i = -q\gamma_u(\vec{u}\times\vec{B})^i. \tag{7.13}$$

Setting $i = 1, 2$ and 3, we can read off

$$F_{12} = -\vec{B}^3, \quad F_{13} = \vec{B}^2, \quad F_{23} = -\vec{B}^1. \tag{7.14}$$

We arrive at the following components of the field-strength tensor in terms of the electric and magnetic fields:

$$F_{\mu\nu} = \begin{pmatrix} 0 & \vec{E}^1/c & \vec{E}^2/c & \vec{E}^3/c \\ -\vec{E}^1/c & 0 & -\vec{B}^3 & \vec{B}^2 \\ -\vec{E}^2/c & \vec{B}^3 & 0 & -\vec{B}^1 \\ -\vec{E}^3/c & -\vec{B}^2 & \vec{B}^1 & 0 \end{pmatrix}. \tag{7.15}$$

For reference, if we raise both indices with $\eta^{\mu\nu}$, we obtain

$$F^{\mu\nu} = \begin{pmatrix} 0 & -\vec{E}^1/c & -\vec{E}^2/c & -\vec{E}^3/c \\ \vec{E}^1/c & 0 & -\vec{B}^3 & \vec{B}^2 \\ \vec{E}^2/c & \vec{B}^3 & 0 & -\vec{B}^1 \\ \vec{E}^3/c & -\vec{B}^2 & \vec{B}^1 & 0 \end{pmatrix}. \tag{7.16}$$

As $F^{\mu\nu}$ are the components of a type-$(2,0)$ tensor, we know how they transform under a Lorentz transformation $x'^\mu = \Lambda^\mu{}_\nu x^\nu$:

$$F'^{\mu\nu} = \Lambda^\mu{}_\rho \Lambda^\nu{}_\sigma F^{\rho\sigma}. \tag{7.17}$$

This tells us how the electric and magnetic fields are related in different inertial frames; for a standard Lorentz boost we have

$$\vec{E}' = \begin{pmatrix} \vec{E}^1 \\ \gamma(\vec{E}^2 - v\vec{B}^3) \\ \gamma(\vec{E}^3 - v\vec{B}^2) \end{pmatrix}, \qquad \vec{B}' = \begin{pmatrix} \vec{B}^1 \\ \gamma(\vec{B}^2 + v\vec{E}^3/c^2) \\ \gamma(\vec{B}^3 - v\vec{E}^2/c^2) \end{pmatrix}. \tag{7.18}$$

7.2 Maxwell's Equations

The relativistic form of the Lorentz force law has led us to introduce the field-strength tensor, whose components in any inertial frame encode the electric and magnetic fields there. We now seek to express Maxwell's equations in terms of the spacetime derivative of the field-strength tensor (since the 3D Maxwell's equations have first-order space and time derivatives). However, before we do this we need to consider the source terms ρ and $\vec{J}$.

7.2.1 Current 4-Vector

Since a static charge density in some inertial frame S will transform to a moving charge density, i.e., possess a current, under a change in inertial frame, we might expect to be able to assemble a 4-vector from ρ and $\vec{J}$.

Consider a current distribution $\vec{J}$ formed from a charge density ρ moving with 3-velocity $(v,0,0)$ in inertial frame S, so that $\vec{J} = \rho(v,0,0)$. Let S' be the rest-frame of the charges, which is in standard configuration with S, and let the charge density in this frame be ρ'. The current vanishes in S' since the charges are at rest.

We can relate the charge density and current in the two frames by the following physical argument. Take some given volume V' in S' and mark those charges that lie within V'. The same charges occupy a smaller volume V'/γ in S by length contraction (see Fig. 7.1), but the amount of charge is the same. It follows that charge density is larger in S by a factor of γ so

$$\rho = \gamma\rho_0 \quad \text{and} \quad \vec{J} = \gamma\rho_0(v,0,0). \tag{7.19}$$

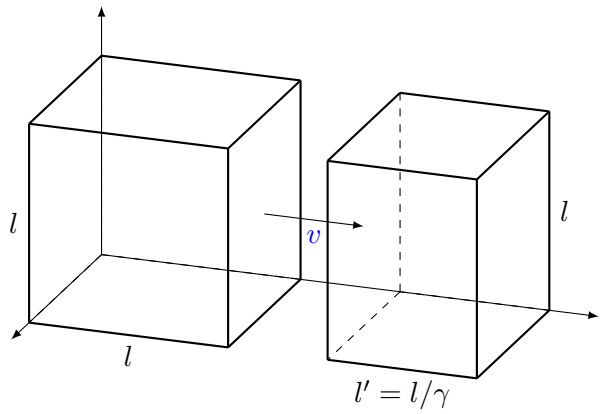

Fig. 7.1: Length contraction of the volume occupied by a given set of charges between their rest frame volume (left) and that in a frame in which they are moving with speed v (right).

However, these are just the transformations that would follow if ρc and $\vec{J}$ were the time and space components of a *current* 4*-vector*

$$\boxed{j^\mu = \left(c\rho, \vec{J}\right).} \tag{7.20}$$

This is because in S' we would have $j'^\mu = (c\rho_0, \vec{0})$, and, on Lorentz transforming,

$$\begin{aligned} j^0 &= \gamma(j'^0 + \beta j'^1) \\ \implies \rho &= \gamma\rho_0, \end{aligned} \tag{7.21}$$

and

$$\begin{aligned} j^1 &= \gamma(j'^1 + \beta j'^0) \\ \implies \vec{J}^1 &= \gamma v\rho_0. \end{aligned} \tag{7.22}$$

7.2.2 Relativistic Field Equations

We want to relate the field-strength tensor to the current 4-vector and we expect this relation to be linear in spacetime derivatives. If we contract the type-$(2,0)$ tensor $F^{\mu\nu}$ with the covariant derivative (i.e., form the covariant divergence), we necessarily form a 4-vector. We therefore consider a tensor equation of the form

$$\nabla_\mu F^{\mu\nu} = kj^\nu, \tag{7.23}$$

for some constant scalar k.

Let us consider this equation in Cartesian inertial coordinates, so that the covariant derivative becomes a partial derivative and we have

$$\partial_\mu F^{\mu\nu} = kj^\nu. \tag{7.24}$$

Taking the divergence and using the antisymmetry of $F^{\mu\nu}$ gives

$$\begin{aligned} k\partial_\nu j^\nu &= \partial_\nu\partial_\mu F_{\mu\nu} = 0 \\ \implies \partial_\mu j^\mu &= 0. \end{aligned} \tag{7.25}$$

This is the just the continuity equation in 4D language since, using $j^\mu = (c\rho, \vec{J})$,

$$\frac{\partial j^0}{\partial(ct)} + \sum_i \frac{\partial j^i}{\partial x^i} = 0 \quad \Longrightarrow \quad \frac{\partial \rho}{\partial t} + \vec{\nabla}\cdot\vec{J} = 0. \tag{7.26}$$

Eq. (7.23) therefore looks promising, but we still need to show that it does return the correct Maxwell's equations.

Consider the 0-component:

$$\begin{aligned} &\frac{\partial F^{00}}{\partial(ct)} + \sum_i \frac{\partial F^{i0}}{\partial x^i} = kj^0 \\ \Longrightarrow\ &\frac{1}{c}\vec{\nabla}\cdot\vec{E} = kc\rho. \end{aligned} \tag{7.27}$$

Recalling that $\epsilon_0\mu_0 = 1/c^2$, we see that we recover the Maxwell equation $\vec{\nabla}\cdot\vec{E} = \rho/\epsilon_0$ if we take $k = \mu_0$. Consider now the spatial components:

$$\frac{\partial F^{0i}}{\partial(ct)} + \sum_j \frac{\partial F^{ji}}{\partial x^j} = \mu_0 j^i. \tag{7.28}$$

Taking $i = 1$, we have

$$-\frac{1}{c^2}\frac{\partial \vec{E}^1}{\partial t} + \underbrace{\frac{\partial \vec{B}^3}{\partial x^2} - \frac{\partial \vec{B}^2}{\partial x^3}}_{(\vec{\nabla}\times\vec{B})^1} = \mu_0 \vec{J}^1. \tag{7.29}$$

Repeating for the other components, and rearranging, we recover

$$\vec{\nabla}\times\vec{B} = \mu_0\vec{J} + \mu_0\epsilon_0\frac{\partial \vec{E}}{\partial t}. \tag{7.30}$$

Eq. (7.23) is therefore the tensor version of the two sourced Maxwell's equations, but what of the other two?

We need to introduce a further (homogeneous) tensor equation to capture the two source-free equations,

$$\vec{\nabla}\times\vec{E} = -\frac{\partial \vec{B}}{\partial t}, \tag{7.31}$$

$$\vec{\nabla}\cdot\vec{B} = 0. \tag{7.32}$$

These are four equations in total, so we look for a tensor equation involving only the covariant derivative of the field-strength that has four independent components. In 4D, a totally-antisymmetric type-$(0,3)$ tensor has $4\times 3\times 2/3! = 4$ independent components, so we consider

$$\nabla_{[\mu}F_{\nu\rho]} = 0, \tag{7.33}$$

where, recall, the square brackets denote the antisymmetric part. Since $F_{\mu\nu}$ is itself antisymmetric, this can be written explicitly as

$$\nabla_\mu F_{\nu\rho} + \nabla_\nu F_{\rho\mu} + \nabla_\rho F_{\mu\nu} = 0. \tag{7.34}$$

Again, let us consider this tensor equation in Cartesian inertial coordinates, in which case it reduces to[1]

$$\partial_\mu F_{\nu\rho} + \partial_\nu F_{\rho\mu} + \partial_\rho F_{\mu\nu} = 0. \tag{7.35}$$

The four independent choices of indices are

$$(\mu,\nu,\rho) = (0,1,2)\,,\ (0,1,3\,,\ (0,2,3)\,,\ (1,2,3). \tag{7.36}$$

Consider these in turn; for $(\mu,\nu,\rho) = (0,1,2)$ we have

$$-\frac{\partial \vec{B}^3}{\partial(ct)} - \frac{1}{c}\underbrace{\frac{\partial \vec{E}^2}{\partial x^1} - \frac{\partial \vec{E}^1}{\partial x^2}}_{\left(\vec{\nabla}\times\vec{E}\right)^3} = 0, \tag{7.37}$$

which is the 3-component of Eq. (7.31), and the other two cases with $\mu = 0$ give the remaining components. For $(\mu,\nu,\rho) = (1,2,3)$, we have

$$-\frac{\partial \vec{B}^1}{\partial x^1} - \frac{\partial \vec{B}^2}{\partial x^2} - \frac{\partial \vec{B}^3}{\partial x^3} = 0, \tag{7.38}$$

which is $\vec{\nabla}\cdot\vec{B} = 0$.

To summarise, the four Maxwell's equations in any inertial frame are the components in Cartesian inertial coordinates of the two tensor equations

$$\nabla_\mu F_{\mu\nu} = \mu_0 j^\nu, \tag{7.39}$$

$$\nabla_\mu F_{\nu\rho} + \nabla_\nu F_{\rho\mu} + \nabla_\rho F_{\mu\nu} = 0. \tag{7.40}$$

Being tensor equations, these are valid in any coordinate system covering Minkowski space. Charge conservation is built into these equations: in Cartesian inertial coordinates the conservation equation takes the form $\partial_\mu j^\mu = 0$, and in general coordinates this becomes the tensor equation[2]

$$\nabla_\mu j^\mu = 0. \tag{7.44}$$

[1] The equation actually takes this form in any coordinate system due to the symmetry of the metric connection.

[2] A direct proof of this result in a general coordinate system is as follows. From Eq. (7.23), we have

$$\begin{aligned}\mu\nabla_\nu j^\nu &= \nabla_\nu\nabla_\mu F^{\mu\nu}\\ &= \partial_\nu(\nabla_\mu F^{\mu\nu}) + \Gamma^\nu_{\nu\rho}(\nabla_\mu F^{\mu\rho})\\ &= \partial_\nu(\nabla_\mu F^{\mu\nu}) + \frac{1}{2}(\partial_\rho \ln|g|)(\nabla_\mu F^{\mu\rho}),\end{aligned} \tag{7.41}$$

where we used the result $\Gamma^\nu_{\nu\rho} = (\partial_\rho \ln|g|)/2$ (See Chapter 6). We now use

$$\begin{aligned}\nabla_\mu F^{\mu\nu} &= \partial_\mu F^{\mu\nu} + \Gamma^\mu_{\mu\rho}F^{\rho\nu} + \underbrace{\Gamma^\nu_{\mu\rho}F^{\mu\rho}}_{0}\\ &= \partial_\mu F^{\mu\nu} + \frac{1}{2}(\partial_\rho \ln|g|)F^{\rho\nu}\end{aligned} \tag{7.42}$$

to find

$$\begin{aligned}\mu\nabla_\nu j^\nu &= \partial_\nu\partial_\mu F^{\mu\nu} + \frac{1}{2}(\partial_\nu\partial_\rho \ln|g|)F^{\rho\nu} + \frac{1}{2}(\partial_\rho \ln|g|)\partial_\nu F^{\rho\nu} + \frac{1}{2}(\partial_\rho \ln|g|)\partial_\mu F^{\mu\rho} + \frac{1}{4}(\partial_\rho \ln|g|)(\partial_\sigma \ln|g|)F^{\sigma\rho}\\ &= 0,\end{aligned} \tag{7.43}$$

where we have used the commutativity of partial derivatives and the antisymmetry of $F^{\mu\nu}$.

7.2.3 The 4-Vector Potential

The Maxwell equation $\nabla_{[\mu}F_{\nu\rho]} = 0$, written in Cartesian inertial coordinates, implies that the type-$(0,2)$ field strength tensor can be derived from a (dual)-vector potential A_μ. If we take

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \tag{7.45}$$

then $\nabla_{[\mu}F_{\nu\rho]} = 0$ is identically satisfied since

$$\begin{aligned} \partial_\mu F_{\nu\rho} + \partial_\nu F_{\rho\mu} + \partial_\rho F_{\mu\nu} &= \partial_\mu(\partial_\nu A_\rho - \partial_\rho A_\nu) \\ &\quad + \partial_\nu(\partial_\rho A_\mu - \partial_\mu A_\rho) \\ &\quad + \partial_\rho(\partial_\mu A_\nu - \partial_\nu A_\mu) \\ &= 0. \end{aligned} \tag{7.46}$$

This is analogous to a (globally) curl-free vector being expressible as the gradient of a scalar in 3D Euclidean space.

The 4-vector potential is not uniquely determined by the field-strength tensor - there is a residual *gauge freedom* to add the gradient of a scalar ψ to A_μ without altering $F_{\mu\nu}$:

$$\begin{aligned} F_{\mu\nu} &\to \partial_\mu(A_\nu + \partial_\nu\psi) - \partial_\nu(A_\mu\partial_\mu\psi) \\ &= \partial_\mu A_\nu - \partial_\nu A_\mu = F_{\mu\nu}. \end{aligned} \tag{7.47}$$

A convenient gauge choice, which has the virtue of being Lorentz invariant, is to take A_μ to be divergence free,

$$\partial_\mu A^\mu = 0. \tag{7.48}$$

This gauge choice is called the *Lorenz gauge* (Lorenz was different to Lorentz!). The sourced 4D Maxwell equation can be written directly in terms of the 4-vector potential. It is convenient to lower the index and write

$$\nabla_\mu F_{\mu\nu} = \mu_0 j_\nu, \tag{7.49}$$

or, in Cartesian inertial coordinates,

$$\eta^{\mu\rho}\partial_\rho(\partial_\mu A_\nu - \partial_\nu A_\mu) = \mu_0 j_\nu. \tag{7.50}$$

In the Lorenz gauge, this simplifies further to

$$\nabla^2 A_\nu = \mu_0 j_\nu \tag{7.51}$$

where $\nabla^2 A_\nu = \eta^{\mu\rho}\partial_\mu\partial_\rho A_\nu$ is the 4D Laplacian in Cartesian inertial coordinates. The Laplacian in Minkowski spacetime is the wave operator, with wave speed c:

$$\nabla^2 = \frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \vec{\nabla}^2. \tag{7.52}$$

In the absence of charges and currents, Eq. (7.51) therefore admits wavelike solutions that travel at the speed of light. In the presence of a variable source, Eq. (7.51) describes the generation of electromagnetic fields that asymptotically far from the source describe radiation fields.

In Cartesian inertial coordinates, the components of A^μ are the familiar scalar and magnetic-vector potentials of 3D electromagnetism:

$$A^\mu = \left(\phi/c, \vec{A}\right) \quad \text{or} \quad A_\mu = \left(\phi/c, -\vec{A}\right). \tag{7.53}$$

Using the relation Eq. (7.15) between the components of the field-strength tensor and the electric and magnetic fields, we find

$$\begin{aligned} F_{0i} &= -\frac{\partial \vec{A}^i}{\partial(ct)} - \frac{1}{c}\frac{\partial \phi}{\partial x^i} \\ \implies \vec{E} &= -\frac{\partial \vec{A}}{\partial t} - \vec{\nabla}\phi. \end{aligned} \tag{7.54}$$

Similarly, the spatial components of $F_{\mu\nu}$ give, for example,

$$F_{12} = -\frac{\partial \vec{A}^2}{\partial x^1} + \frac{\partial \vec{A}^1}{\partial x^2} = -\left(\vec{\nabla} \times \vec{A}\right)^3, \tag{7.55}$$

so that

$$\vec{B} = \vec{\nabla} \times \vec{A}. \tag{7.56}$$

The source-free 3D Maxwell equations (7.31) and (7.32) are identically satisfied by the relations (7.54) and (7.56).Finally, we note that the generalisation of Eq. (7.45) to arbitrary coordinates is

$$F_{\mu\nu} = \nabla_\mu A_\nu - \nabla_\nu A_\mu. \tag{7.57}$$

However, due to the symmetry of the metric connection, this is actually equivalent to $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ in all coordinate systems.

7.3 Electromagnetism in Curved Spacetime

We have seen that in Minkowski spacetime, the tensor equations

$$\nabla_\mu F_{\mu\nu} = \mu_0 j^\nu, \tag{7.58}$$

$$\nabla_\mu F_{\nu\rho} + \nabla_\nu F_{\rho\mu} + \nabla_\rho F_{\mu\nu} = 0. \tag{7.59}$$

are equivalent to the usual 3D Maxwell equations when expressed in global Cartesian coordinates. In curved spacetime, electromagnetism is described by exactly the same tensor equations. This is because if we express these tensor equations in local inertial coordinates at some point, they reduce to the same form there as in Minkowski space in inertial coordinates, as required by the equivalence principle.

Chapter 8

Spacetime Curvature

The equivalence principle has led us to formulate physical laws as tensor equations that reduce to their usual special-relativistic form in local-inertial coordinates. If we could construct such coordinates globally across spacetime, there would be no manifestations of gravity. Gravity enters only via our inability to construct such coordinates, i.e., through the curvature of spacetime.

In general relativity, gravity is no longer regarded as a force in the conventional sense, but rather as a manifestation of spacetime curvature, where the curvature is itself due to the presence of matter. In this topic we shall develop further the idea of the intrinsic curvature of a manifold, and in the next we shall learn how general relativity relates the curvature to the matter that is present.

8.1 Gravity as Spacetime Curvature

To capture gravitational effects, we need to model spacetime as a 4D manifold that is more complicated than Minkowski space. However, we are restricted to pseudo-Euclidean manifolds by the equivalence principle: we must be able to find coordinates X^μ locally such that the line element reduces to the Minkowski form

$$\mathrm{d}s^2 \approx \eta_{\mu\nu}\,\mathrm{d}X^\mu\,\mathrm{d}X^\nu\,, \tag{8.1}$$

and this is only the case when the interval is quadratic in the coordinate differentials. In curved spacetime, the equivalence principle tells us that the equation of motion of a massive particle is

$$\frac{\mathrm{D}u^\mu}{\mathrm{D}\tau} = 0, \tag{8.2}$$

with the 4-velocity $u^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$, since this tensor equation reduces to the special-relativistic form $\mathrm{d}^2X^\mu/\mathrm{d}\tau^2 = 0$ in local-inertial coordinates. In other words, the wordline of a particle freely-falling under gravity is a geodesic in curved spacetime.

8.1.1 Local-Inertial Coordinates

Close to any point P, we can always construct coordinates X^μ on a pseudo-Euclidean 4D manifold such that

$$g_{\mu\nu}(P) = \eta_{\mu\nu} \quad \text{and} \quad (\partial_\rho g_{\mu\nu})_P = 0; \tag{8.3}$$

these imply that the metric connection vanishes at P. Physically, these local-inertial coordinates correspond to a free-falling, non-rotating, Cartesian reference frame over some limited region of spacetime. In such coordinates, the coordinates basis vectors $\mathbf{e}_\mu \equiv \partial/\partial X^\mu$ are orthonormal at P:

$$\mathbf{g}(\mathbf{e}_\mu, \mathbf{e}_\nu) = \eta_{\mu\nu}. \tag{8.4}$$

The local-inertial coordinates are not unique at P; there are infinitely many such coordinate systems related by Lorentz transformations at P. Close to the point P, in local-inertial coordinates

$$g_{\mu\nu} = \eta_{\mu\nu} + \frac{1}{2}\left(\frac{\partial^2 g_{\mu\nu}}{\partial X^\rho \partial X^\sigma}\right)_P [x^\rho - X^\rho(P)][X^\sigma - X^\sigma(P)] + \cdots . \tag{8.5}$$

The size of the second derivative of the metric determines the extent of the region over which physics locally looks like special relativity in these coordinates.

8.1.1.1 Fermi-Normal Coordinates

It is possible to construct coordinates all along a timelike geodesic such that $g_{\mu\nu} = \eta_{\mu\nu}$ and $\Gamma^\rho_{\mu\nu} = 0$ on the geodesic. These coordinates, called *Fermi-normal coordinates*, extend the idea of local-inertial coordinates from the vicinity of a point to the vicinity of a timelike geodesic. Fermi-normal coordinates can be constructed physically by the following procedure:

- Take a free-falling observer carrying an orthonormal frame of vectors, $\hat{\mathbf{e}}_0(\tau)$ equal to their 4-velocity and three spacelike vectors $\{\hat{\mathbf{e}}_i(\tau)\}$, where τ is proper time for the observer. All four vectors are parallel transported along the observer's geodesic worldline (e.g., the $\hat{\mathbf{e}}_i(\tau)$ could be defined by the direction of gyroscopes supported at their centre of mass).
- At every proper time τ, the observer constructs a family of spacelike geodesics that have unit tangent vectors at the observer constructed as linear combinations of the $\hat{\mathbf{e}}_i(\tau)$.
- Any point close to the observer's worldline will lie on exactly one such spacelike geodesic $\mathcal{C}$; assign coordinates to this point with $T = \tau$ and the X^i equal to the products of the direction cosines of the unit tangent vector to $\mathcal{C}$ at the observer and the proper distance along $\mathcal{C}$.

The resulting coordinates can be shown to have the properties advertised above. The $\hat{\mathbf{e}}_0$ and $\hat{\mathbf{e}}_i$ are the (orthonormal) coordinate basis vectors of the Fermi-normal coordinates along the observer's geodesic. Local measurements made by the observer correspond to projections of tensors onto this orthonormal frame.

8.1.2 Newtonian Limit for a Free-Falling Particle

Let us verify that we can recover the correct Newtonian limit for free-falling particles from the geodesic equation in the limit of low speeds and weak gravitational fields. In the absence of gravity, spacetime is Minkowski space; for weak gravitational fields we expect to be able to find global coordinates where the metric is close to Minkowski (spacetime is only "weakly curved"):

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \quad \text{where} \quad |h_{\mu\nu}| \ll 1. \tag{8.6}$$

We shall also assume that the metric is stationary in these coordinates, $\partial h_{\mu\nu}/\partial x^0 = 0$, which corresponds to our usual intuition of a static gravitational field. For slow-moving particles relative to this coordinate system, $\left|\mathrm{d}x^i/\mathrm{d}t\right| \ll c$, where $ct = x^0$, and so

$$\left|\frac{\mathrm{d}x^i}{\mathrm{d}\tau}\right| \ll \frac{\mathrm{d}x^0}{\mathrm{d}\tau}. \tag{8.7}$$

In the geodesic equation,

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{\rho\sigma}\frac{\mathrm{d}x^\rho}{\mathrm{d}\tau}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\tau} = 0, \tag{8.8}$$

we can therefore ignore $\mathrm{d}x^i/\mathrm{d}\tau$ terms relative to $\mathrm{d}x^0/\mathrm{d}\tau$ in the connection part, so that

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{00}c^2\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 \approx 0. \tag{8.9}$$

The relevant connection coefficients to first-order in $h_{\mu\nu}$ are

$$\begin{aligned}\Gamma^\mu_{00} &= \frac{1}{2}g^{\mu\nu}\left(2\frac{\mathrm{d}g_{v0}}{\mathrm{d}x^0} - \frac{\mathrm{d}g_{00}}{\mathrm{d}x^\nu}\right)\\ &\approx -\frac{1}{2}\eta^{\mu\nu}\frac{\mathrm{d}g_{00}}{\mathrm{d}x^\nu}\\ &\approx -\frac{1}{2}\sum_i \eta^{\mu i}\frac{\mathrm{d}h_{00}}{\mathrm{d}x^i},\end{aligned} \tag{8.10}$$

where we used that derivatives of the metric are first order in $h_{\mu\nu}$ in passing to the second line on the right. It follows that

$$\Gamma^0_{00} \approx 0 \quad \text{and} \quad \Gamma^i_{00} \approx \frac{1}{2}\frac{\mathrm{d}h_{00}}{\mathrm{d}x^i}. \tag{8.11}$$

The 0-component of the geodesic equation gives $\mathrm{d}^2t/\mathrm{d}\tau^2 \approx 0$, so that $\mathrm{d}t/\mathrm{d}\tau = \text{const}$. The ith component of the geodesic equation then becomes

$$\begin{aligned}\frac{\mathrm{d}^2x^i}{\mathrm{d}\tau^2} &\approx -\frac{c^2}{2}\frac{\mathrm{d}h_{00}}{\mathrm{d}x^i}\left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2\\ \implies \frac{\mathrm{d}^2x^i}{\mathrm{d}t^2} &\approx -\frac{c^2}{2}\frac{\mathrm{d}h_{00}}{\mathrm{d}x^i}.\end{aligned} \tag{8.12}$$

This has the form of the Newtonian equation of motion in Cartesian coordinates,

$$\frac{\mathrm{d}^2x^i}{\mathrm{d}t^2} = -\frac{\Phi}{x^i}, \tag{8.13}$$

where Φ is the Newtonian gravitational potential, provided that we make the identification $h_{00} \approx 2\Phi/c^2$ or

$$g_{00} \approx \left(1 + \frac{2\Phi}{c^2}\right). \tag{8.14}$$

The assumption of a small perturbation to the metric holds provided that $\Phi/c^2 \ll 1$. This is an excellent approximation even for dense objects; for example, $\Phi/c^2 \sim 10^{-4}$ at the surface of a white dwarf. However, the weak-field approximation does break down in some extreme situations of astrophysical interest, such as close to the event horizon of a black hole. In the next section, we shall use the requirement (8.14) to fix the field equations in general relativity that relate the geometry of spacetime to the matter that is present.

8.2 Intrinsic Curvature of a Manifold

A manifold, or some extended region of one, is *flat* if it possible to find Cartesian coordinates X^a such that, throughout this region, the line element takes the Euclidean form

$$\mathrm{d}s^2 = \epsilon_1(\mathrm{d}X^1)^2 + \epsilon_2(\mathrm{d}X^2)^2 + \cdots + \epsilon_N(\mathrm{d}X^N)^2, \quad \epsilon_a = \pm 1. \tag{8.15}$$

Given points labelled with arbitrary coordinates in some manifold, how can we tell whether it is flat? For example, we know that the 3D space with line element

$$\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2 \tag{8.16}$$

is just 3D Euclidean space written in spherical polar coordinates, and so is flat, but how could we tell this without spotting the coordinate transformation to Cartesian coordinates?

Fortunately, we can construct a tensor-valued measure of curvature, the *Riemann curvature tensor*, which allows us to test for curvature in an arbitrary coordinate system. The physical relevance of curvature for general relativity is as follows: if a region of spacetime is flat, then we can construct global inertial coordinates, in which the metric is $\eta_{\mu\nu}$, and we recover special relativity *globally*, i.e., there is no gravitational field.

8.2.1 Riemann Curvature Tensor

We defined the covariant derivative such that its action was commutative on scalar fields, $\nabla_a\nabla_b\phi = \nabla_b\nabla_a\phi$. This is not the case for vector fields. Consider a dual-vector field v_a, and take two covariant derivatives:

$$\begin{aligned}\nabla_a\nabla_b v_c &= \partial_a(\nabla_b v_c) - \Gamma^d_{ab}\nabla_d v_c - \Gamma^d_{ac}\nabla_b v_d \\ &= \partial_a(\partial_b v_c - \Gamma^d_{bc}v_d) - \Gamma^d_{ab}(\partial_d v_c - \Gamma^e_{dc}v_e) - \Gamma^d_{ac}(\partial_b v_d - \Gamma^e_{bd}v_e).\end{aligned} \tag{8.17}$$

Now switch a and b and subtract; the term involving Γ^d_{ab} then cancels as do all terms involving derivatives of v_a, leaving

$$\nabla_a\nabla_b v_c = \nabla_b\nabla_a v_c = -\partial_a\Gamma^d_{bc}v_d + \partial_b\Gamma^d_{ac}v_d + \Gamma^e_{ac}\Gamma^d_{be}v_d - \Gamma^e_{bc}\Gamma^d_{ae}v_d. \tag{8.18}$$

We can therefore write

$$\boxed{\nabla_a\nabla_b v_c = \nabla_b\nabla_a v_c = R_{abc}{}^d v_d,} \tag{8.19}$$

where

$$\boxed{R_{abc}{}^d = -\partial_a\Gamma^d_{bc} + \partial_b\Gamma^d_{ac} + \Gamma^e_{ac}\Gamma^d_{be} - \Gamma^e_{bc}\Gamma^d_{ae}.} \tag{8.20}$$

The quotient theorem tells us that this must be a type-$(1,3)$ tensor, which we call the *Riemann curvature tensor*.

Note how the curvature tensor is constructed from the metric and its first and second derivatives (via the connection). If a manifold is flat, the Riemann curvature tensor vanishes since we can always choose Cartesian coordinates such that the connection vanishes; as the components of the Riemann curvature tensor vanish in such coordinates, the tensor itself is zero. The converse is also true: if the Riemann tensor vanishes throughout some region of a manifold, the manifold is flat in that region.

8.2.2 Symmetries of the Curvature Tensor

It follows directly from the definition (8.19) that the Riemann tensor is antisymmetric in its first two indices:

$$R_{abc}{}^{d} == -R_{bac}{}^{d}. \tag{8.21}$$

The explicit construction (8.20) reveals the cyclic symmetry,

$$R_{abc}{}^{d} + R_{cab}{}^{d} + R_{bca}{}^{d} = 0, \tag{8.22}$$

in which the first three indices are cyclically permuted. Given the antisymmetry on the first two indices, this can also be written as $R_{[abc]}{}^{d} = 0$.

There are remaining symmetries that are most easily seen by lowering the final index to form the type-$(0,4)$ tensor R_{abcd}. On a general curved manifold, let us work in local Cartesian coordinates at an arbitrary point P; then the connection vanishes at P and we have

$$(R_{abcd})_P = -(g_{de}\partial_a\Gamma^e_{bc} - g_{de}\partial_b\Gamma^e_{ac})_P. \tag{8.23}$$

The (metric) connection is, generally,

$$\Gamma^e_{bc} = \frac{1}{2}g^{ef}(\partial_b g_{cf} + \partial_c g_{bf} - \partial_f g_{bc}), \tag{8.24}$$

so that, in local Cartesian coordinates,

$$(g_{de}\partial_a\Gamma^e_{bc})_P = \frac{1}{2}(\partial_a\partial_b g_{cd} + \partial_a\partial_c g_{bd} - \partial_a\partial_d g_{bc})_P. \tag{8.25}$$

Using this in Eq. (8.23), we find

$$(R_{abcd})_P = \frac{1}{2}(\partial_a\partial_b g_{bc} + \partial_b\partial_c g_{ad} - \partial_a\partial_c g_{bd} + \partial_b\partial_d g_{ac})_P. \tag{8.26}$$

Since symmetries are preserved under general coordinate transformations, and given that the point P is arbitrary, any symmetry of the components R_{abcd} in local Cartesian coordinates at P will imply a general symmetry of the Riemann tensor. Inspection of Eq. (8.26) reveals two further symmetries:

$$R_{abcd} = -R_{abdc} \tag{8.27}$$

$$R_{abcd} = R_{cdab} \tag{8.28}$$

the first implying antisymmetry on the third and fourth indices (as for the first and second), and the second implying symmetry under swapping the first pair of indices with the second pair.

Let us consider what these symmetries mean for the number of independent components of the curvature tensor. In 1D, the curvature tensor necessarily vanishes since there is only one possible component R_{1111}, but this vanishes by antisymmetry.[1] In 2D, antisymmetry implies there is only one independent component, say R_{1212}. In 3D, there are six independent components, as the following argument shows:

[1] You might think a line can be curved, and indeed it can, but this curvature reflects the embedding in the plane and is not *intrinsic curvature*. We can always use the length along the curve as a coordinate, in which case the metric is trivially $g_{11} = 1$ everywhere implying intrinsic flatness.

- There are three distinct combinations of the first pair of indices that give non-zero components of the curvature tensor: 12, 13 and 23. The same is true for the second pair.

- Given the symmetry under swapping the first and second pair of indices, there are three independent curvature components where the pairs are the same and three where they are different (like the diagonal and off-diagonal components of a 3×3 matrix, respectively). This gives six independent components.

- Finally, we should check that the cyclic symmetry (22) does not imply any further dependencies amongst these six components. Generally, the cyclic symmetry $R_{[abc]d} = 0$ is trivially satisfied if any of abc are equal. Moreover, it is also trivial if the fourth index is equal to any of the first three, by virtue of the other symmetries (antisymmetry in the first pair and second pair of indices and symmetry under swapping these pairs). So in 3D the cyclic symmetry implies no new constraints and there are six independent components of the curvature tensor.

In 4D, there are 20 independent components. The argument is similar to that in 3D, but now there are six distinct combinations for either the first or second pair of indices, and so 21 independent components before we consider the cyclic symmetry. However, in 4D the cyclic symmetry is not trivial since it is possible for all four indices to be different. This implies one further constraint, of the form (letting indices run from $0-3$)

$$R_{0123} + R_{1203} + R_{2013} = 0, \tag{8.29}$$

which reduces the number of independent components of the curvature tensor from 21 to 20. Generally, in ND the number of independent components of the curvature tensor is $N^2(N^2-1)/12$, which is exactly the number of physical degrees of freedom in the second derivative of the metric (i.e., after accounting for the freedom to perform coordinate transformations; see Chapter 2).

8.2.3 The Bianchi Identity

The curvature tensor satisfies the differential *Bianchi identity*, which is very important for the development of general relativity:

$$\boxed{\nabla_a R_{bcd}{}^{e} + \nabla_b R_{cad}{}^{e} + \nabla_c R_{abd}{}^{e} = 0.} \tag{8.30}$$

Note that this is a tensor identity since it involves the covariant derivative. The Bianchi identity can be written in the equivalent form

$$\nabla_{[a} R_{bc]d}{}^{e} = 0 \tag{8.31}$$

using the antisymmetry of the curvature tensor in its first pair of indices.

It is simplest to prove the Bianchi identity by working in local Cartesian coordinates at some arbitrary point P; as it is a tensor identity, if we can show that it holds in one

coordinate system it will necessarily hold in all. Since the covariant derivative reduces to a simple covariant derivative in local Cartesian coordinates, we have

$$\begin{aligned}(\nabla_a R_{bcd}{}^e)_P &= \Big(\partial_a\Big[-\partial_b\Gamma^e_{cd} + \partial_c\Gamma^e_{bd} + \Gamma^f_{bd}\Gamma^e_{cf} - \Gamma^f_{cd}\Gamma^e_{bf}\Big]\Big)_P \\ &= (-\partial_a\partial_b\Gamma^e_{cd} + \partial_a\partial_c\Gamma^e_{bd})_P\end{aligned} \tag{8.32}$$

Adding in the cyclic permutations of a, b and c, the righthand side vanishes thus proving the Bianchi identity.

8.2.4 Ricci Tensor and Ricci Scalar

Lower-rank tensors can be formed from the curvature tensor by contraction. Given the antisymmetry $R_{abcd} = R_{[ab]cd} = R_{ab[cd]}$, the only option is to contract across the first and second pair; the contraction on the first and last indices[2] defines the Ricci tensor

$$R_{ab} \equiv R_{cab}{}^c. \tag{8.33}$$

The Ricci tensor is symmetric, $R_{ab} = R_{ba}$, which follows from contracting the cyclic identity:

$$\begin{aligned}0 &= \delta^c_d\Big(R_{abc}{}^d + R_{cab}{}^d + R_{bca}{}^d\Big) \\ &= R_{ab} - R_{ba},\end{aligned} \tag{8.34}$$

where we used $R_{abcd} = R_{[ab]cd} = R_{ab[cd]}$. We can contract the Ricci tensor to obtain the *Ricci scalar* (or curvature scalar):

$$R \equiv g^{ab}R_{ab}. \tag{8.35}$$

If a manifold is flat in some region, the curvature tensor will vanish and hence so will the Ricci tensor and scalar. However, it is possible for the Ricci tensor to vanish but for the full curvature tensor to be non-zero and the manifold to be curved – we shall see that this situation generally arises in vacuum regions of spacetime, where only tidal gravitational effects arise.

Contracting the Bianchi identity gives

$$\begin{aligned}0 &= \delta^b_e(\nabla_a R_{bcd}{}^e + \nabla_b R_{cad}{}^e + \nabla_c R_{abd}{}^e) \\ &= \nabla_a R_{cd} - \nabla_c R_{ad} + \nabla^b R_{cadb}.\end{aligned} \tag{8.36}$$

Taking a further contraction over a and d gives

$$\begin{aligned}0 &= g^{ad}\Big(\nabla_a R_{cd} - \nabla_c R_{ad} + \nabla^b R_{cadb}\Big) \\ &= \nabla^d R_{cd} - \nabla_c R + \nabla^b R_{cb} \\ &= 2\nabla^d R_{cd} - \nabla_c R.\end{aligned} \tag{8.37}$$

This gives the *contracted Bianchi identity*:

$$\boxed{\nabla^a\left(R_{ab} - \frac{1}{2}g_{ab}R\right) = 0,} \tag{8.38}$$

[2] This choice is not universal in the literature; often the first and third are chosen instead, which reverses the sign of the Ricci tensor.

which involves the divergence of the *Einstein tensor*

$$\boxed{G_{ab} \equiv R_{ab} - \frac{1}{2} g_{ab} R.} \tag{8.39}$$

The Einstein tensor is symmetric and divergence-free. We shall see that the contracted Bianchi identity is related to the conservation of energy and momentum in general relativity.

8.3 Physical Manifestations of Curvature

8.3.1 Curvature and Parallel Transport

We noted in Chapter 5 that parallel transport is generally path dependent, the exception being when the manifold is flat. We can relate the curvature of the manifold, as expressed by the curvature tensor, to the path dependence of parallel transport by considering an infinitesimal loop defined by a curve $\mathcal{C}$. Let $\mathcal{C}$ be parameterised by $x^a(u)$, and consider parallel transporting a vector $\mathbf{v}$ around this loop from the point P back to itself. The equation of parallel transport is

$$\frac{\mathrm{d}v^a}{\mathrm{d}u} = -\Gamma^a_{bc} \frac{\mathrm{d}x^b}{\mathrm{d}u} v^c, \tag{8.40}$$

and so, at the point with coordinates $x^a(u)$, the result of parallel transporting $\mathbf{v}$ from P is the vector with components

$$v^a(u) = v^a_P - \int_{u_P}^{u} \Gamma^a_{bc} \frac{\mathrm{d}x^b}{\mathrm{d}u'} v^c \, \mathrm{d}u' \,. \tag{8.41}$$

Since the closed loop is small, we can expand the connection and components v^a in the right-hand side to first order in the coordinate differences $x^a(u) - x^a_P$ as

$$\Gamma^a_{bc}(u) = (\Gamma^a_{bc})_P + (\partial_d \Gamma^a_{bc})_P [x^d(u) - x^d_P] + \cdots \,, \tag{8.42}$$

$$v^c(u) = v^c_P - (\Gamma^c_{ef})_P \, v^f_P [x^e(u) - x^e_P] + \cdots \,. \tag{8.43}$$

Substituting these into Eq. (8.41), for the closed path

$$\Delta v^a = -(\partial_d \Gamma^a_{bc} - \Gamma^a_{be}\Gamma^e_{dc})_P \, v^c_P \oint x^d \, \mathrm{d}x^b \,, \tag{8.44}$$

where we have retained terms to first order in $x^a(u) - x^a_P$ and used $\oint \mathrm{d}x^b = 0$. We can simplify this further by noting that

$$\oint \mathrm{d}(x^b x^d) = 0 \quad \Longrightarrow \quad \oint x^b \, \mathrm{d}x^d = -\oint x^d \, \mathrm{d}x^b \tag{8.45}$$

so that

$$\Delta v^a = (\partial_d \Gamma^a_{bc} - \Gamma^a_{be}\Gamma^e_{dc})_P \, v^c_P \oint x^{[b} \, \mathrm{d}x^{d]} \,, \tag{8.46}$$

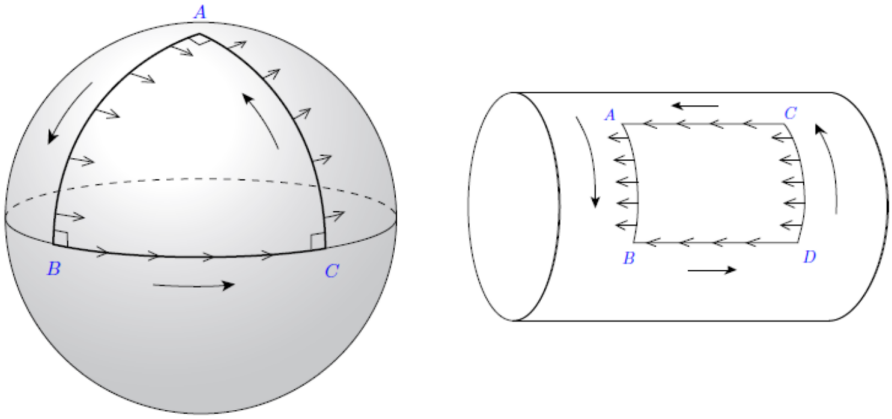

Fig. 8.1: Parallel transport of a vector around closed loops on the 2-sphere (left) and the surface of a cylinder embedded in $\mathbb{R}^3$ (right). The 2-sphere has (constant) intrinsic curvature and as a result a vector is rotated after undergoing parallel transport around a closed loop. In contrast, the cylinder has vanishing intrinsic curvature and a vector is unchanged by parallel transport around a closed loop.

We can now antisymmetrise over the indices b and d in the first-term on the right to find

$$\Delta v^a = \frac{1}{2}\underbrace{(\partial_d \Gamma^a_{bc} - \partial_b \Gamma^a_{dc} - \Gamma^a_{be}\Gamma^e_{dc} + \Gamma^a_{de}\Gamma^e_{bc})_P}_{-(R_{dbc}{}^a)_P}\, v^c_P \oint x^{[b}\, \mathrm{d}x^{d]}\,. \tag{8.47}$$

Finally, relabelling indices, we have

$$\Delta v^a = \frac{1}{2}(R_{bcd}{}^a)_P\, v^d_P \oint x^{[b}\, \mathrm{d}x^{c]}\,. \tag{8.48}$$

For an infinitesimal loop, the integral $\oint x^{[b}\, \mathrm{d}x^{c]}$ is a type-$(2,0)$ tensor at P that encodes the planar area of the loop, so the right-hand side of Eq. (8.48) (48) is a vector there, as required. We see that the vector $\mathbf{v}$ does not change on parallel transport around a closed loop near P if the curvature tensor vanishes there. These ideas are illustrated in Fig. 8.1 for the case of a curved manifold (the 2-sphere) and a manifold with no intrinsic curvature (the surface of a cylinder embedded in $\mathbb{R}^3$).

8.3.2 Curvature and Geodesic Deviation

A further important consequence of curvature is that two nearby geodesics that are initially parallel will either converge or diverge depending on the local curvature. Consider two nearby affinely-parameterised geodesics, $\mathcal{C}$ given by $x^a(u)$, and $\bar{\mathcal{C}}$ given by $\bar{x}^a(u)$. Let the initial values of the affine parameters be chosen so that the coordinate difference $\xi^a(u) = \bar{x}^a(u) - x^a(u)$ is infinitesimal (see Fig. 8.2). For infinitesimal separations, the ξ^a form the components of a vector.

Let us consider how $\xi^a(u)$ changes with u. Since $\mathcal{C}$ and $\bar{\mathcal{C}}$ are geodesics, we have

$$\frac{\mathrm{d}^2 x^a}{\mathrm{d}u^2} + \Gamma^a_{bc}\frac{\mathrm{d}x^b}{\mathrm{d}u}\frac{\mathrm{d}x^c}{\mathrm{d}u} = 0 \quad \text{and} \quad \frac{\mathrm{d}^2 \bar{x}^a}{\mathrm{d}u^2} + \Gamma^a_{bc}\frac{\mathrm{d}\bar{x}^b}{\mathrm{d}u}\frac{\mathrm{d}\bar{x}^c}{\mathrm{d}u} = 0, \tag{8.49}$$

where the bar on the metric connection denotes that it is evaluated at the point with coordinates $\bar{x}^a(u)$. Taking the difference of the barred and unbarred geodesic equations,

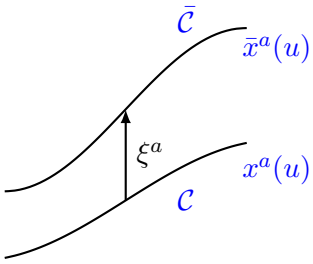

Fig. 8.2: Two nearby affinely-parameterised geodesics, $\mathcal{C}$ given by $x^a(u)$, and $\bar{\mathcal{C}}$ given by $\bar{x}^a(u)$. The initial values of the affine parameters are chosen so that the coordinate difference $\xi^a(u) = \bar{x}^a(u) - x^a(u)$ is infinitesimal.

we have

$$\frac{\mathrm{d}^2\xi^a}{\mathrm{d}u^2} + \bar{\Gamma}^a_{bc}\dot{\bar{x}}^b\dot{\bar{x}}^c - \Gamma^a_{bc}\dot{x}^b\dot{x}^c = 0, \tag{8.50}$$

where overdots denote derivatives with respect to u. Expanding to first order in $\xi^a(u)$, we have

$$\bar{\Gamma}^a_{bc}(u) = \Gamma^a_{bc}(u) + \partial_d\Gamma^a_{bc}\xi^d, \tag{8.51}$$

and so

$$\frac{\mathrm{d}^2\xi^a}{\mathrm{d}u^2} + 2\Gamma^a_{bc}\dot{x}^b\dot{\xi}^c + \partial_d\Gamma^a_{bc}\dot{x}^b\dot{x}^c\xi^d = 0. \tag{8.52}$$

We now look to write this in terms of objects that are manifestly tensor-valued; we use

$$\begin{aligned}\frac{\mathrm{D}}{\mathrm{D}u}\left(\frac{\mathrm{D}\xi^a}{\mathrm{D}u}\right) &= \frac{\mathrm{d}}{\mathrm{d}u}\left(\dot{\xi}^a + \Gamma^a_{bc}\dot{x}^b\xi^c\right) + \Gamma^a_{bc}\dot{x}^b\left(\dot{\xi}^c + \Gamma^c_{de}\dot{x}\xi^e\right)\\ &= \frac{\mathrm{d}^2\xi^a}{\mathrm{d}u^2} + \partial_d\Gamma^a_{bc}\dot{x}^b\dot{x}^d\xi^c + \Gamma^a_{bc}\ddot{x}^b\xi^c + 2\Gamma^a_{bc}\dot{x}^b\dot{\xi}^c + \Gamma^a_{bc}\Gamma^c_{de}\dot{x}^b\dot{x}^d\xi^e.\end{aligned} \tag{8.53}$$

Eliminating $\ddot{x}^b$ with the geodesic equation, and substituting into Eq. (8.52), after some index relabelling we find

$$\frac{\mathrm{D}}{\mathrm{D}u}\left(\frac{\mathrm{D}\xi^a}{\mathrm{D}u}\right) + \underbrace{\partial_d\Gamma^a_{bc} - \partial_b\Gamma^a_{dc} - \Gamma^e_{dc}\Gamma^a_{be} + \Gamma^e_{bc}\Gamma^a_{de}}_{-R_{dbc}{}^a}\dot{x}^b\dot{x}^c\xi^d = 0. \tag{8.54}$$

We see that the evolution of the connecting vector is given by the equation of *geodesic deviation*:

$$\frac{\mathrm{D}}{\mathrm{D}u}\left(\frac{\mathrm{D}\xi^a}{\mathrm{D}u}\right) - R_{dbc}{}^a\dot{x}^b\dot{x}^c\xi^d = 0. \tag{8.55}$$

If a manifold is flat over some region, the curvature tensor vanishes there and, in Cartesian coordinates, the intrinsic derivative reduces to the ordinary derivative $\mathrm{d}/\mathrm{d}u$ and Eq. (8.55) reduces to $\mathrm{d}^2\xi^a/\mathrm{d}u^2 = 0$ – the Cartesian components of ξ^a grow linearly with u, as expected. However, if there is curvature, two geodesics that are initially parallel (i.e., $\mathrm{D}\xi^a/\mathrm{D}u = 0$) will converge or diverge due to the curvature. This latter behaviour is familiar on the 2-sphere, where neighbouring lines of longitude are geodesics that intersect at the poles, but are parallel at the equator.

In spacetime, the geodesic deviation equation describes the *relative* acceleration of neighbouring free-falling particles due to *tidal* gravitational effects. Consider free-falling

particles with geodesics $x^\mu(\tau)$ and $\bar{x}^\mu(\tau)$, where τ is proper time for each particle. The connecting vector $\xi^u(\tau) \equiv \bar{x}^u(\tau) - x^u(\tau)$ evolves according to

$$\frac{\mathrm{D}}{\mathrm{D}u}\left(\frac{\mathrm{D}\xi^u}{\mathrm{D}\tau}\right) = \underbrace{R_{\nu\alpha\beta}{}^{\mu} u^\alpha u^\beta}_{S_\nu{}^\mu} \xi^\nu, \tag{8.56}$$

where $u^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$ is the 4-velocity and we have introduced the tidal tensor $S_\nu{}^\mu$. Note that $S_\nu{}^\mu$ is symmetric. Eq. (8.56) is analogous to the result for the tidal acceleration in Newtonian gravity, as we now show.

Take two free-falling particles with neighbouring trajectories $x^i(t)$ and $\bar{x}^i(t)$ in Cartesian coordinates, and define their connecting vector as $\xi^i(t) \equiv \bar{x}^i(t) - x^i(t)$. The particles free-fall in a gravitational potential Φ as

$$\frac{\mathrm{d}^2 x^i}{\mathrm{d}t^2} = -\left(\frac{\partial\Phi}{\partial x^i}\right)_{x(t)} \quad \text{and} \quad \frac{\mathrm{d}^2 \bar{x}^i}{\mathrm{d}t^2} = -\left(\frac{\partial\Phi}{\partial x^i}\right)_{\bar{x}(t)} \tag{8.57}$$

Taking the difference, and expanding to first-order in ξ^i, gives

$$\frac{\mathrm{d}^2 \xi^i}{\mathrm{d}t^2} \approx -\left(\frac{\partial^2\Phi}{\partial x^i \partial x^j}\right)\xi^j. \tag{8.58}$$

In free space, where $\vec{\nabla}^2\Phi = 0$, the tidal tensor $\partial_i\partial_j\Phi$ is symmetric and trace-free and generates a volume-preserving distortion of a set of free-falling particles. It can be shown in the weak-field limit, and for slow speeds, the Newtonian tidal equation (8.58) and the appropriate components of the geodesic deviation equation (8.55) are equivalent.

Chapter 9

The Gravitational Field Equations

We have seen how to formulate the laws of physics on a curved spacetime as tensor equations, thus ensuring consistency with the equivalence principle. In this chapter, we complete the development of general relativity by specifying how the curvature of spacetime is related to the matter that is present. This will lead us to the Einstein equations, relating the energy–momentum tensor of the matter to the Einstein tensor.

In terms of components, these are non-linear, second-order partial differential equations for the metric functions, much as Maxwell's equations can be regarded as (linear) second-order equations relating the 4-vector potential to the current 4-vector.

9.1 The Energy–Momentum Tensor

In Newtonian gravity, the Poisson equation relates the Laplacian of the gravitational potential to the mass density. In general relativity, we need to find an appropriate tensor that generalises the mass density to describe the relativistic energy at each event in spacetime.

First consider the case of non-interacting particles, each of rest mass m, with no velocity dispersion. Such matter is usually referred to as dust; it has the property that at each event P all particles present there have the same 4-velocity $u^\mu(x)$. At P, the dust has *energy density* ρc^2 when measured in some local-inertial frame there, and the particles all have 3-velocity $\vec{u}$. In particular, it is possible to find a local-inertial frame in which the particles at P are at rest; in this instantaneous rest frame, let the number density of particles be n_0 so that the energy density is $\rho_0 c^2 = m n_0 c^2$.

Transforming to the local-inertial frame S in which the 3-velocity of the dust is $\vec{u}$ at P, the number density will be $\gamma_u n_0$ (length contraction) and the energy of each particle $\gamma_u m c^2$. It follows that in this frame, the energy density is

$$\rho c^2 = (\gamma_u n_0)\gamma_u m c^2 = \gamma_u^2 \rho_0 c^2. \tag{9.1}$$

We see that energy density is not a Lorentz scalar; rather it transforms like the 00 component of the type-$(2,0)$ tensor

$$T^{\mu\nu}(x) = \rho_0(x) u^\mu(x) u^\nu(x). \tag{9.2}$$

That this is a tensor follows from the fact that $\rho_0 c^2$ is a scalar field (it is *defined* to be the energy density in the instantaneous rest frame), and u^μ are the components of a 4-vector. In the local-inertial frame S, in which the 3-velocity is $\vec{u}$, the 0-component of u^μ is $\gamma_u c$, so that

$$T^{00} = \gamma_u^2 \rho_0 c^2, \tag{9.3}$$

as advertised.

What is the physical interpretation of the other components of $T^{\mu\nu}$? Again, consider observing in the local-inertial frame S in which the 3-velocity at P is $\vec{u}$; the 4-velocity of the dust has components $u^\mu = \gamma_u(c, \vec{u})$ and so

$$\begin{aligned} T^{i0} &= mn_0(\gamma_u \vec{u}^i)(\gamma_u c) \\ &= c(\gamma_u n_0)(m\gamma_u \vec{u}^i). \end{aligned} \tag{9.4}$$

This is the product of the number density of particles and the 3-momentum of each, and so is the *momentum density* (times c). An alternative interpretation is as the energy flux in the ith direction, since

$$\text{energy flux} = (\gamma_u^2 n_0 mc^2)\vec{u}^i = cT^{i0}, \tag{9.5}$$

i.e., the product of the energy density and the particle 3-velocity. Finally, consider the ij components:

$$\begin{aligned} T^{ij} &= mn_0(\gamma_u \vec{u}^i)(\gamma_u \vec{u}^j) \\ &= (\gamma_u^2 mn_0 \vec{u}^i)\vec{u}^j. \end{aligned} \tag{9.6}$$

This is the ith component of the 3-momentum density multiplied by the jth component of the 3-velocity, i.e., the *flux* of the ith component of 3-momentum along the jth direction.

Collecting these results together, we have the following physical interpretation of the components of the tensor $T^{\mu\nu}$ in a local-inertial frame:

1. T^{00} energy density,
2. T^{i0} ith component of 3-momentum density (times c),
3. T^{ij} flux of i-component of 3-momentum in j-direction.

This association of a type-$(2,0)$ tensor $T^{\mu\nu}$ with the energy and momentum density and their fluxes generalises to other sources, such as the electromagnetic field. The properties of these densities and fluxes under Lorentz transformations ensure that they also form the components of a tensor for any source. The tensor $T^{\mu\nu}$ is called the *energy–momentum tensor* (alternatively, the stress–energy tensor) and, as we shall see, acts as the source for spacetime curvature. Note that the energy–momentum tensor is always symmetric, $T^{\mu\nu} = T^{\nu\mu}$; this can be shown to be necessary to ensure angular momentum conservation in all inertial frames.

9.1.1 Energy–Momentum Tensor of an Ideal Fluid

The energy density that appears in the energy–momentum tensor must include all sources of energy, for example, the kinetic energy of the particles in a gas due to their velocity dispersion and any interaction energies. Similarly, the energy flux and momentum density must include contributions from heat conduction in the gas, as well as from bulk motions (the former giving non-zero components T^{i0} in the instantaneous rest frame). For the flux of momentum, we must include effects from the velocity dispersion in the gas and any shear stresses that may be present.

In this course, we shall only consider *ideal fluids*. These have the property that at any event one can find a local-inertial frame (the instantaneous rest frame) in which $T^{i0} = 0$ and in which the spatial components are isotropic: $T^{ij} \propto \delta^{ij}$. In the case of a gas of particles, this requires the mean-free path in the gas to be short compared to the length scale of any temperature or velocity gradients (so that conduction and shear stresses are negligible). In the instantaneous rest frame, the components of the energy–momentum tensor for an ideal fluid have to take the form

$$T^{\mu\nu} = \mathrm{diag}(\rho c^2, p, p, p), \tag{9.7}$$

where ρc^2 is the rest-frame energy density (we now drop the subscript 0 on the rest-frame energy density that we included earlier) and p is the *isotropic pressure*. We can write this in tensor form, and so valid in any coordinate system, as

$$\boxed{T^{\mu\nu} = \left(\rho + \frac{p}{c^2}\right) u^\mu u^\nu - p g^{\mu\nu},} \tag{9.8}$$

where u^μ is the 4-velocity of the fluid (defined by its instantaneous rest frame). In the instantaneous rest frame, $u^\mu = (c, \vec{0})$ and $g^{\mu\nu} = \eta^{\mu\nu}$, and so we recover the correct diagonal components [Eq. (9.7)] from Eq. (9.8). Note that the energy density ρc^2 and isotropic pressure p are scalar fields as they are defined in the instantaneous rest frame. Note also that for $p \ll \rho c^2$, the energy–momentum tensor for an ideal fluid reduces to that for dust.

9.1.2 Conservation of Energy and Momentum

Recall that in electromagnetism, the conservation of charge is expressed through the (covariant) continuity equation $\nabla_\mu j^\mu = 0$. The energy–momentum tensor satisfies a similar continuity equation:

$$\boxed{\nabla_\mu T^{\mu\nu} = 0.} \tag{9.9}$$

In local-inertial coordinates at some event P, this reduces to

$$\frac{\partial T^{00}}{\partial t} + c \sum_i \frac{\partial T^{i0}}{\partial x^i} = 0, \tag{9.10}$$

$$\frac{\partial T^{i0}/c}{\partial t} + \sum_j \frac{\partial T^{ij}}{\partial x^j} = 0. \tag{9.11}$$

The first, Eq. (9.10) expresses conservation of energy through a continuity equation of the form

$$\frac{\partial}{\partial t}(\text{energy density}) + \vec{\nabla} \cdot (\text{energy flux}) = 0. \tag{9.12}$$

The second, Eq. (9.11), expresses conservation of momentum in the form

$$\frac{\partial}{\partial t}(\text{momentum density}) + \vec{\nabla} \cdot (\text{momentum flux}) = 0. \tag{9.13}$$

The covariant continuity equation (9.9) thus encodes both conservation of energy and 3-momentum.

9.1.2.1 Energy and Momentum Conservation for the Ideal Fluid

For the ideal fluid, with energy–momentum tensor given by Eq. (9.8), conservation of the energy–momentum tensor gives

$$\nabla_\mu \left[\left(\rho + \frac{p}{c^2} \right) u^\mu u^\nu = p g^{\mu\nu} \right] = 0, \tag{9.14}$$

so that

$$u^\nu u^\mu \nabla_\mu \left(\rho + \frac{p}{c^2} \right) + \left(\rho + \frac{p}{c^2} \right) (\nabla_\mu u^\mu) u^\nu + \left(\rho + \frac{p}{c^2} \right) u^\mu \nabla_\mu u^\nu - \nabla^\nu p = 0, \tag{9.15}$$

where we have used $\nabla_\mu g^{\mu\nu} = 0$. We can project this parallel and perpendicular to u^ν. Contracting with u^ν extracts the parallel component:

$$\begin{aligned} c^2 u^\mu \nabla_\mu \left(\rho + \frac{p}{c^2} \right) + c^2 \left(\rho + \frac{p}{c^2} \right) (\nabla_\mu u^\mu) - u^\mu \nabla_\mu p = 0 \\ \implies \nabla_\mu (\rho u^\mu) + \frac{p}{c^2} \nabla_\mu u^\mu = 0. \end{aligned} \tag{9.16}$$

Here, we have used that $u^\mu \nabla_\mu u^\nu$ is orthogonal to u^ν because of $g_{\nu\rho} u^\nu u^\rho = c^2$:

$$\begin{aligned} u^\mu \nabla_\mu (g_{\nu\rho} u^\nu u^\rho) = 0 \\ \implies g_{\nu\rho} (u^\mu \nabla_\mu u^\nu) u^\rho + g_{\nu\rho} u^\nu (u^\mu \nabla_\mu u^\rho) = 0 \\ \implies 2 u_\nu (u^\mu \nabla_\mu u^\nu) = 0. \end{aligned} \tag{9.17}$$

The part of Eq. (9.15) that is perpendicular to u^ν can now be extracted by subtracting u^ν times the parallel component; we find

$$\left(\rho + \frac{p}{c^2} \right) u^\mu \nabla_\mu u^\nu = \left(g^{\mu\nu} = \frac{u^\mu u^\nu}{c^2} \right) \nabla_\mu p. \tag{9.18}$$

These parallel and perpendicular projections describe energy and momentum conservation, respectively.

We can gain physical insight into these equations by adopting local-inertial coordinates in the vicinity of some event P. The metric is $\eta_{\mu\nu}$ at P, and is close to this in the immediate vicinity of P (with the extent of this region determined by the spacetime curvature). For the 4-velocity of the fluid, we have

$$u^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} = \frac{\mathrm{d}t}{\mathrm{d}\tau} \left(c, \frac{\mathrm{d}x^i}{\mathrm{d}t} \right) = \frac{\mathrm{d}t}{\mathrm{d}\tau}. \tag{9.19}$$

We can determine $\mathrm{d}t/\mathrm{d}\tau$ in terms of the 3-velocity $\vec{u}^i$ using the normalisation $g_{\mu\nu} u^\mu u^\nu = c^2$. Consider the Newtonian limit $|\vec{u}^i| \ll c$; then we have $\mathrm{d}t/\mathrm{d}\tau \approx 1$ and

$$u^\mu \approx (c, \vec{u}^i). \tag{9.20}$$

We further assume that $p \ll \rho c^2$ (i.e., the speed of all particles in the rest-frame of the fluid are much less than c). Since the metric connection vanishes at P, Eq. (9.16) reduces there to

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x^i} \left(\rho \vec{u}^i \right) \approx 0. \tag{9.21}$$

Given that ρ is approximately the mass density in this Newtonian limit, we see that we recover the *continuity equation* of Newtonian fluid mechanics. For the perpendicular projection, Eq. (9.18), we have at P

$$\rho u^\mu \frac{\partial u^\nu}{\partial x^\mu} = \left(\eta^{\mu\nu} - \frac{u^\mu u^\nu}{c^2}\frac{\partial p}{\partial x^\mu}\right) \tag{9.22}$$

The $\nu = 0$ component vanishes identically at leading order in $p/(\rho c^2)$ and $|\vec{u}^i|/c$. For $\nu = i$, we have

$$\rho\left(\frac{\partial \vec{U}^i}{\partial t} + \sum_j \vec{u}^j \frac{\partial \vec{u}^i}{\partial x^j}\right) \approx -\sum_j \delta^{ij}\frac{\partial p}{\partial x^j} - \frac{\vec{u}^i}{c^2}\frac{\partial p}{\partial t}, \tag{9.23}$$

which, on retaining the leading-order terms, gives

$$\frac{\partial \vec{u}}{\partial t} + \vec{u}\cdot\vec{\nabla}\vec{u} \approx -\frac{\vec{\nabla} p}{\rho}. \tag{9.24}$$

This is the *Euler equation* (for an ideal fluid) of Newtonian fluid mechanics.

9.2 The Einstein Equations

Recall that in Newtonian gravity, the gravitational potential Φ and mass density ρ are related by Poisson's equation

$$\nabla^2\Phi = 4\pi G\rho. \tag{9.25}$$

We also know from Subsection 8.1.2 that in the weak-field limit, where $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ with $|h_{\mu\nu}| \ll 1$, we recover the correct Newtonian equation of motion for a test particle from the geodesic equation if

$$g_{00} \approx \left(1 + \frac{2\Phi}{c^2}\right). \tag{9.26}$$

We have seen that the proper relativistic treatment of the distribution of matter is through the energy–momentum tensor, and that for a slow-moving fluid $T_{00} \approx \rho c^2$ in local-inertial coordinates. Putting these results together, the Poisson equation in the limit of weak fields and non-relativistic speeds is equivalent to

$$\vec{\nabla}^2 g_{00} \approx \frac{8\pi G}{c^4} T_{00}. \tag{9.27}$$

The second derivative of the metric is a measure of the curvature of spacetime, so this suggests we look for a relativistic field equation of the form

$$K_{\mu\nu} = \kappa T_{\mu\nu} \quad \text{and} \quad \kappa \equiv \frac{8\pi G}{c^4}, \tag{9.28}$$

where $K_{\mu\nu}$ is a symmetric type-$(0,2)$ tensor related to the curvature of spacetime (with dimensions of inverse-squared length). The energy–momentum tensor is conserved, $\nabla_\mu T^{\mu\nu} = 0$, and if this is to be consistent with the field equation (9.28) we must have

$$\nabla^\mu K_{\mu\nu} = 0. \tag{9.29}$$

We know a good candidate for $K_{\mu\nu}$: the contracted Bianchi identity (see Subsection 8.2.3) in spacetime is

$$\nabla^{\mu}\left(R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R\right) = 0, \tag{9.30}$$

so the (symmetric) Einstein tensor, $G_{\mu\nu} \equiv R_{\mu\nu} - g_{\mu\nu}R/2$, identically satisfies

$$\nabla^{\mu}G_{\mu\nu} = 0. \tag{9.31}$$

This suggests the field equation

$$\boxed{G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = -\kappa T_{\mu\nu}.} \tag{9.32}$$

The constant of proportionality (and the minus sign) on the right is required for consistency with the weak-field limit, as we shall prove shortly.

Eq. (9.32) is the *Einstein field equation* of general relativity. In spacetime, the symmetric tensors $G_{\mu\nu}$ and $T_{\mu\nu}$ have 10 independent components so the Einstein field equations are really 10 non-linear partial differential equations for the metric functions $g_{\mu\nu}$ (and so are hard to solve in general!). In contrast, Newtonian gravity has the single Poisson equation, which is linear in the potential Φ.

We can obtain an alternative (but equivalent) form of the Einstein equations by contracting with $g^{\mu\nu}$:

$$\begin{aligned} & g^{\mu\nu}\left(R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R\right) = -\kappa g^{\mu\nu}T^{\mu\nu} \\ \Longrightarrow\ & R - 2R = -\kappa T, \end{aligned} \tag{9.33}$$

where

$$T \equiv g^{\mu\nu}T_{\mu\nu} = T^{\nu}{}_{\nu} \tag{9.34}$$

is the trace of the energy–momentum tensor. It follows that

$$R_{\mu\nu} = -\kappa\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\right). \tag{9.35}$$

9.2.1 The Einstein Equations in Empty Space

The energy–momentum tensor must include all sources of energy and momentum. In some empty region of spacetime (i.e., vacuum), the energy–momentum tensor vanishes and we have

$$R_{\mu\nu} = 0 \qquad \text{(vacuum)}. \tag{9.36}$$

Generally, the vanishing of the Ricci tensor does not imply that spacetime is flat in that region; it is still possible for the Riemann tensor to be non-zero and so there to be gravitational tidal effects in the region. Interestingly, four spacetime dimensions is the minimum in which gravitational effects do not necessarily vanish in empty space.

Interestingly, four spacetime dimensions is the minimum in which gravitational effects do not necessarily vanish in empty space. To see this, note that general relativity in ND

in empty space would imply $N(N+1)/2$ field equations (the independent components of $R_{ab} = 0$), while the Riemann curvature tensor would have $N^2(N^2-1)/12$ independent components. Only for $N \geq 4$ is the number of independent components in the curvature tensor larger than the number of field equations, which is necessary to have a non-zero Riemann tensor.

9.3 Weak-Field Limit of Einstein's Equations

We aim to show that we recover Poisson's equation (9.25) from the Einstein equations in the weak-field limit. Our starting point is

$$g_{00} = -\kappa\left(T_{00} - \frac{1}{2}g_{00}T\right). \tag{9.37}$$

For the matter source, we consider a non-relativistic fluid ($p \ll \rho c^2$) that is slowly moving in coordinates in which the metric takes the weak-field form, $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$. Moreover, we assume that the mass distribution and metric are independent of time in these coordinates (a *stationary* situation).

Neglecting p compared to ρc^2, the energy-momentum tensor is

$$T_{\mu\nu} \approx \rho u_\mu u_\nu, \tag{9.38}$$

and so

$$T \approx g^{\mu\nu}\rho u_\mu u_\nu = \rho c^2. \tag{9.39}$$

Also, using $u^\mu \approx (c, \vec{u}^i)$, we have

$$u_0 = g_{0\mu}u^\mu \approx g_{00}c \approx c, \tag{9.40}$$

so that

$$T_{00} \approx \rho u_0 u_0 \approx \rho c^2. \tag{9.41}$$

Putting these pieces together, Eq. (9.37) reduces to

$$R_{00} \approx -\frac{1}{2}\kappa\rho c^2. \tag{9.42}$$

For the approximate Ricci tensor, we use

$$R_{00} = -\partial_\mu\Gamma^\mu_{00} + \partial_0\Gamma^\mu_{\mu 0} + \Gamma^\nu_{\mu 0}\Gamma^\mu_{0\nu} - \Gamma^\nu_{00}\Gamma^\mu_{\mu\nu}. \tag{9.43}$$

The connection coefficients are first-order in $h_{\mu\nu}$ since they vanish for the Minkowski metric, and so we can neglect the last two terms involving products of the connection. It follows that for a stationary metric, we have

$$R_{00} \approx -\sum_i \frac{\partial\Gamma^i_{00}}{\partial x^i}. \tag{9.44}$$

We showed in Subsection 8.1.2 that

$$\Gamma^i_{00} \approx \frac{1}{2}\frac{\partial h_{00}}{\partial x^i}, \tag{9.45}$$

and so

$$R_{00} = -\frac{1}{2}\vec{\nabla}^2 h_{00}. \tag{9.46}$$

Finally, the 00 component of the Einstein equation becomes

$$\vec{\nabla}^2 h_{00} \approx \frac{8\pi G}{c^2}\rho, \tag{9.47}$$

which, on recalling the identification $h_{00} \approx 2\Phi/c^2$ that is required from the geodesic equation, gives the usual Poisson equation

$$\vec{\nabla}^2 \Phi = 4\pi G\rho. \tag{9.48}$$

This justfies our choice of proportionality constant on the right-hand side of the Einstein field equation (9.32).

9.4 The Cosmological Constant

The Einstein equation (9.32) is the simplest field equation that we can write down if we demand that the tensor on the left has the following properties:

1. divergence free;
2. constructed from the metric and its first two derivatives; and
3. linear in second derivatives of the metric.

However, the Einstein tensor is not the only such tensor that satisfies these conditions. It can be shown (*Lovelock's theorem*) that the only other possibility is to add a constant multiple of the metric tensor, giving a new field equation of the form

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \Lambda g_{\mu\nu} = -\kappa T_{\mu\nu}. \tag{9.49}$$

The tensor on the left is still divergence free since $\nabla_\rho g_{\mu\nu} = 0$.

The quantity Λ is called the *cosmological constant*; it would be a new universal constant of nature with dimensions of inverse length squared. Einstein originally included this term in the field equations in order to construct static cosmological solutions (see Chapter 13). However, with the discovery of the expanding universe he dismissed the new term $\Lambda g_{\mu\nu}$, describing its introduction as his "biggest blunder". We now know from several different cosmological observations (including the observed fluxes of distant supernova, the clustering of galaxies, and the fluctuations in the cosmic microwave background) that Λ is non-zero, with

$$\begin{aligned} \Lambda &\approx 1.1 \times 10^{-52}\mathrm{m}^{-2} \\ &= (3.04\mathrm{Gpc})^{-2} \end{aligned} \tag{9.50}$$

where 1Gpc (Gigaparsec) equals 3.1×10^{25}m. This is a very large length scale – comparable to the size of the observable universe – so the cosmological constant is a small correction

in a system much smaller than this (i.e., where the curvature scale is small compared to $\Lambda^{-1/2}$).

Contracting Eq. (9.49) with $g^{\mu\nu}$, we have

$$R - 4\Lambda = \kappa T, \tag{9.51}$$

so that

$$R_{\mu\nu} = -\kappa\left(t_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\right) + \Lambda g_{\mu\nu}. \tag{9.52}$$

Repeating the calculation of the weak-field limit in the presence of small Λ gives

$$\vec{\nabla}^2\Phi = 4\pi G\rho - \Lambda c^2 \tag{9.53}$$

The solution for a point source of mass M at rest at the origin gives

$$-\vec{\nabla}\Phi(\vec{x}) = -\frac{GM}{|\vec{x}|^3}\vec{x} + \frac{\Lambda c^2}{3}\vec{x}, \tag{9.54}$$

showing that Λ provides a *gravitational repulsion* that increases linearly with distance from the central mass. Although the cosmological constant is irrelevant on many scales of interest, it is important at current times on the scale of the entire universe. The repulsive nature of the cosmological constant may be responsible for the observed late-time acceleration in the expansion of our universe.[1]

9.4.1 The Cosmological Constant as Vacuum Energy

The energy–momentum tensor for an ideal fluid is

$$T^{\mu\nu} = \left(\rho + \frac{p}{c^2}\right)u^\mu u^\nu - pg^{\mu\nu}. \tag{9.55}$$

Consider a fluid with the strange property that $p = -\rho c^2$ implying that it has a large negative pressure (i.e., tension); then

$$T_{\mu\nu} = -pg_{\mu\nu} = \rho c^2 g_{\mu\nu}. \tag{9.56}$$

In *any* local-inertial frame, $T^{00} = \rho c^2$, so *all* observers measure an energy density ρc^2 (and pressure $p = -\rho c^2$). Such an energy–momentum tensor would therefore have to be a fundamental property of the vacuum.

The term $\Lambda g_{\mu\nu}$ in the Einstein equation (9.49) has exactly the form of the energy–momentum tensor with $p = -\rho c^2$. We can write

$$\begin{aligned} R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R &= -\kappa\left(t_{\mu\nu} + \frac{\Lambda}{\kappa}g_{\mu\nu}\right) \\ &= -\kappa(T_{\mu\nu} + T^{\text{vac}}_{\mu\nu}), \end{aligned} \tag{9.57}$$

[1] Other alternatives are also being actively considered including scalar fields (*quintessence*) and modifications to the laws of gravitation on cosmological scales.

where we have introduced the energy–momentum tensor of the vacuum

$$T^{\rm vac}_{\mu\nu} = \rho_{\rm vac} c^2 g_{\mu\nu}, \quad \text{with} \quad \rho_{\rm vac} c^2 = \frac{\Lambda c^4}{8\pi G}. \tag{9.58}$$

Should we expect the vacuum to have a non-zero energy density? Indeed, we should because of the quantum zero-point energies of all the fields in nature.

Recall that a simple harmonic oscillator with classical frequency ω has a quantum ground-state energy $\hbar\omega/2$. In quantum field theory, the Fourier modes of free quantum fields behave like quantum harmonic oscillators with a frequency $\omega(k)$ that depends on the Fourier wavenumber k. The ground-state configuration of the field (i.e., the vacuum) has a non-zero energy due to the zero-point energies of each of these harmonic oscillators. For example, for the electromagnetic field, we expect a ground-state energy

$$\rho_{\rm vac,EM}\, c^2 = \frac{2}{(2\pi)^3} \int \frac{\hbar c k}{2}\, {\rm d}^3\vec{k}\,, \tag{9.59}$$

where we have used $\omega = ck$. The integral is clearly divergent, but we cannot expect quantum field theory to be valid on all scales, particularly as $1/k$ approaches the Planck length.[2] If we cut off the integral in Eq. (9.59) at the Planck scale, and include all known fields, we predict a vacuum energy density some 120 orders of magnitude larger than current cosmological limits! We do not (yet) understand why the energy density of the vacuum, as inferred from the observed cosmological constant, is so much smaller than the naive theoretical prediction.

[2]The Planck scale is where we expect quantum gravity effects to become important; it is given by

$$l_{Pl} = \sqrt{\frac{\hbar G}{c^3}} = 1.62 \times 10^{-35}{\rm m}. \tag{9.60}$$

Chapter 10

The Schwarzschild Solution

The Einstein field equations are non-linear partial differential equations and are therefore hard to solve. However, some exact solutions are known in situations where the spacetime possesses symmetries. In this chapter, we shall look at the first exact solution of Einstein's equations that was found, representing the spacetime in the vacuum region outside a spherically-symmetric mass distribution. This solution was found by Karl Schwarzschild in 1915, the same year that Einstein published his General Theory of Relativity in its final form, while serving in the German army on the Russian front during World War I.

10.1 Spherically-Symmetric Spacetimes

In other areas of physics, we are used to thinking of symmetries of field configurations in an *active* sense: the fields and their sources can be actively transformed (e.g., rotated about some origin) and the resulting configuration is identical to the original if the transformation is a symmetry. However, we can also adopt a *passive* viewpoint: we can change our coordinate system without changing the *functional form* of the fields on our coordinates. In general relativity, we shall adopt such a passive viewpoint.

A spacetime possesses a symmetry if under some coordinate transformation $x^\mu \to x'^\mu$, the new components of the metric expressed as functions of the new coordinates, $g'_{\mu\nu}(x')$, have the same functional dependence as the original metric components on the original coordinates, $g_{\mu\nu}(x)$. This means that the line element in the new coordinates has exactly the same dependence on x'^μ and $\mathrm{d}x'^\mu$ as the original line element does on x^μ and $\mathrm{d}x^\mu$.

For the specific case of spherical symmetry, Cartesian-like coordinates $x^i (i = 1, 2, 3)$ must exist such that under the *constant* coordinate transformation

$$x'^i = \sum_{j=1}^{3} (O^{-1})^i{}_j x^j, \tag{10.1}$$

where $O^i{}_j$ is an orthogonal matrix, the functional form of the line element is unchanged. Writing $\vec{x} = (x^1, x^2, x^3)$, and the original line element as[1]

$$\mathrm{d}s^2 = g_{00}(t, \vec{x})\,\mathrm{d}t^2 + 2g_{0i}(t, \vec{x})\,\mathrm{d}t\,\mathrm{d}x^i + g_{ij}(t, \vec{x})\,\mathrm{d}x^i\,\mathrm{d}x^j\,, \tag{10.2}$$

the line element in the rotated coordinates is

$$\mathrm{d}s^2 = g_{00}(t, \mathbf{O}\vec{x}')\,\mathrm{d}t^2 + 2g_{0i}(t, \mathbf{O}\vec{x}')\,\mathrm{d}t\,O^i{}_j\,\mathrm{d}x'^j + g_{ij}(t, \mathbf{O}\vec{x}')O^i{}_k O^j{}_l\,\mathrm{d}x'^k\,\mathrm{d}x'^l\,. \tag{10.3}$$

[1] Here, we extend the summation convention to include implicit summation over repeated "spatial" indices.

If the spacetime is spherically-symmetric, this must have the same functional form as the original line element so that (dropping the primes)

$$g_{00}(t,\vec{x}) = g_{00}(t,\mathbf{O}\vec{x}), \tag{10.4}$$

$$g_{0j}(t,\vec{x})\,\mathrm{d}t\,\mathrm{d}x^j = g_{0i}(t,\mathbf{O}\vec{x})\,\mathrm{d}t\,O^i{}_j\,\mathrm{d}x^j\,, \tag{10.5}$$

$$g_{kl}(t,\vec{x})\,\mathrm{d}x^k\,\mathrm{d}x^l = g_{ij}(t,\mathbf{O}\vec{x})O^i{}_kO^j{}_l\,\mathrm{d}x^k\,\mathrm{d}x^l\,. \tag{10.6}$$

To satisfy these constraints, the following must be true:

- $g_{00}(t,\vec{x})$ can only depend on $\vec{x}$ through the rotational invariant $r=\sqrt{\vec{x}\cdot\vec{x}}$ so we can write
$$g_{00}(t,\vec{x}) = A(t,r); \tag{10.7}$$
- $g_{0i}(t,\vec{x})\,\mathrm{d}t\,\mathrm{d}x^i$ can only involve the invariant $\vec{x}\cdot\mathrm{d}\vec{x}$ multiplying a function of t and r, so that
$$g_{0i}(t,\vec{x})\,\mathrm{d}t\,\mathrm{d}x^i = -B(t,r)\vec{x}\cdot\mathrm{d}\vec{x}\,; \tag{10.8}$$
- $g_{ij}(t,\vec{x})\,\mathrm{d}x^i\,\mathrm{d}x^j$ must be of the form
$$g_{ij}(t,\vec{x})\,\mathrm{d}x^i\,\mathrm{d}x^j = -C(t,r)(\vec{x}\cdot\mathrm{d}\vec{x})^2 - D(t,r)\,\mathrm{d}\vec{x}\cdot\mathrm{d}\vec{x}\,. \tag{10.9}$$

We shall later discuss the following additional symmetries:

- *Stationary*: a spacetime is stationary if it has constant time shifts $t\to t+\text{const.}$ as a symmetry (for all constant shifts), which means that $g_{\mu\nu}$ cannot depend on t.
- Static: a spacetime is static if it additionally has time reversal, $t\to -t$, as a symmetry; this requires $g_{0i}=0$.

To understand the distinction between a static spacetime and a stationary one, consider the analogous situation in electromagnetism. There, a steady charge and current distribution is stationary (it looks the same for all time) and the fields and sources are independent of t. However, under time reversal the current changes sign at a given point in space and this reverses the sign of the magnetic field everywhere. For the configuration to be static, i.e., invariant under time reversal, would require that the current be zero (so the charges are at rest).

Returning to the most general spherically-symmetric spacetime, it is convenient to switch from the Cartesian-like spatial coordinates x^i to spherical-polar coordinates with

$$x^1 = r\sin\theta\cos\phi,\quad x^2 = r\sin\theta\sin\phi,\quad x^3 = r\cos\theta. \tag{10.10}$$

It follows that

$$\vec{x}\cdot\mathrm{d}\vec{x} = r\,\mathrm{d}r\,, \tag{10.11}$$

$$\mathrm{d}\vec{x}\cdot\mathrm{d}\vec{x} = \mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2\,, \tag{10.12}$$

where $\mathrm{d}\Omega^2$ is the spherical line element on the unit 2-sphere:

$$\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\,. \tag{10.13}$$

Using Eq.s (10.7-10.9), after regrouping terms and absorbing factors of r into redefined functions A, B, C and D, the spherically-symmetric line element takes the form

$$\mathrm{d}s^2 = A(t,r)\,\mathrm{d}t^2 - 2B(t,r)\,\mathrm{d}t\,\mathrm{d}r - C(t,r)\,\mathrm{d}r^2 - D(t,r)\,\mathrm{d}\Omega^2\,. \tag{10.14}$$

We can simplify this further by applying various coordinate transformations. First, we switch to a new radial coordinate $\bar{r}$ defined by

$$\bar{r}^2 = D(t,r), \tag{10.15}$$

and so use coordinates $(t,\bar{r},\theta,\phi)$. We can, in principle, express r as a function of $\bar{r}$ and t so that, for example, $A(t,r)$ becomes some (different) function of $\bar{r}$ and t. Also,

$$\bar{r}^2 = D(t,r) \quad \Longrightarrow \quad 2\bar{r}\,\mathrm{d}\bar{r} = \frac{\partial D}{\partial t}\,\mathrm{d}t + \frac{\mathrm{d}D}{\mathrm{d}r}\,\mathrm{d}r\,, \tag{10.16}$$

so that

$$\mathrm{d}r = \frac{1}{\partial D/\partial r}\left(2\bar{r}\,\mathrm{d}\bar{r} - \frac{\partial D}{\partial t}\,\mathrm{d}t\right). \tag{10.17}$$

Substituting into the line element (10.14), and collecting terms and redefining functions A, B and C, we find

$$\mathrm{d}s^2 = A(t,\bar{r})\,\mathrm{d}t^2 - 2B(t,\bar{r})\,\mathrm{d}t\,\mathrm{d}\bar{r} - C(t,\bar{r})\,\mathrm{d}\bar{r}^2 - \bar{r}^2\,\mathrm{d}\Omega^2\,. \tag{10.18}$$

Note that $\bar{r}$ has a clear physical definition: it is an *area coordinate* in the sense that in the $t=$ const. hypersurface, the 2D surface defined by $\bar{r}=$ const. is a 2-sphere of area $4\pi\bar{r}^2$.

We can make a further coordinate transformation to remove the cross-term involving $\mathrm{d}t\,\mathrm{d}\bar{r}$ in the line element as follows. Noting that

$$A(t,\bar{r})\,\mathrm{d}t^2 - 2B(t,\bar{r})\,\mathrm{d}t\,\mathrm{d}\bar{r} = \frac{1}{A(t,\bar{r})}[A(t,\bar{r})\,\mathrm{d}t - B(t,\bar{r})\,\mathrm{d}\bar{r}]^2 - \frac{B^2(t,\bar{r})}{A(t,\bar{r})}\,\mathrm{d}\bar{r}^2\,, \tag{10.19}$$

we introduce a new time coordinate $\bar{t}$, which is a function of t and $\bar{r}$, such that

$$\mathrm{d}\bar{t} = \Phi(t,\bar{r})[A(t,\bar{r})\,\mathrm{d}t - B(t,\bar{r})\,\mathrm{d}\bar{r}]. \tag{10.20}$$

The function $\Phi(t,\bar{r})$ is an integrating factor to ensure that the combination on the right-hand side is an exact differential. In terms of $(\bar{t},\bar{r},\theta,\phi)$, the line element becomes

$$\mathrm{d}s^2 = \frac{1}{A\Phi^2}\,\mathrm{d}\bar{t}^2 - \left(\frac{B^2}{A} + C\right)\mathrm{d}\bar{r}^2 - \bar{r}^2\,\mathrm{d}\Omega^2\,, \tag{10.21}$$

where all functions depend implicitly on $\bar{t}$ and $\bar{r}$ through their arguments t and $\bar{r}$. Re-defining these functions, and dropping the bars, we obtain the diagonal form of the line element with spherical symmetry (the *isotropic* line element):

$$\mathrm{d}s^2 = A(t,r)\,\mathrm{d}t^2 - B(t,r)\,\mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2\,. \tag{10.22}$$

Finally, we shall look for solutions that are static in these coordinates, so that the two functions A and B do not depend on time. In this case, the static, isotropic line element takes the form

$$\boxed{\mathrm{d}s^2 = A(r)\,\mathrm{d}t^2 - B(r)\,\mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2\,.} \tag{10.23}$$

10.2 Solution of the Field Equations in Vacuum

Any static spherically-symmetric spacetime has a line element that can be written in the form (10.23). The functions $A(r)$ and $B(r)$ are determined by solving the Einstein field equations given some (spherical and static) matter distribution. We shall consider the important case of the vacuum region outside of such a spherical mass distribution. In this case, the energy–momentum tensor vanishes and (ignoring the cosmological constant term) the Einstein field equations reduce to

$$R_{\mu\nu} = 0, \tag{10.24}$$

where the Ricci tensor can be written in terms of the metric connection as

$$R_{\mu\nu} = -\partial_\rho \Gamma^\rho_{\mu\nu} + \partial_\mu \Gamma^\rho_{\rho\nu} + \Gamma^\rho_{\sigma\nu}\Gamma^\sigma_{\mu\rho} - \Gamma^\rho_{\mu\nu}\Gamma^\sigma_{\sigma\rho}. \tag{10.25}$$

We can easily calculate the connection from the metric and its inverse; instead of using numerical coordinate labels $0, 1, 2, 3$, let us use the more transparent labels t, r, θ, ϕ, in which case

$$\begin{aligned} g_{tt} &= A(r), & g^{tt} &= 1/A(r), \\ g_{rr} &= -B(r), & g^{rr} &= -1/B(r), \\ g_{\theta\theta} &= -r^2, & g^{\theta\theta} &= -1/r^2, \\ g_{\phi\phi} &= -r^2 \sin^2\theta, & g^{\phi\phi} &= -1/(r^2 \sin^2\theta). \end{aligned} \tag{10.26}$$

The non-zero, independent connection coefficients are

$$\begin{aligned} \Gamma^t_{tr} &= A'/(2A), & \Gamma^r_{tt} &= A'/(2B), \\ \Gamma^r_{rr} &= B'/(2B), & \Gamma^r_{\theta\theta} &= -r/B, \\ \Gamma^r_{\phi\phi} &= -r\sin^2\theta/B, & \Gamma^\theta_{r\theta} &= 1/r, \\ \Gamma^\theta_{\phi\phi} &= -\sin\theta\cos\theta, & \Gamma^\phi_{r\phi} &= 1/r, \\ \Gamma^\phi_{\theta_\phi} &= \cot\theta, \end{aligned} \tag{10.27}$$

where primes denote derivatives with respect to r. It is now a matter of tedious, but routine, algebra to calculate the components of the Ricci tensor. A useful intermediate result is that

$$\Gamma^\rho_{\rho\sigma} = \left(\frac{A'}{2A} + \frac{B'}{2B} + \frac{2}{r}\right)\delta^r_\sigma + \cot\theta\,\delta^\theta_\sigma. \tag{10.28}$$

The off-diagonal components of the Ricci tensor vanish, while the diagonal components are

$$R_{tt} = -\frac{A''}{2B} + \frac{A'}{4B}\left(\frac{A'}{A} + \frac{B'}{B}\right) - \frac{A'}{rB}, \tag{10.29}$$

$$R_{rr} = \frac{A''}{2A} - \frac{A'}{4A}\left(\frac{A'}{A} + \frac{B'}{B}\right) - \frac{B'}{2B}. \tag{10.30}$$

$$R_{\theta\theta} = \frac{1}{B} - 1 + \frac{r}{2B}\left(\frac{A'}{A} - \frac{B'}{B}\right), \tag{10.31}$$

$$R_{\phi\phi} = \sin^2\theta R_{\theta\theta} \tag{10.32}$$

In vacuum, all these components must vanish.

Forming $AR_{rr}/B + R_{tt} = 0$ gives

$$\frac{A'}{A} + \frac{B'}{B} = 0 \quad \Longrightarrow \quad AB = \alpha, \tag{10.33}$$

where α is an integration constant. Substituting in $r_{\theta\theta} = 0$ then gives

$$rA' + a = \alpha, \tag{10.34}$$

so that

$$rA = \alpha(r + k), \tag{10.35}$$

where k is a further integration constant. It follows that

$$A(r) = \alpha\left(1 + \frac{k}{r}\right), \qquad B(r) = \left(1 + \frac{k}{r}\right)^{-1}. \tag{10.36}$$

Note that if we substitute $B = \alpha/A$ into $R_{tt} = 0$ or $R_{rr} = 0$, we have

$$rA'' + 2A' = 0 \quad \Longrightarrow \quad (r^2A')' = 0, \tag{10.37}$$

so that $r^2A' = \text{const.}$; this is automatically satisfied by the solution in Eq. (10.36).

We can determine the constants α and k by considering the metric at large r, where the field becomes weak. Recall that in the weak-field limit, consistency with Newtonian theory demands that the line element take the for

$$\mathrm{d}s^2 \approx \left(1 + \frac{2\Phi}{c^2}\right)\mathrm{d}(ct)^2 + \cdots, \tag{10.38}$$

where Φ is the Newtonian potential. Outside a spherical mass of total mass M, we have $\Phi = -GM/r$ and so

$$A(r)\,\mathrm{d}t^2 \to c^2\left(1 - \frac{2GM}{c^2 r}\right)\mathrm{d}t^2, \tag{10.39}$$

which requires $\alpha = c^2$ and $k = -2GM/c^2$.

We this obtain the *Schwarzschild solution* for the vacuum region outside a static, spherically-symmetric body of mass M:

$$\boxed{\mathrm{d}s^2 = c^2\left(1 - \frac{2\mu}{r}\right)\mathrm{d}t^2 - \left(1 - \frac{2\mu}{r}\right)^{-1}\mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2,} \tag{10.40}$$

where $\mu \equiv GM/c^2$. Note the following properties of this solution.

- The solution is valid only down to the surface of the spherical body; in the interior region, $r_{\mu\nu} \neq 0$ and the solution is not valid there.
- The metric is singular at $r = 2\mu$. This turns out to be only a coordinate singularity, although the hypersurface $r = 2\mu$ does have interesting properties (if the radius of the body is smaller than 2μ, in which case we have a blackhole; see Chapter 12). For the moment we restrict attention to $r > 2\mu$.
- As $r \to \infty$, the metric tends to the Minkowski metric and the spacetime is said to be *asymptotically flat*.

10.2.1 Birkhoff's Theorem

Suppose we dropped the requirement that the metric be static; then

$$\mathrm{d}s^2 = A(t,r)\,\mathrm{d}t^2 - B(t,r)\,\mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2\,. \tag{10.41}$$

If we repeated the calculation above, we would find additional terms in the connection and the components of the Ricci tensor, but solving $r_{\mu\nu} = 0$ would still lead to the same Schwarzschild solution. This leads to *Birkhoff's theorem*:

> Any spherically-symmetric solution of the Einstein field equations in vacuum is given by the Schwarzschild solution; it is static and asymptotically flat.

Birkhoff's theorem implies that a spherical star undergoing radial pulsations has a static external metric and so, for example, cannot emit gravitational waves. Birkhoff's theorem is similar to Gauss's theorem in Newtonian theory; in particular, any *redistribution* of the mass in some spherical system that preserves the total mass has a static external gravitational field.

10.3 Geodesics in Schwarzschild Spacetime

We now consider the motion of free-falling particles in the Schwarzschild solution. It is simplest to follow the alternative "Lagrangian" approach of Subsection 5.4.4, so we consider

$$L = c^2\left(1 - \frac{2\mu}{r}\right)\dot{t}^2 - \left(1 - \frac{2\mu}{r}\right)^{-1}\dot{r}^2 - r^2\dot{\theta}^2 - r^2\sin^2\theta\dot{\phi}^2, \tag{10.42}$$

where overdots denote derivatives with respect to the affine parameter (which we shall denote here by λ). The Euler–Lagrange equations are

$$\frac{\partial L}{\partial x^\mu} = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\partial L}{\partial \dot{x}^\mu}\right), \tag{10.43}$$

where $x^0 = t, x^1 = r, x^2 = \theta$ and $x^3 = \phi$.

Consider first the equation for θ:

$$\begin{aligned} -2r^2\sin\theta\cos\theta\dot{\phi}^2 &= -2\frac{\mathrm{d}(r^2\dot{\theta})}{\mathrm{d}\lambda} \\ \implies \ddot{\theta} + \frac{2}{r}\dot{r}\dot{\theta} - \sin\theta\cos\theta\dot{\phi}^2 &= 0. \end{aligned} \tag{10.44}$$

A possible solution of this is $\theta = \pi/2$, i.e., planar motion in the equatorial plane; given the spherical symmetry we can always consider motion in this plane without loss of generality. The Euler–Lagrange equation for t reduces to $\partial L/\partial\dot{t} = $ const. since L has no explicit dependence on t, a consequence of the solution being stationary. We can therefore write

$$\left(1 - \frac{2\mu}{r}\right)\dot{t} = k, \tag{10.45}$$

where k is a constant. Note that this conservation law is equivalent to $t_0 = \text{const.}$, where $t^\mu = \dot{x}^\mu$ is the tangent vector to the wordline.

The Lagrangian has no explicit dependence on ϕ and so $\partial L \big/ \partial \dot{\phi} = \text{const}$. In the plane $\theta = \pi/2$, this reduces to

$$r^2 \dot{\phi} = h \tag{10.46}$$

for constant h, which is equivalent to $t_3 = \text{const}$.

Finally, for r the Euler–Lagrange equation gives

$$\left(1 - \frac{2\mu}{r}\right)^{-1} \ddot{r} + \frac{\mu c^2}{r^2} \dot{r}^2 - \left(1 - \frac{2\mu}{r}\right)^{-2} \frac{\mu}{r^2} \dot{r}^2 - r\dot{\phi}^2 = 0. \tag{10.47}$$

However, it is often more convenient to use a further first integral of the motion, which follows directly from $L = c^2$ for a massive particle, and $L = 0$ for a massless one:

$$\left(1 - \frac{2\mu}{r}\right) c^2 \dot{t}^2 - \left(1 - \frac{2\mu}{r}\right)^{-1} \dot{r}^2 - r^2 \dot{\phi}^2 = \begin{cases} c^2 & \text{massive,} \\ 0 & \text{massless.} \end{cases} \tag{10.48}$$

10.3.1 Interpretation of the integration constants k and h

The constant k in $(1 - 2\mu/r)\dot{r} = k$ is related to the energy of the particle as measured by a stationary observer. Consider an observer at rest in the (r, θ, ϕ) coordinates. Their 4-velocity is of the form $u^\mu = \mathcal{A}\delta^\mu_0$, with $\mathcal{A}$ determined from

$$c^2 = g_{\mu\nu} u^\mu u^\nu \quad \Longrightarrow \quad \mathcal{A} = \left(1 - \frac{2\mu}{r}\right)^{-1/2}. \tag{10.49}$$

The energy of the particle with 4-momentum $\mathbf{p}$ as measured by this observer is[2]

$$E = \mathbf{g}(\mathbf{u}, \mathbf{p}) = g_{00} \mathcal{A} p^0, \tag{10.50}$$

which evaluates to

$$E = mc^2 \dot{t} \left(1 - \frac{2\mu}{r}\right)^{1/2} = kmc^2 \left(1 - \frac{2\mu}{r}\right)^{-1/2} \tag{10.51}$$

for the massive case. We see that kmc^2 is the energy of the particle as measured by a stationary observer as $r \to \infty$. For the massive particle to reach spatial infinity we require $k \geq 1$ (since the measured energy cannot be less than mc^2).

In the case of a massless particle, $p^0 = \dot{t}$ and we have

$$E = c^2 \dot{t} \left(1 - \frac{2\mu}{r}\right)^{1/2} = kc^2 \left(1 - \frac{2\mu}{r}\right)^{-1/2}. \tag{10.52}$$

We require $k \geq 0$ for the massless particle to reach infinity. The constant h in $r^2\dot{\phi} = h$ arises from the symmetry of the spacetime under rotations about the z-axis, and can be interpreted as the *specific angular momentum*.

[2] In the rest-frame of the observer, in local inertial coordinates, the observer's 4-velocity has components $u'^\mu = (c, 0, 0, 0)$ and $p'^0 = E/c$. It follows that $E = \eta_{\mu\nu} u'^\mu p'^\nu = g_{\mu\nu} u^\mu p^\nu$.

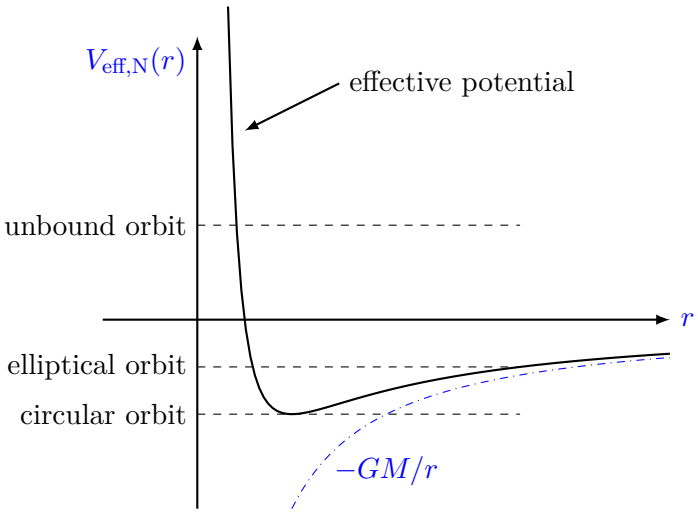

Fig. 10.1: Effective potential in Newtonian theory (for non-zero angular orbital momentum). Note how there is an angular momentum barrier that prevents particles from reaching $r = 0$.

10.3.2 The Energy Equation and Effective Potential

In Newtonian theory, we can understand a lot about orbital motion by considering the effective potential. Let us begin by recalling how this works in Newtonian theory, before extending to general relativity. In Newtonian theory, conservation of energy (kinetic plus gravitational potential) for motion in the equatorial plane gives

$$\frac{m}{2}\left(\dot{r}^2 + r^2\dot{\phi}^2\right) - \frac{GMm}{r} = E_{\text{N}}, \tag{10.53}$$

where E_{N} is the Newtonian energy. Here r is radial distance from the mass M and the overdots are with respect to time t. Angular momentum conservation takes the form $r^2\dot{\phi} = h$, and using this in Eq. (10.53) gives

$$\frac{1}{2}\dot{r}^2 - \frac{GM}{r} + \frac{h^2}{2r^2} = \frac{E_{\text{N}}}{m}. \tag{10.54}$$

We write this as

$$\frac{1}{2}\dot{r}^2 + V_{\text{eff,N}}(r) = \frac{E_{\text{N}}}{m}, \tag{10.55}$$

where the Newtonian effective potential is

$$V_{\text{eff,N}}(r) = -\frac{GM}{r} + \frac{h^2}{2r^2}. \tag{10.56}$$

This is plotted in Fig. 10.1 for non-zero h.

The main features of the Newtonian effective potential are as follows.

- There is a "centrifugal barrier" at small r, which prevents the particle reaching $r = 0$ (for non-zero angular momentum h) for any value of the energy E_{N}.
- Bound orbits have $E_{\text{N}} < 0$ so the particle cannot reach spatial infinity. In this case, there are two turning points in r during the orbit (corresponding to the extrema of an elliptical orbit) given by the solutions of $V_{\text{eff,N}}(r) = E_{\text{N}}$.

- The effective potential has a single turning point at $r = h^2/(GM)$. It is a minimum and corresponds to a *stable circular orbit* at this radius.

10.3.2.1 Massive Particles in General Relativity

In general relativity, an analogous energy equation can be derived by eliminating $\dot{t}$ and $\dot{\phi}$ (with their conservation equations (10.45) and (10.46)) from

$$\left(1 - \frac{2\mu}{r}\right)c^2\dot{t}^2 - \left(1 - \frac{2\mu}{r}\right)^{-1}\dot{r}^2 - r^2\dot{\phi}^2 = c^2. \tag{10.57}$$

which, recall, follows from $L = c^2$. This results in

$$\frac{1}{2}\dot{r}^2 - \frac{GM}{r} + \frac{h^2}{2r^2}\left(1 - \frac{2\mu}{r}\right) = \frac{1}{2}c^2(k^2 - 1), \tag{10.58}$$

which we write as

$$\frac{1}{2}\dot{r}^2 + V_{\text{eff}}(r) = \frac{1}{2}c^2(k^2 - 1), \tag{10.59}$$

where the relativistic effective potential is

$$V_{\text{eff}}(r) = -\frac{GM}{r} + \frac{h^2}{2r^2}\left(1 - \frac{2\mu}{r}\right). \tag{10.60}$$

Eq. (10.59) has the same structure as in Newtonian theory, but it must be remembered that the time derivative is with respect to the proper time of the particle, and the coordinate r does not simply measure radial distance from the origin (it is an area coordinate). The effective potential (10.60) is also very similar to the Newtonian case, but the centrifugal term is modified by the factor $(1 - 2\mu/r)$.

This modification has a significant effect on the effective potential at small r for non-zero h, reversing the sign of the centrifugal barrier as shown in Fig. 10.2. The figure shows that the form of the effective potential depends strongly on the angular momentum h; to understand these dependencies, consider the stationary points of $V_{\text{eff}}(r)$. These can be found from

$$\frac{\mathrm{d}V_{\text{eff}}}{\mathrm{d}r} = \frac{\mu c^2}{r^2} + \frac{h^2}{r^3}\left(\frac{3\mu}{r} - 1\right), \tag{10.61}$$

where we have used $\mu \equiv GM/c^2$. Solving for the locations of the extrema gives

$$r_\pm = \frac{h}{2\mu c^2}\left(h \pm \sqrt{h^2 - 12\mu^2c^2}\right), \tag{10.62}$$

so that for $h > \sqrt{12}\mu c$ there are two stationary points, but none for smaller values of h. As $h/(\mu c) \to \infty$, the locations of the stationary points tend to

$$r_+ \to \infty \quad \text{and} \quad r_- \to 4\mu \,(\text{from above}), \tag{10.63}$$

while as $h \to \sqrt{12}\mu c$ they merge at $r_\pm = 6\mu$. The nature of the stationary points follows from

$$\frac{\mathrm{d}^2V_{\text{eff}}}{\mathrm{d}r^2} = -\frac{2\mu c^2}{r^3} + \frac{3h^2}{r^4}\left(1 - \frac{4\mu}{r}\right). \tag{10.64}$$

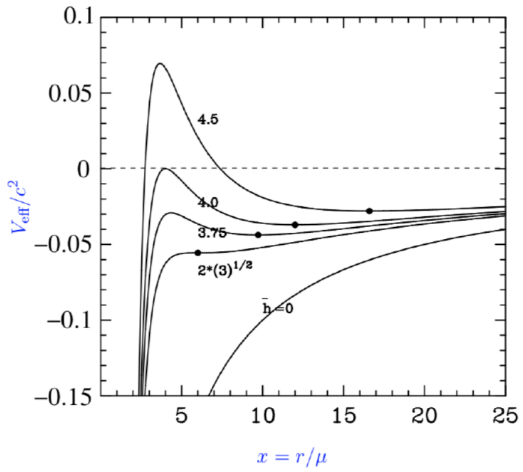

Fig. 10.2: Effective potential in general relativity for several values of the dimensionless angular momentum $\bar{h} = h/(c\mu)$. The dots show the locations of stable circular orbits.

Evaluating at $r_\pm$ gives

$$\left.\frac{\mathrm{d}^2 V_{\rm eff}}{\mathrm{d}r^2}\right|_{r_\pm} = \frac{h^2}{r_\pm^5}(r_\pm - 6\mu), \tag{10.65}$$

so that r_- corresponds to a local maximum and r_+ to a minimum. The effective potential at the stationary points evaluates to

$$V_{\rm eff}(r_\pm) = \frac{h^2}{2r_\pm^3}(4\mu - r_\pm), \tag{10.66}$$

which is always negative at the minimum (r_+), but is positive at the maximum for $r_- < 4\mu$ (which occurs for $h > 4\mu c$). For $h = \sqrt{12}\mu c$ we have only an inflection point at $r = 6\mu$ and the effective potential there takes the value $V_{\rm eff} = -c^2/18$. The implications of the shape of $V_{\rm eff}(r)$ for orbital motion are as follows:

- The reversal in sign of the centrifugal barrier at small r means that a particle with sufficient energy (i.e., large enough k) can always reach $r = 0$ for any h, unlike the case in Newtonian gravity. Such an in-falling particle spirals into $r = 0$.
- For $h > \sqrt{12}\mu c$, circular orbits are possible at two radii (the locations of the stationary points, $r_\pm$), with the innermost orbit being unstable and the outermost being stable.
  - The radius of the unstable circular orbit is always in the range $3\mu < r_- \leq 6\mu$, while the radius of the stable circular orbit has $r_+ > 6\mu$.
  - The *innermost stable circular orbit* is at $r = 6\mu$ and the particle has angular moment $h = \sqrt{12}\mu c$ in this orbit.
  - The circular orbits with $r > 4\mu$ are bound ($k < 1$) – the effective potential is negative there, and if the particle's motion is perturbed at constant h it cannot reach infinity.

10.3.2.2 Accretion Power

Gas in an *accretion disc* around a compact object moves in quasi-circular orbits. Due to viscosity, a packet of gas in the disc loses angular momentum causing it to move slowly inwards until it can no longer follow a stable circular orbit, at which point it falls into the compact object. Vast amounts of energy can be radiated as gas moves though the disc, which we can estimate as follows.

First consider the Newtonian calculation: the total energy (kinetic plus gravitational potential energy) of a particle of mass m in a circular orbit at radius r is

$$E_{\rm N} = -\frac{GMm}{2r} = -\frac{\mu}{2r}mc^2. \tag{10.67}$$

The difference in $E_{\rm N}$ between an orbit at $r = \infty$ and one at r can be radiated as the particle moves through the accretion disc in a series of quasi-static orbits. It follows that

$$\frac{\Delta E_{\rm N}}{mc^2} = \frac{\mu}{2r}. \tag{10.68}$$

This can be a significant fraction of the rest-mass energy mc^2 if the particle descends to an orbit with r comparable to μ.

However, we cannot use Newtonian arguments for such compact orbits, so let us perform the analogous calculation relativistically. We imagine a particle in a circular orbit at large r with parameter k, being disturbed in such a way that it moves to a radius r preserving the constant k (this would be like the Newtonian particle conserving its total energy). At r, it will be moving too fast to enter a circular orbit there, and this excess energy can be shed as radiation. We can figure out the parameter k required for a circular orbit at r as follows. The angular momentum for a circular orbit at radius r follows from $\mathrm{d}^2V_{\rm eff}/\mathrm{d}r^2 = 0$, which gives

$$\mu c^2 r^2 = h^2(r - 3\mu). \tag{10.69}$$

The value of k in the orbit follows from

$$\begin{aligned}\frac{1}{2}c^2(k^2-1) &= V_{\rm eff}(r)\\ &= -\frac{\mu c^2}{r} + \frac{h^2}{2r^2}\left(1 - \frac{2\mu}{r}\right)\\ &= -\frac{h^2}{2r^2}\left(1 - \frac{4\mu}{r}\right),\end{aligned} \tag{10.70}$$

where we have used Eq. (10.69) to substitute for μc^2 in passing to the final line. It follows that

$$k^2 - 1 = -\frac{\mu}{r}\frac{(1-4\mu/r)}{(1-3\mu/r)}, \tag{10.71}$$

which, on solving for k, gives

$$k = \frac{(1-2\mu/r)}{(1-3\mu/r)^{1/2}}. \tag{10.72}$$

The particle starts in a circular orbit at $r \gg \mu$, so that $k \approx 1$. When it reaches r, the energy of the particle as measured locally by a stationary observer is, from Eq. (10.51)

with $k = 1$,

$$E = mc^2\left(1 - \frac{2\mu}{r}\right)^{-1/2}. \tag{10.73}$$

In contrast, the energy for a particle in a circular orbit there is

$$E = mc^2\left(\frac{1 - 2\mu/r}{1 - 3\mu/r}\right)^{1/2}, \tag{10.74}$$

where we have used k from Eq. (10.72). The difference between these two energies must be lost if the particle is to enter a circular orbit at r, which gives

$$\frac{\Delta E}{mc^2} = \left(1 - \frac{2\mu}{r}\right)^{-1/2} - \left(\frac{1 - 2\mu/r}{1 - 3\mu/r}\right)^{1/2}. \tag{10.75}$$

In the limit $\mu \ll r$, we have

$$\frac{\Delta E}{mc^2} \approx \frac{\mu}{2r}, \tag{10.76}$$

which recovers the Newtonian result (10.68). Evaluating ΔE for $r = 6\mu$, corresponding to the innermost stable circular orbit, we find

$$\frac{\Delta E}{mc^2} = \frac{\sqrt{3}}{6}(3\sqrt{2} - 4) = 0.07. \tag{10.77}$$

It follows that around 7% of the rest-mass energy must be radiated as a packet of gas moves through the accretion disc to the inner-most edge at $r = 6\mu$. This is around 10 times as efficient as fusing hydrogen into helium (which releases 0.7% of the initial rest mass). Such accretion powers some of the most extreme phenomena known in the Universe.

10.3.2.3 Massless Particles in General Relativity

For massless particles in general relativity, our starting point is

$$\left(1 - \frac{2\mu}{r}\right)c^2\dot{t}^2 - \left(1 - \frac{2\mu}{r}\right)^{-1}\dot{r}^2 - r^2\dot{\phi}^2 = 0, \tag{10.78}$$

which follows from $L = 0$. Eliminating $\dot{t}$ and h, as in the massive case, gives

$$\frac{1}{2}\dot{r}^2 + \frac{h^2}{2r^2}\left(1 - \frac{2\mu}{r}\right) = \frac{1}{2}c^2k^2, \tag{10.79}$$

which we write as

$$\frac{1}{2}\dot{r}^2 + V_{\text{eff}}(r) = \frac{1}{2}c^2k^2, \tag{10.80}$$

where the relativistic effective potential is

$$V_{\text{eff}}(r) = \frac{h^2}{2r^2}\left(1 - \frac{2\mu}{r}\right). \tag{10.81}$$

The effective potential has the same shape for all nonzero h; it is plotted in Fig. 10.3. We have

$$\frac{\mathrm{d}V_{\text{eff}}}{\mathrm{d}r} = -\frac{h^2}{r^3}\left(1 - \frac{3\mu}{r}\right), \tag{10.82}$$

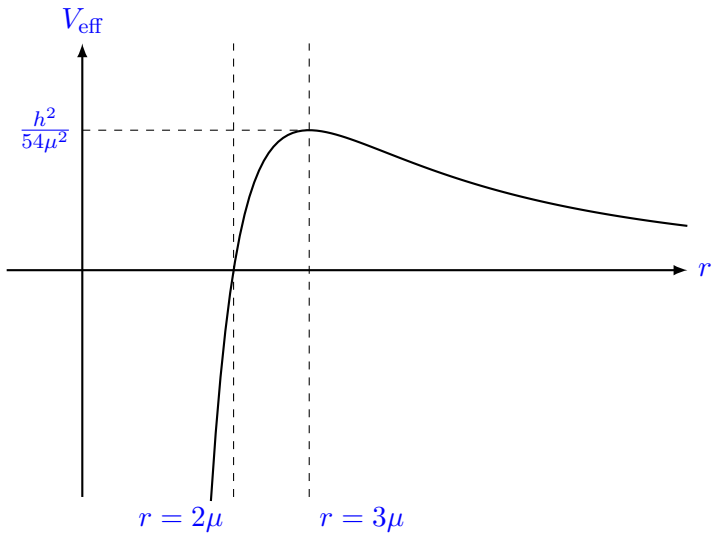

Fig. 10.3: Relativistic effective potential for massless particles.

so there is a single stationary point at $r = 3\mu$, which is a maximum. The value of V_{eff} there is

$$V_{\text{eff}}(3\mu) = \frac{h^2}{54\mu^2}, \tag{10.83}$$

and is the global maximum of the function. We see that massless particles have a single circular orbit at $r = 3\mu$ and it is unstable.

Now consider more general orbits, in particular, a photon moving inwards from large radii with angular momentum h. If $c^2k^2/2 < h^2/(54\mu^2)$, the photon cannot cross the barrier in V_{eff} at $r = 3\mu$; there will be a single turning point of closest approach, where $c^2k^2/2 = V_{\text{eff}}(r)$, and the photon will escape to infinity again. However, if $c^2k^2/2 > h^2/(54\mu^2)$, the photon will be captured by the massive body and will spiral inwards to $r = 0$. Note that the behaviour of the photon orbit depends on the parameters k and h only through the ratio h/k.

The physical interpretation of this parameter is that it is the impact parameter of the orbit (see Fig. 10.4):

$$b = \frac{h}{ck}. \tag{10.84}$$

To see this, we combine the energy equation (10.80) with $r^2\dot{\phi} = h$ to find the shape equation

$$\begin{aligned}\left(\frac{\mathrm{d}r}{\mathrm{d}\phi}\right)^2 &= \frac{\dot{r}^2}{\dot{\phi}^2} = \frac{r^4\dot{r}^2}{h^2} \\ &= r^2\left[\frac{c^2k^2}{h^2}r^2 - \left(1 - \frac{2\mu}{r}\right)\right],\end{aligned} \tag{10.85}$$

so that

$$\frac{\mathrm{d}r}{\mathrm{d}\phi} = \pm r\left[\frac{c^2k^2}{h^2}r^2 - \left(1 - \frac{2\mu}{r}\right)\right]^{1/2}. \tag{10.86}$$

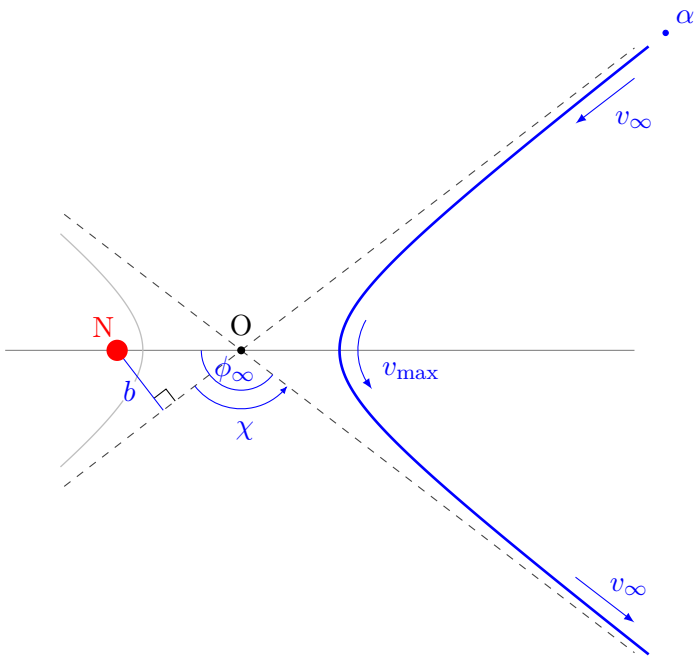

Fig. 10.4: The impact parameter b is the distance by which a body α, if it continued on an unperturbed path, would miss the central body N at its closest approach. With bodies experiencing classical inverse square law forces (e.g. newtonian gravity) and following hyperbolic trajectories it is equal to the semi-minor axis of the hyperbola. The total angle of deflection is $\chi = \pi - 2\phi_\infty$, and is determined by the asymptote angles ϕ_∞ and the velocity at infinity v_∞ with the relation $\tan^2\phi_\infty = mv_\infty^2 b/A$, where the constant A results from the relevant force law $F = Ar^2$, where $A > 0$ for a repulsive force or $A < 0$ for an attractive force.

At large radii, the spacetime metric tends to the Minkowski metric and we expect photon orbits to approach straight lines there. If the impact parameter is b and, without loss of generality, we assume that $\phi \to 0$ as $r \to \infty$, the straight line path is

$$r \sin\phi = b. \tag{10.87}$$

Differentiating gives

$$\frac{\mathrm{d}r}{\mathrm{d}\phi}\sin\phi + r\cos\phi = 0 \quad \Longrightarrow \quad \frac{\mathrm{d}r}{\mathrm{d}\phi} = \pm r\left(\frac{r^2}{b^2} - 1\right)^{1/2}. \tag{10.88}$$

Comparing with Eq. (10.86) for $r \gg 2\mu$, we see that, indeed, $b = h/(ck)$. It follows that if the impact parameter $b < \sqrt{27}\mu$, the photon will be captured by the massive body.

The reason the orbit depends only on the ratio h/k is because the effects of gravitational fields on massless particles are achromatic (i.e., independent of energy). Since we have taken $p^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}$, considering a particle of different energy amounts to a constant scaling of the affine parameter (preserving its affine character). Under such a change, the constants k and h both scale in the same way, preserving their ratio, but if the orbit depended on k and h separately the orbits would be different for massless particles with the same intitial conditions (i.e., spacetime position and spatial direction of propagation) but different frequencies.

10.4 Gravitational Redshift

Finally, we consider the change in the frequency of a photon, as measured by observers at constant (r, θ, ϕ), as it propagates through Schwarzschild spacetime. We have already seen that the energy of the photon relative to a stationary observer at r is

$$E = c^2 \dot{t}\left(1 - \frac{2\mu}{r}\right)^{1/2} = kc^2\left(1 - \frac{2\mu}{r}\right)^{-1/2}, \tag{10.89}$$

and the energy is related to the observed frequency by $E = h\nu$. As k is constant, if a photon is emitted at r_E, with frequency ν_E as measured by a stationary observer there, and is received by a stationary observer at r_R who measures the frequency to be r_E, we have

$$\frac{\nu_R}{\nu_E} = \left(\frac{1 - 2\mu/r_E}{1 - 2\mu/r_R}\right)^{1/2}. \tag{10.90}$$

This is usually expressed in terms of the *redshift* z, which is the ratio of received to emitted wavelengths:

$$1 + z = (\nu_R/\nu_E)^{-1}. \tag{10.91}$$

For observations at infinity, the redshift becomes

$$1 + z_\infty = \left(1 - \frac{2\mu}{r_E}\right)^{-1/2}. \tag{10.92}$$

Note how this tends to infinity as the point of emission approaches $r = 2\mu$, something we shall return to in Chapter 12.

The calculation of the gravitational redshift generalises to any *static* metric. We then have the "Lagrangian"

$$L = g_{00}\dot{t}^2 + \sum_{ij} g_{ij}\dot{x}^i\dot{x}^j, \tag{10.93}$$

where the metric functions are independent of t. The Euler–Lagrange equation for t now gives

$$g_{00}\dot{t} = kc^2, \tag{10.94}$$

with k a constant. Furthermore, the 4-velocity of a stationary observer has

$$u^\mu = \frac{c}{\sqrt{g_{00}}}\delta^\mu_0, \tag{10.95}$$

and so the energy of the photon as measured by such an observer is

$$E = g_{00}\frac{c}{\sqrt{g_{00}}}\dot{t} = \frac{kc^2}{\sqrt{g_{00}/c^2}} \tag{10.96}$$

The formula for the gravitational redshift in this case generalises Eq. (10.90) to

$$\frac{\nu_R}{\nu_E} = \sqrt{\frac{g_{00}(E)}{g_{00}(R)}}. \tag{10.97}$$

Chapter 11

Classical Tests of General Relativity

General Relativity was conceived by pure thought. Over the 100 years since its formulation, General Relativity has been tested experimentally and observationally over a variety of length scales (from Solar System and smaller to the entire observable universe). No deviation from the predictions of the theory has ever been found! Many experimental tests are based on the Schwarzschild geometry for $r > 2\mu$, and involve the trajectories of massive particles or light.

Here, we shall discuss two such classic tests: the perihelion advance of the planet Mercury; and the bending of light by the Sun. These are classic predictions, dating back to Einstein's seminal papers laying down the theory of General Relativity, of effects that are either absent or differ in value in Newtonian gravity. Many more recent tests of General Relativity are targeting the strong-field region, $\Phi \sim c^2$. A very significant recent breakthrough for the field was the detection of gravitational waves by LIGO (see Chapter 1), which is now allowing tests of General Relativity in truly extreme environments.

11.1 Shapes of Orbits for Massive and Massless Particles

Recall Eq. (10.48) from Section 10.3 that for free-falling massive and massless particles in Schwarzschild spacetime

$$\left(1-\frac{2\mu}{r}\right)c^2\dot{t}^2-\left(1-\frac{2\mu}{r}\right)^{-1}\dot{r}^2-r^2\dot{\phi}^2=\begin{cases}c^2 & \text{massive,}\\ 0 & \text{massless.}\end{cases} \tag{11.1}$$

These combine with the conserved quantities from equations (10.45) and (10.46),

$$\left(1-\frac{2\mu}{r}\right)\dot{t}=k, \tag{11.2}$$

$$r^2\dot{\phi}=h, \tag{11.3}$$

to determine the orbits. Using the conserved quantities k and h in Eq. (11.1), we have

$$\frac{1}{2}\dot{r}^2-\frac{GM}{r}+\frac{h^2}{2r^2}\left(1-\frac{2\mu}{r}\right)=\frac{1}{2}c^2(k^2-1), \tag{11.4}$$

for massive particles, as in Eq. (10.58), and

$$\frac{1}{2}\dot{r}^2+\frac{h^2}{2r^2}\left(1-\frac{2\mu}{r}\right)=\frac{1}{2}c^2k^2, \tag{11.5}$$

for massless particles, as in Eq. (10.79).

Here, we shall be interested in the *shape* of the orbits, i.e., r as a function of ϕ, so we use

$$\dot{r}=\dot{\phi}\frac{\mathrm{d}r}{\mathrm{d}\phi}=\frac{h}{r^2}\frac{\mathrm{d}r}{\mathrm{d}\phi}=-h\frac{\mathrm{d}u}{\mathrm{d}\phi}, \tag{11.6}$$

where $u \equiv 1/r$. Substituting in Eq. (11.4), we have

$$\frac{1}{2}\left(\frac{\mathrm{d}u}{\mathrm{d}\phi}\right)^2 - \frac{GM}{h^2}u + \frac{1}{2}(u^2 - 2\mu u^3) = \frac{c^2}{2h^2}(k^2 - 1). \tag{11.7}$$

Differentiating with respect to ϕ gives

$$\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2} + u - 3\mu u^2 = \frac{GM}{h^2} \qquad \text{(massive)}, \tag{11.8}$$

which determines the shape of the orbit. Repeating for the massless case, we have

$$\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2} + u - 3\mu u^2 = 0 \qquad \text{(massless)}. \tag{11.9}$$

11.1.1 Newtonian Orbits of Massive Particles

Let us recall some of the properties of the orbits of massive particles in Newtonian theory. If we repeat the analysis above for Newtonian dynamics, the orbit equation (for massive particles) is

$$\frac{\mathrm{d}^2u}{\mathrm{d}\phi^2} + u = \frac{GM}{h^2} \qquad \text{(Newtonian)}. \tag{11.10}$$

This differs from the relativistic equation (11.8), which has the additional term $-3\mu u^2$ on the left-hand side. We shall see that this has important consequences for the shape of the orbits in General Relativity. The solution of Eq. (11.10) is

$$u = \frac{GM}{h^2}(1 + e\cos\phi), \tag{11.11}$$

where the constant e is the eccentricity of the orbit and we have chosen the orbit to have $\mathrm{d}u/\mathrm{d}\phi = 0$ at $\phi = 0$ without loss of generality. The Newtonian energy equation

$$\begin{aligned} \frac{E_\mathrm{N}}{m} &= \frac{1}{2}\dot{r}^2 + \frac{1}{2}r^2\dot{\phi}^2 - \frac{GM}{r} \\ \Longrightarrow \quad \frac{E_\mathrm{N}}{mh^2} &= \frac{1}{2}\left(\frac{\mathrm{d}u}{\mathrm{d}\phi}\right)^2 + \frac{1}{2}u^2 - \left(\frac{GM}{h^2}\right)u, \end{aligned} \tag{11.12}$$

relates the eccentricity to the energy:

$$\frac{E_\mathrm{N}}{mh^2} = \frac{1}{2}\left(\frac{GM}{h^2}\right)^2\left(e^2 - 1\right). \tag{11.13}$$

Bound orbits have $E_\mathrm{N} < 0$ and so eccentricity $e < 1$. For $e = 0$, we have circular orbits with radius r_0, where

$$r_0 \equiv \frac{h^2}{GM}. \tag{11.14}$$

For $0 < e < 1$, we have elliptical orbits with the mass M at a focus (see Fig. 11.1). The radial distance at the point of closest approach (*perihelion*) is $r_0/(1+e)$ and the greatest distance (*aphelion*) is $r_0/(1-e)$. If the ellipse has semi-major axis length a (so the distance between perihelion and aphelion is $2a$), we have $a = r_0/(1-e^2)$ or

$$a = \frac{h^2}{GM(1-e^2)}. \tag{11.15}$$

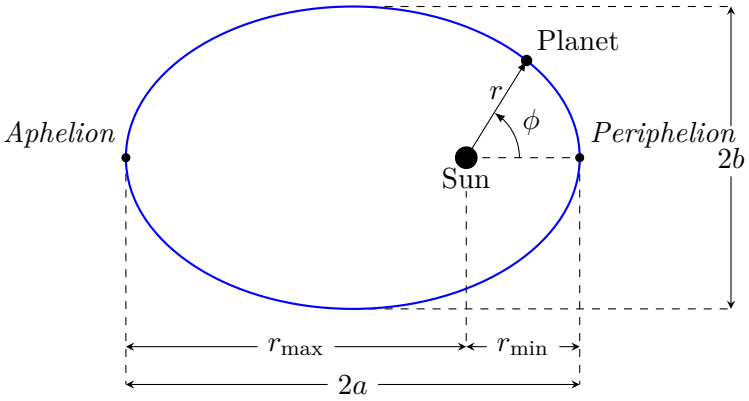

Fig. 11.1: Elliptical orbit of a planet around the Sun in Newtonian gravity. The Sun is at one focus of the ellipse. The semi-major axis length is a and the ellipticity is e. Note that the distances of closest and furthest approach are given by $r_{\min} = a(1+e)$ and $r_{\max} = a(1-e)$, respectively.

11.2 Precession of Planetary Orbits

We shall solve the relativistic equation of motion, which we now write as

$$\frac{\mathrm{d}^2 u}{\mathrm{d}\phi^2} + u = \frac{GM}{h^2} + \frac{3GM}{c^2}u^2, \tag{11.16}$$

in the limit where the relativistic correction is weak ($GM \ll rc^2$). In this case, we can solve the equation by perturbing around a Newtonian orbit. Let us introduce a dimensionless inverse radius, U, so that

$$u = \frac{GM}{h^2}U. \tag{11.17}$$

Writing Eq. (11.16) in terms of U, we have

$$\frac{\mathrm{d}^2 U}{\mathrm{d}\phi^2} + U = 1 + \underbrace{\frac{3(GM)^2}{c^2h^2}}_{\alpha} U^2. \tag{11.18}$$

The dimensionless quantity $\alpha = 3\mu/r_0$ 0 is assumed small (for Mercury, $\alpha = 8\times10^{-8}$, for example). We therefore look for solutions as an expansion in α, i.e.,

$$U = U_0 + \alpha U_1 + \alpha^2 U_2 + \cdots, \tag{11.19}$$

where $U_0 = 1 + e\cos\phi$ is the Newtonian solution.

Substituting in Eq. (11.18), we have

$$\alpha\frac{\mathrm{d}^2 U_1}{\mathrm{d}\phi^2} + \alpha U_1 - \alpha(1+e\cos\phi)^2 + \mathcal{O}\left(\alpha^2\right) = 0, \tag{11.20}$$

so that at first-order in α,

$$\begin{aligned}\frac{\mathrm{d}^2 U_1}{\mathrm{d}\phi^2} + U_1 &= (1+e\cos\phi)^2 \\ &= \left(1+\frac{1}{2}e^2\right) + 2e\cos\phi + \frac{1}{2}e^2\cos 2\phi.\end{aligned} \tag{11.21}$$

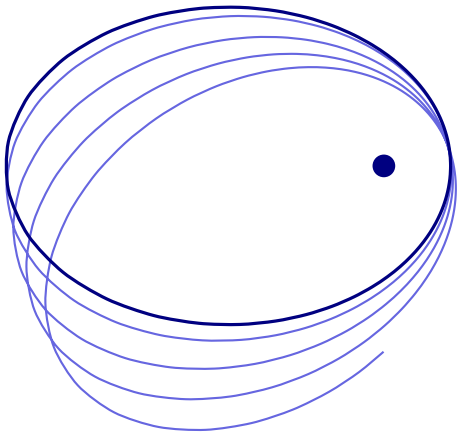

Fig. 11.2: In the non-relativistic Kepler problem, a particle follows the same perfect ellipse (dark blue orbit) eternally. General Relativity introduces a perturbing effect around the Newtonian orbit that causes the body's elliptical orbit to precess (pale blue orbit) in the direction of its rotation; this effect has been measured in Mercury, Venus and Earth to strong agreement with the prediction of General Relativity. The black dot within the orbits represents the center of attraction at a focus of the ellipse.

The particular integral of this equation is

$$U_1(\phi) = \left(1 + \frac{1}{2}e^2\right) + e\phi \sin\phi - \frac{1}{6}e^2 \cos 2\phi. \tag{11.22}$$

The corrections to the Newtonian orbit are very small from the first and third terms on the right since they get multiplied by the small quantity α; however, the amplitude of the second term can grow over many orbits so we retain this.

The general relativistic orbit is then

$$\begin{aligned} u(\phi) &\approx \frac{GM}{h^2}(1 + e\cos\phi + e\alpha\phi\sin\phi) \\ &\approx \frac{GM}{h^2}(1 + e[\cos\phi\cos\alpha\phi + \sin\phi\sin\alpha\phi]), \end{aligned} \tag{11.23}$$

where the second line is correct to first-order in $\alpha\phi$. It follows that

$$u(\phi) \approx \frac{GM}{h^2}\{1 + e\cos\left[\phi(1-\alpha)\right]\}. \tag{11.24}$$

We see from this expression that the orbit is not closed since r is periodic in ϕ with period $2\pi/(1-\alpha)$.

This means that the ellipse *precesses*, with the angle ϕ at perihelion increasing by

$$\Delta\phi = 2\pi\left(\frac{1}{1-\alpha} - 1\right) \approx 2\pi alpha \tag{11.25}$$

per revolution (see Fig. 11.2). Substituting for α, and expressing h^2 in terms of the semi-major axis using Eq. (11.15), we have

$$\Delta\phi = \frac{6\pi GM}{a(1-e^2)c^2}. \tag{11.26}$$

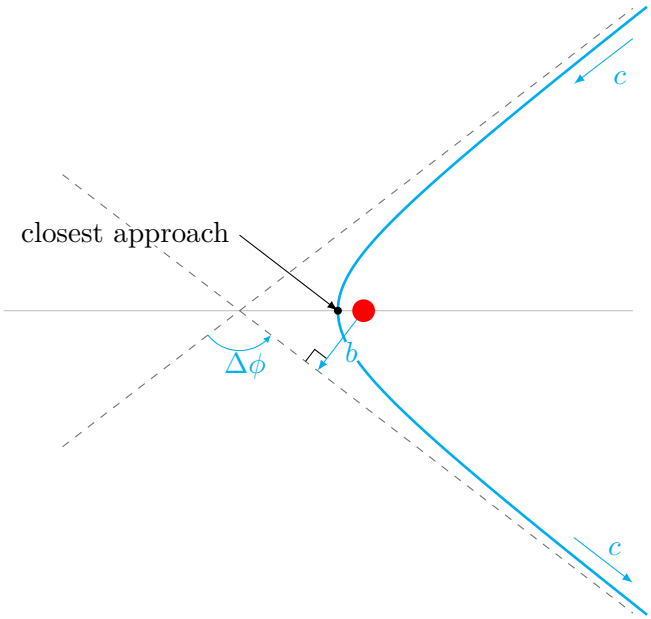

Fig. 11.3: Bending of light with impact parameter b by a spherical mass. The total deflection angle is $\Delta\phi$.

This is largest when the orbit is small and highly eccentric. In the Solar System, the largest effect is for Mercury, which has $a = 5.8 \times 10^{10}$m and $e = 0.2$. Combining with the Solar mass, $M_\odot = 2 \times 10^{30}$kg (so $\mu = 1.5$km), we predict $\Delta\phi = 0.10$ arcsec or

$$\Delta\phi = 43 \text{ arcsec per century} \tag{11.27}$$

using the period of 88 days. The measured precession is

$$\Delta\phi = (574.1 \pm 0.1) \text{ arcsec per century}, \tag{11.28}$$

but most of this is due to the perturbing effect of the other planets. Once these are corrected for, the residual is

$$\Delta\phi = (43.1 \pm 0.5) \text{ arcsec per century}, \tag{11.29}$$

in beautiful agreement with the prediction of General Relativity.

11.3 The Bending of Light

To describe the gravitational deflection of light, we use Eq. (11.9), which we write in the form

$$\frac{\mathrm{d}^2 u}{\mathrm{d}\phi^2} + u = \frac{3GM}{c^2} u^2. \tag{11.30}$$

For $M = 0$, this is solved by the straight line

$$u = \frac{\sin\phi}{b}, \tag{11.31}$$

where b is the impact parameter (see Fig. 11.3). For $\beta \ll 1$, where

$$\beta \equiv \frac{3GM}{c^2 b}, \tag{11.32}$$

we can proceed perturbatively writing

$$b\,u(\phi) = \sin\phi + \beta U_1(\phi) + \beta^2 U_2(\phi) + \cdots, \tag{11.33}$$

where the ${}_U n$ are dimensionless. To first order in β, we have

$$\begin{aligned} \frac{\mathrm{d}^2 U_1}{\mathrm{d}\phi^2} + U_1 &= \sin^2\phi \\ &= \frac{1}{2}(1 - \cos 2\phi). \end{aligned} \tag{11.34}$$

This is solved by

$$U_1(\phi) = C_1 \sin\phi + C_2 \cos\phi + \frac{1}{2}\left(1 + \frac{1}{3}\cos 2\phi\right), \tag{11.35}$$

where C_1 and C_2 are integration constants. For impact parameter b, we require that $bu \to \sin\phi$ as $\phi \to \pi$, so that $C_1 = 0$ and $C_2 = 2/3$. It follows that

$$u(\phi) = \frac{\sin\phi}{b} + \frac{3GM}{c^2 b^2}\left[\frac{2}{3}\cos\phi + \left(1 + \frac{1}{3}\cos 2\phi\right)\right]. \tag{11.36}$$

At the far end of the light path, $u \to 0$ as $\phi \to -\Delta\phi$ (with $|\Delta\phi| \ll 1$) where

$$-\frac{\Delta\phi}{b} + \frac{3GM}{c^2 b^2} \times \frac{4}{3} = 0 \tag{11.37}$$

to first order in β. It follows that the total deflection of the light ray is[1]

$$\Delta\phi = \frac{4GM}{c^2 b}. \tag{11.38}$$

For light grazing the Sun, $b = R_\odot = 6.96 \times 10^5$km, and the total deflection predicted is

$$\Delta\phi = 1/75 \text{ arcsec}. \tag{11.39}$$

This *prediction* of General Relativity was first verified in the famous 1919 eclipse expeditions led by Arthur Eddington. High-precision tests of light bending have subsequently been made using extragalactic radio sources rather than stars in our Galaxy (as used by Eddington). The radio sources (quasars) can be measured close to the Sun even when there is no lunar eclipse and are not affected by the atmosphere; the angular shift in the position of quasars when they are eclipsed by the Sun has been used to test General Relativity with a relative accuracy approaching 10^{-4}. More extreme manifestations of light bending have now been observed in astrophysical systems. For example, if a distant source is sufficiently aligned with a foreground galaxy, it is possible for multiple light paths to connect the source to the observer. This phenomenon, known as strong gravitational lensing, leads to strong distortion of the image of the source and even multiple images. An example of strong lensing is the Cosmic Horseshoe; see Fig. 11.4.

[1] This is twice the value obtained with a Newtonian calculation for a particle travelling at c. This reflects the fact that for a massless particle both the g_{tt} and g_{rr} metric perturbations contribute to the particle dynamics, while for a slowly-moving massive particle only g_{tt} is relevant.

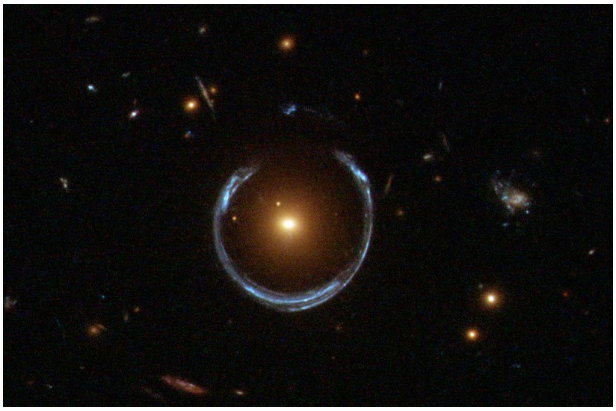

Fig. 11.4: The *Cosmic Horseshoe* is a beautiful example of strong gravitational lensing. A distant galaxy (blue) lies directly behind a foreground luminous red galaxy on our line of sight. The light from the former is bent by the massive foreground galaxy. Due to the close alignment, multiple light paths from the background galaxy can reach us on Earth, giving rise to the extreme ring-like distortion (called an *Einstein ring*) in the image of the background galaxy.

Chapter 12

Schwarzschild Black Holes

We noted in Chapter 10 that the Schwarzschild metric is singular at $r = 2\mu$ (and also at $r = 0$). So far, we have been concerned with motion in the region $r > 2\mu$, but what should we make of the region $r < 2\mu$ and the hypersurface $r = 2\mu$? In this topic, we discuss the causal structure of the Schwarzschild solution. We shall show that the region$r < 2\mu$ represents a *black hole* – a region of spacetime from which no particle can escape to spatial infinity. Moreover, we shall show that the surface $r = 2\mu$, while being perfectly regular, does have important physical significance as an *event horizon*. To uncover these global properties of the Schwarzschild solution, we shall have to introduce new coordinates that allow us to join the regions $r < 2\mu$ and $r > 2\mu$ in a continuous manner.

12.1 Singularities in the Schwarzschild Metric

The Schwarzschild line element from Eq. (10.40) is

$$\mathrm{d}s^2 = c^2\left(1 - \frac{2\mu}{r}\right)\mathrm{d}t^2 - \left(1 - \frac{2\mu}{r}\right)^{-1}\mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2\,, \tag{12.1}$$

where, recall, $\mu \equiv GM/c^2$. The metric is singular at $r = 0$ and at $r = 2\mu$. The latter is generally called the *Schwarzschild radius*

$$r_s = 2\mu = \frac{2GM}{c^2}. \tag{12.2}$$

The Schwarzschild solution is a vacuum solution of the Einstein field equations, so only holds down to the surface of the spherical massive body. If the radius of the body exceeds r_s, the singularities in the Schwarzschild solution are of no consequence (the spacetime inside the body is described by some non-vacuum solution of the field equations). For example, for the Sun, $r_s \approx 3$km but the Solar radius $R_\odot = 7 \times 10^5$km $\gg r_s$. However, if the central body is so compact that it is smaller than its Schwarzschild radius, we must take (part of) the region $r < 2\mu$ seriously.

To investigate the nature of the singularities at $r = 0$ and $r = 2\mu$ we should consider coordinate-independent properties of spacetime there. Invariants formed from the Riemann curvature tensor are suitable candidates. Since $R_{\mu\nu} = 0$ by construction, the first non-trivial invariant to consider is

$$R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} \propto \frac{\mu^2}{r^6}. \tag{12.3}$$

This scalar (called the *Kretschmann scalar*) is telling us something about the magnitude of tidal effects; it has the same dependence on

M

and r as the square of the tidal tensor in Newtonian gravity. We see that the Kretschmann scalar is regular at $r = 2\mu$ but singular at $r = 0$. The apparent singularity in the metric at $r = 2\mu$ is just an artefact of our particular choice of coordinate system and so is a *coordinate singularity*. It can be removed by adopting different coordinates (as we show later). However, the curvature of spacetime is infinite at $r = 0$ so this is a genuine intrinsic singularity.

- Example: Coordinate singularities for the 2-sphere.

  In Chapter 3, we considered the 2-sphere described in cylindrical coordinates (ρ, ϕ), for which the line element is

  $$\mathrm{d}s^2 = \frac{a^2\,\mathrm{d}\rho^2}{(a^2 - \rho^2)} + \rho^2\,\mathrm{d}\phi^2\,, \tag{12.4}$$

  where a is the radius of the sphere. The coordinate ranges are $0 \leq \rho \leq a$ and $0 \leq \phi \leq 2\pi$ and these cover just one hemisphere. The metric is singular at $\rho = a$, but we know the 2-sphere is perfectly regular everywhere with nothing odd happening at the equator. The curvature invariants are all regular (and actually are the same everywhere) as is necessary for $\rho = a$ to be only a coordinate singularity.

The region $r < 2\mu$ of the Schwarzschild solution has some very odd properties. Let us denote the region of spacetime with $r > 2\mu$ as *region I* and $r < 2\mu$ as *region II*. In region I, the coordinate basis vector $\mathbf{e}_0 \equiv \partial/\partial t$ is timelike,

$$\mathbf{g}(\mathbf{e}_0, \mathbf{e}_0) = g_{00} = c^2\left(1 - \frac{2\mu}{r}\right) > 0. \tag{12.5}$$

This means that a curve at constant (r, θ, ϕ) is timelike. Also, in region I the spatial basis vectors are spacelike. However, in region II the $\mathbf{e}_0$ basis vector is spacelike since $g_{00} < 0$. Moreover, the basis vector $\partial/\partial r$ is timelike! This means that a particle cannot stay at fixed (r, θ, ϕ) in region II, no matter what they do, as their worldline would then be spacelike rather than timelike. It is as if the time and radial coordinates swap characters for $r < 2\mu$.

12.2 Causal Structure

As timelike curves travel inside the lightcone, and null curves on the lightcone, the lightcones display the causal structure of spacetime (i.e., which events can influence others). We can construct the lightcones in Schwarzschild spacetime by considering radial null geodesics.

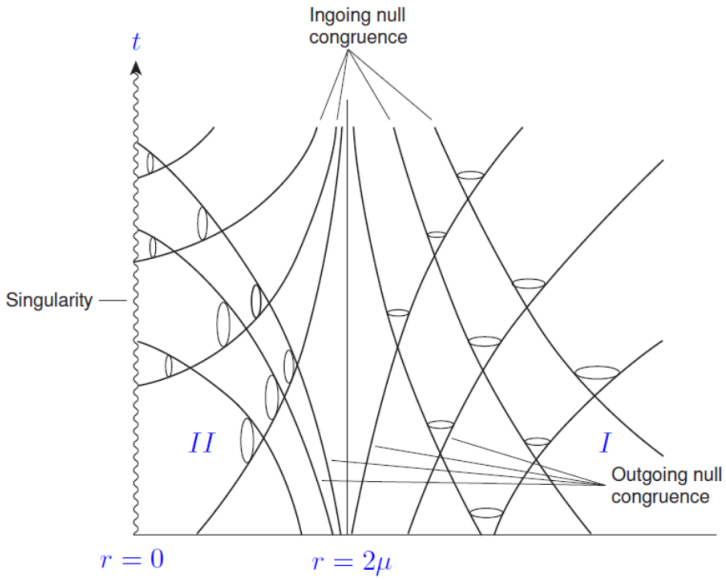

Fig. 12.1: Lightcone structure of the Schwarzschild solution. Ingoing and outgoing radial null geodesics are shown in the (t, r) plane. Both are discontinuous at the Schwarzschild radius $r = 2\mu$.

12.2.1 Radial Null Geodesics

Radial geodesics have $\mathrm{d}\theta = \mathrm{d}\phi = 0$. For null geodesics, it follows that

$$0 = \mathrm{d}s^2 = c^2\left(1 - \frac{2\mu}{r}\right)\mathrm{d}t^2 - \left(1 - \frac{2\mu}{r}\right)^{-1}\mathrm{d}r^2$$
$$\implies \frac{\mathrm{d}(ct)}{\mathrm{d}r} = \pm\left(1 - \frac{2\mu}{r}\right)^{-1}. \tag{12.6}$$

With the + sign, this is solved by

$$ct = r + 2\mu \ln\left|\frac{r}{2\mu} - 1\right| + \text{const.}, \tag{12.7}$$

where the absolute value means we can consider both $r > 2\mu$ and $r < 2\mu$. In region I, these are a family of *outgoing* radial null geodesics. With the − sign, Eq. (12.6) is solved by

$$ct = -r - 2\mu \ln\left|\frac{r}{2\mu} - 1\right| + \text{const.} \tag{12.8}$$

In region I, these are a family of ingoing radial null geodesics. Note the ingoing and outgoing solutions are related by time reversal, $t \to -t$. (Since the Schwarzschild metric is symmetric under time reversal, the time reverse of a solution of the geodesic equations will also be a solution.) These paths of radial null geodesics are plotted in Fig. 12.1 along with the lightcones that they generate.

The main features in region I are as follows.

- For $r \gg 2\mu$, the radial null geodesics are straight lines in the $r - ct$ plane with gradient ± 1, corresponding to the usual lightcones in Minkowski space.

- As $r \to 2\mu$, the lightcones are squashed in the radial direction.
- It *seems* to take an infinite coordinate time for an ingoing photon to reach $r = 2\mu$.
- Similarly, all outgoing rays seem to originate from $r = 2\mu$ and $ct \to -\infty$.

The behaviour as $r \to 2\mu$ is misleading since an ingoing photon actually needs only a finite change in its affine parameter to cross $r = 2\mu$. To see this, recall from the geodesic equations that

$$\frac{\mathrm{d}t}{\mathrm{d}\lambda} = k\left(1 - \frac{2\mu}{r}\right)^{-1}, \tag{12.9}$$

where λ is an affine parameter. It follows that

$$\frac{\mathrm{d}r}{\mathrm{d}\lambda} = k\frac{\mathrm{d}r}{\mathrm{d}t}\left(1 - \frac{2\mu}{r}\right)^{-1} = \pm kc, \tag{12.10}$$

so that

$$r = \pm ck\lambda + \text{const.} \tag{12.11}$$

We see that an ingoing photon reaches $r = 2\mu$ from some $r_0 > 2\mu$ for a finite increment in the affine parameter.

What happens when the photon gets to $r = 2\mu$? Eq. (12.11) tells us that the photon passes straight through continuing to the singularity at $r = 0$ as λ increases further. Moreover, since $k > 0$, Eq. (12.9) implies that $\mathrm{d}t/\mathrm{d}\lambda < 0$ in region II. We thus see that future-directed ingoing null geodesics in region I reach $r = 2\mu$ and $t = \infty$ at a finite value of their affine parameter, and then extend into region II moving towards $r = 0$ with decreasing t! It is worth pausing to restate what we have achieved here: region I (for $r \neq 0$) and region II are spherically-symmetric vacuum solutions of the Einstein field equations, each covering part of spacetime, but because of the coordinate singularity it is not immediately clear how these fit together. By extending ingoing null geodesics from region I, we see that the causal future of this region of the Schwarzschild solution includes a type-II region in which the forward lightcone is tipped over towards $r = 0$. Any particle that falls into this type-II region will inevitably fall towards the singularity at $r = 0$ *no matter what they do to try and avoid it.*

The hypersurface $r = 2\mu$ is an example of an *event horizon* - the outermost boundary of a region of spacetime from which no particle can escape to spatial infinity. The region II in the causal future of region I is a *black hole* - a region of spacetime from which no particle can escape to spatial infinity. The event horizon acts like a one-way membrane, particles can fall through it from $r > 2\mu$ to $r < 2\mu$, but not the other way around. We shall shortly construct a non-singular coordinate system (*ingoing Eddington–Finkelstein coordinates*) that covers both region I and the region II in its causal future.

12.2.1.1 The Causal Past of Region I

What if we had considered outgoing null geodesics instead in region I? These seem to emerge from $r = 2\mu$ at $t = -\infty$, but, actually, the affine parameter changes by a finite amount in travelling from $r = 2\mu$ to some $r_0 > 2\mu$. If we try and extend such outgoing

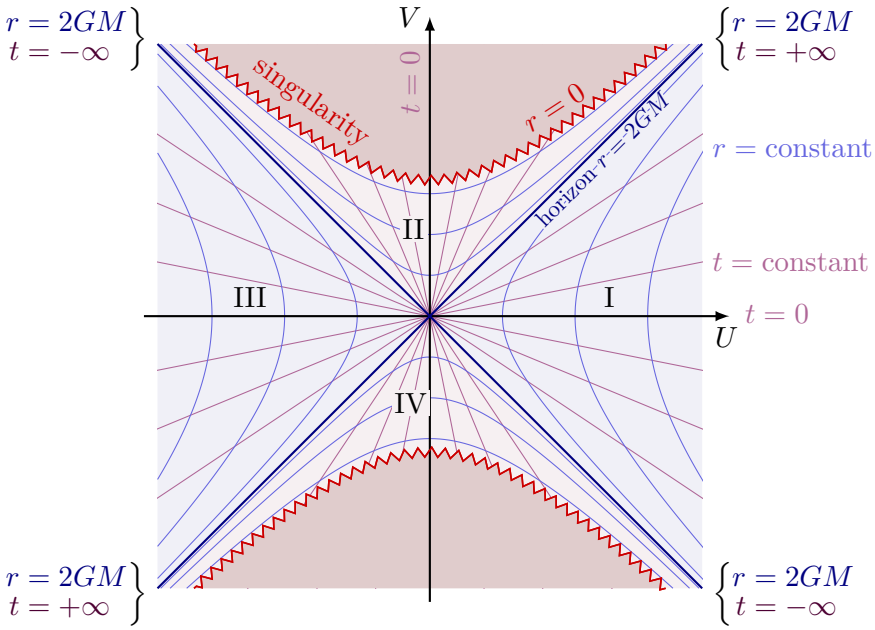

Fig. 12.2: Kruskal–Szekeres diagram. The quadrants are the black hole interior (II), the white hole interior (IV) and the two exterior regions (I and III). The bold dark blue 45° lines, which separate these four regions, are the event horizons. The zigzag dark red hyperbolas which bound the top and bottom of the diagram are the physical singularities. The pale blue hyperbolas represent contours of the Schwarzschild r coordinate, and the straight pale purple lines through the origin represent contours of the Schwarzschild t coordinate. The definitions of the Kruskal-Szekeres coordinates are given in Fig. 12.3

geodesics back to smaller values of the affine parameter λ, we find that they join onto geodesics in region II that have r increasing with λ and t decreasing. This is a type-II region, but it cannot be the same as that in the causal future of region I since the forward lightcone is now directed away from $r = 0$. Such a region is a *white hole* in which particles are inevitably expelled to $r > 2\mu$!

The original Schwarzschild solution, with its single regions I and II separated by a coordinate singularity, do not cover the entire spacetime. Rather, it can be shown that the entire spacetime contains a black hole, a white hole and *two* type-I regions. It is possible to construct coordinates that cover all of these regions in a non-singular manner (*Kruskal–Szekeres* coordinates, see Fig. 12.2 and 12.3). However, the existence of white holes as a physical reality is very doubtful. Black holes can form as the result of gravitational collapse, but white holes require the singularity to exist in the past.

12.2.2 Radially-Infalling Particles

We can repeat the analysis above for the timelike worldlines of massive infalling particles. For such particles, setting $h = 0$ in the radial effective potential for Eq. (10.59), we have

$$\frac{1}{2}\dot{r}^2 - \frac{GM}{r} = \frac{1}{2}c^2(k^2 - 1), \tag{12.12}$$

$$\left.\begin{aligned} U &= \sqrt{\frac{r}{2GM}-1}\, e^{\frac{r}{4GM}} \cosh\left(\frac{t}{4GM}\right) \\ V &= \sqrt{\frac{r}{2GM}-1}\, e^{\frac{r}{4GM}} \sinh\left(\frac{t}{4GM}\right) \end{aligned}\right\} \quad \text{for } r > 2GM$$

$$\left.\begin{aligned} U &= \sqrt{1-\frac{r}{2GM}}\, e^{\frac{r}{4GM}} \sinh\left(\frac{t}{4GM}\right) \\ V &= \sqrt{1-\frac{r}{2GM}}\, e^{\frac{r}{4GM}} \cosh\left(\frac{t}{4GM}\right) \end{aligned}\right\} \quad \text{for } r < 2GM$$

$$\Rightarrow \begin{cases} U^2 - V^2 = \left(\dfrac{r}{2GM}-1\right) e^{\frac{r}{2GM}} & \\ V = \tanh\left(\dfrac{t}{4GM}\right) U & \text{for } r > 2GM \\ V = \coth\left(\dfrac{t}{4GM}\right) U & \text{for } r < 2GM \end{cases}$$

Fig. 12.3: Kruskal-Szekeres coordinates for Schwarzschild spacetime: Kruskal–Szekeres coordinates on a black hole geometry are defined, from the Schwarzschild coordinates (t, r, θ, ϕ), by replacing t and r by a new timelike coordinate V and a new spacelike coordinate U, defined seperately for the exterior region outside the event horizon and the interior region.

where overdots denote differentiation with respect to proper time. Recall also that $\dot{t}(1 - 2\mu/r) = k$. Taking a further time derivative of Eq. (12.12) gives

$$\ddot{r} = -\frac{GM}{r^2}, \tag{12.13}$$

which has exactly the same form as the Newtonian equation of motion (but the time variable and r mean different things).

Consider a particle that starts at rest at infinity, so that $k = 1$. For an in-falling particle,

$$\dot{r} = -\sqrt{\frac{2\mu c^2}{r}}, \tag{12.14}$$

so that

$$c(\tau - \tau_0) = \frac{2}{3}\left[\left(\frac{r_0^3}{2\mu}\right)^{1/2} - \left(\frac{r^3}{2\mu}\right)^{1/2}\right], \tag{12.15}$$

where we have taken $r(\tau_0) = r_0$. This gives us the proper time as a function of r. It takes finite amounts of proper time to reach $r = 2\mu$ and $r = 0$. We can also determine coordinate time t as a function of r using

$$\frac{\mathrm{d}(ct)}{\mathrm{d}r} = \frac{c\dot{t}}{\dot{r}} = -\sqrt{\frac{r}{2\mu}}\left(1 - \frac{2\mu}{r}\right)^{-1}. \tag{12.16}$$

The solution of this is

$$\begin{aligned} c(t - t_0) &= -2\mu \int_{r_0/(2\mu)}^{r/(2\mu)} \frac{x^{3/2}}{x-1}\,\mathrm{d}x \\ &= -2\mu\left[\frac{2}{3}x^{3/2} + 2\sqrt{x} + \ln\left|\frac{\sqrt{x}-1}{\sqrt{x}+1}\right|\right]_{r_0/(2\mu)}^{r/(2\mu)}, \end{aligned} \tag{12.17}$$

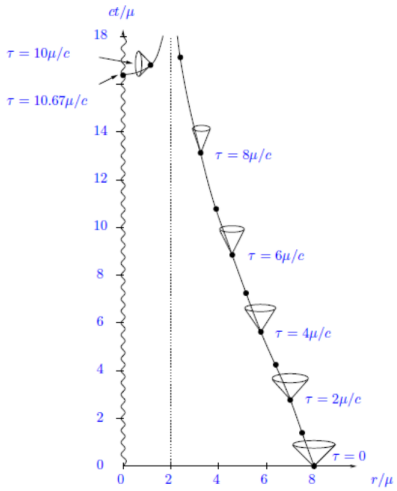

Fig. 12.4: Trajectory of a radially-infalling particle released from rest at infinity. The dots correspond to unit intervals of $c\tau/\mu$, where τ is the particle's proper time. We have taken $\tau = t_0 = 0$ at $r_0 = 8\mu$. The particle reaches the singularity at $r = 0$ at $c\tau = 32\mu/3$.

where we have taken $t(r_0) = t_0$. The integral is improper as $r \to 2\mu$ and diverges logarithmically, so that $t \to \infty$ as $r \to 2\mu$. For smaller r, coordinate time then decreases with increasing τ. The resulting wordline is plotted in Fig. 12.4 2 for a particle with $\tau_0 = 0$, $r_0 = 8\mu$ and $t_0 = 0$.

Consider the in-falling particle as observed by a stationary observer at $r = \infty$. Light from the particle only reaches the distant observer if it is emitted at $r > 2\mu$. Since t is proper time for the distant observer (where the spacetime tends to Minkowski), they perceive that it takes an infinite amount of their proper time for the particle to reach $r = 2\mu$ from any finite $r > 2\mu$. The light signals emitted as $r \to 2\mu$ are also infinitely redshifted when they reach the distant observer.[1] As seen from the distant observer, the in-falling particle is never seen to cross $r = 2\mu$; rather, it seems to hover there, forever becoming redder (and dimmer) and redder.

12.3 Eddington–Finkelstein Coordinates

The time coordinate of the Schwarzschild solution is useful and physically meaningful as $r \to \infty$ (as proper time experienced by a stationary observer there), but is inappropriate as $r \to 2\mu$ and beyond. Instead, let us try and construct coordinates that cover region I and the type-II region in its causal future without coordinate singularities. We do this by adopting coordinates that are adapted to radially-infalling photons in such a way that their worldlines are continuous through $r = 2\mu$. Recall from Eq. (12.8) that

$$ct = -r - 2\mu \ln\left|\frac{r}{2\mu} - 1\right| + \text{const.} \tag{12.18}$$

[1] We shall work out the details later. For the moment, recall from Chapter 10 that the gravitational redshift between static observers becomes infinite as $r \to 2\mu$. Moreover, the signal is further redshifted for an in-falling emitter due to the Doppler effect.

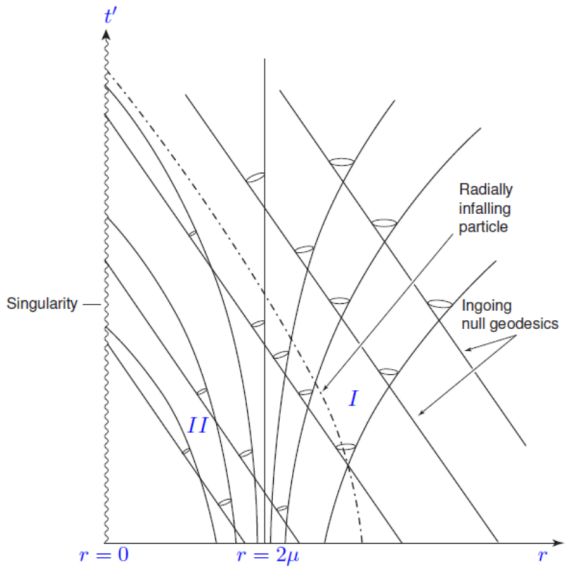

Fig. 12.5: Lightcone structure of Schwarzschild spacetime in ingoing Eddington–Finkelstein coordinates. Ingoing radial null geodesics are straight lines at 45° to the coordinate axes. The path of a massive radially-infalling particle is also shown (dot-dashed line). Outgoing radial null geodesics in region I are still discontinuous at $r = 2\mu$.

for such photons. Let us define a new time coordinate t' by

$$ct' \equiv ct + 2\mu \ln\left|\frac{r}{2\mu} - 1\right|. \tag{12.19}$$

In terms of t' and r, ingoing radial null geodesics have

$$ct' = -r + \text{const.}, \tag{12.20}$$

and so are straight lines at 45° to the coordinate axes (as in Minkowski space).

For outgoing photons, recall from Eq. (12.7) that in the original coordinates

$$ct = r + 2\mu \ln\left|\frac{r}{2\mu} - 1\right| + \text{const.}, \tag{12.21}$$

so that

$$ct' = r + 4\mu \ln\left|\frac{r}{2\mu} - 1\right| + \text{const.} \tag{12.22}$$

The spacetime diagram of these null geodesics and the associated forward lightcones are shown in Fig. 12.5. The ingoing geodesics are now continuous at $r = 2\mu$, with the lightcones tilted over inwards there and in the black hole (region II). To find the line element in these coordinates, we use

$$c\,\mathrm{d}t' = c\,\mathrm{d}t + \left(\frac{r}{2\mu} - 1\right)^{-1} \mathrm{d}r\,, \tag{12.23}$$

so that

$$\mathrm{d}s^2 = \left(1 - \frac{2\mu}{r}\right)\left[c\,\mathrm{d}t' - \left(\frac{2\mu}{r} - 1\right)^{-1}\mathrm{d}r\right]^2 - \left(1 - \frac{r}{2\mu}\right)^{-1}\mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2\,. \tag{12.24}$$

Simplifying, we find

$$\mathrm{d}s^2 = c^2\left(1 - \frac{2\mu}{r}\right)\mathrm{d}t'^2 - \frac{4\mu c}{r}\,\mathrm{d}t'\,\mathrm{d}r - \left(1 + \frac{2\mu}{r}\right)\mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2\,. \tag{12.25}$$

This is no longer singular at $r = 2\mu$ and, indeed, is regular for the whole range $0 < r < \infty$. The time component of the metric does vanish at $r = 2\mu$, but this is not problematic since the metric is nondiagonal; the determinant is non-zero for $r > 0$ and so the inverse metric exists. The coordinates (t', r, θ, ϕ) are called *ingoing Eddington–Finkelstein coordinates* (sometimes *advanced* is used instead of ingoing).

12.3.1 Outgoing Eddington–Finkelstein Coordinates

In ingoing Eddington–Finkelstein coordinates, the outgoing geodesics in region I are discontinuous at $r = 2\mu$ (see Fig. 12.5). They seem to originate from $r = 2\mu$ at $t \to -\infty$ but, as in the original Schwarzschild coordinates, the affine parameter is finite there. As we discussed earlier, this is because the causal past of region I is really another type-II region, but with the character of a white hole rather than a black hole.The ingoing Eddington–Finkelstein coordinates do not cover this type-II region of spacetime. However, we can always construct *outgoing Eddington-Finkelstein coordinates* that cover region I and the white hole in its causal past by taking

$$ct^* \equiv ct - 2\mu \ln\left|\frac{r}{2\mu} - 1\right|. \tag{12.26}$$

In terms of r and t^*, the outgoing radial null geodesics have

$$ct^* = r + \text{const.} \tag{12.27}$$

while the ingoing geodesics have

$$ct^* = -r - 4\mu \ln\left|\frac{r}{2\mu} - 1\right| + \text{const.} \tag{12.28}$$

The outgoing null geodesics are now straight lines at 45° to the coordinate axes and the lightcones tip over to point outwards at $r = 2\mu$ and within the white hole region $r < 2\mu$.

Of course, the ingoing geodesics are now discontinuous at $r = 2\mu$, with $t^* \to \infty$ as $r \to 2\mu$ but finite affine parameter. These ingoing geodesics can be extended into the black hole region in the causal future of region I, but this region is not covered by the outgoing Eddington–Finkelstein coordinates.

To find the line element in these coordinates, we use

$$c\,\mathrm{d}t^* = c\,\mathrm{d}t - \left(\frac{r}{2\mu} - 1\right)^{-1}\mathrm{d}r\,, \tag{12.29}$$

so that

$$\mathrm{d}s^2 = c^2\left(1 - \frac{2\mu}{r}\right)\mathrm{d}t^{*2} + \frac{4\mu c}{r}\,\mathrm{d}t^*\,\mathrm{d}r - \left(1 + \frac{2\mu}{r}\right)\mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2\,. \tag{12.30}$$

This line element is also regular for the whole range $0 < r < \infty$. The coordinates (t^*, r, θ, ϕ) are called *outgoing Eddington–Finkelstein coordinates* (sometimes *retarded* is used instead of outgoing). As noted earlier, the global structure of Schwarzschild spacetime can be shown to consist of a black hole, a white hole and two type-I regions. The ingoing and outgoing Eddington–Finkelstein coordinates each cover part of this, but it is possible to combine the two coordinate systems to form *Kruskal–Szekeres coordinates* that cover the global spacetime in a non-singular way.

12.4 Formation of Black Holes

The Schwarzschild solution is a highly idealised configuration, with vanishing energy momentum everywhere except at the singularity at $r = 0$. However, we believe that region I and the type-II region in its causal future are realised as the endpoint of stellar evolution of sufficiently massive stars. After a star has expended its nuclear fuel, it will cool and lose pressure support causing it to contract under the influence of gravity. If the star contracts to a radius smaller than $2GM/c^2$, it must inevitably collapse to form a singularity resulting in a black hole.

In practice, electron degeneracy pressure becomes important for a cold star that collapses to sufficient density, and this can hold the star up against gravitational collapse if the mass is below around $1.4M_\odot$ (known as the *Chandrasekhar limit*). Such *white dwarfs* are comparable in size to the Earth, but with the mass of the Sun. For more massive stars, electron degeneracy pressure is insufficient to halt gravitational collapse. In this case, the star collapses further until it reaches such densities that it becomes energetically favourable for electrons and protons to form (more massive) neutrons[2] and neutrinos forming a *neutron star*. There is uncertainty over the equation of state for matter at the extreme densities found at the centre of a neutron star, but we think that the maximum mass of a neutron star is around $4M_\odot$ and the radius would be around only 10km. A star more massive than this will inevitably collapse to form a black hole.

12.4.1 Spherically-Symmetric Collapse of Dust

As a toy-model for the formation of a black hole, consider the spherically-symmetric collapse of a cloud of pressure-free dust. As there is no pressure support, the dust follows geodesics in spacetime. By Birkhoff's theorem (see Subsection 10.2.1, the spacetime *outside* the dust cloud will be described by the Schwarzschild solution, so we can treat the outside edge of the cloud as a massive particle free-falling radially in Schwarzschild spacetime. For simplicity, we shall assume that the collapse starts from rest at infinity ($k = 1$), and consider how the collapse appears to a stationary observer at rest at large radius $r \gg \mu$. The trajectory of the edge of the dust cloud is shown in Eddington–Finkelstein coordinates in Fig. 12.6. As discussed earlier, the distant observer will never see the surface of the dust cloud pass through $r = 2\mu$. Light emitted after the surface crosses $r = 2\mu$ will instead end up at the singularity at $r = 0$. Moreover, the light is increasingly

[2] Although the neutron is more massive than the sum of the proton and electron masses, the Fermi energy of neutrons is lower than electrons at the same number density as their mass is so much higher.

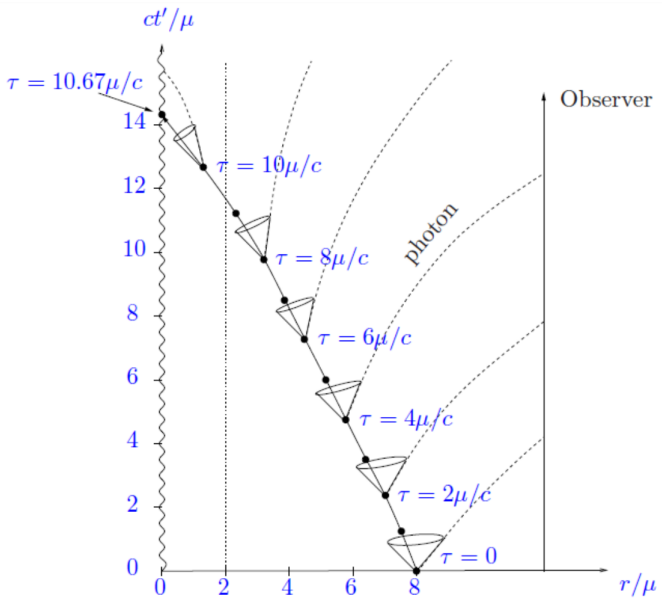

Fig. 12.6: Collapse of the surface of a pressure-free dust cloud to form a black hole in ingoing Eddington–Finkelstein coordinates. The cloud's surface started at rest at infinity, and we have chosen $\tau = t' = 0$ at $r = 8\mu$.

redshifted as $r \to 2\mu$ so the frequency of light and the arrival rate of photons as measured by the distant observer fall to zero. It follows that the distant observer sees the collapse slow down as $r \to 2\mu$, with the cloud becoming more and more red and more and dim.

Let us now work out some of the quantitative details. Before the surface of the star crosses $r = 2\mu$, we can describe the collapse in terms of the standard Schwarzschild coordinates (t, r, θ, ϕ). Suppose the edge of the cloud emits a radially outgoing photon at coordinates (t_E, r_E), which is received at (t_R, r_R) by the distant stationary observer. Since these coordinates lie on an outgoing null geodesic, we have

$$ct_R - r_R - 2\mu \ln\left(\frac{r_R}{2\mu} - 1\right) = ct_E - r_E - 2\mu \ln\left(\frac{r_E}{2\mu} - 1\right). \tag{12.31}$$

The relation between ct_E and r_E is given by the path of the edge of the dust cloud; using Eq. (12.17), as $r_E \to 2\mu$, we have

$$ct_E = 2\mu \ln\left(\frac{\sqrt{r_E/2\mu} + 1}{\sqrt{r_E/2\mu} - 1}\right) + \text{const.} \tag{12.32}$$

Using this in Eq. (12.31), we have

$$\begin{aligned} ct_R &\to 2\mu \ln\left(\frac{\sqrt{r_E/2\mu} + 1}{\sqrt{r_E/2\mu} - 1}\right) - 2\mu \ln\left(\frac{r_E}{2\mu} - 1\right) + \text{const.} \\ &= 2\mu \ln\left(\sqrt{r_E/2\mu} - 1\right)^{-2} + \text{const.} \\ &\approx -4\mu \ln\left(\frac{r_E}{2\mu} - 1\right) + \text{const.} \end{aligned} \tag{12.33}$$

It follows that

$$r_E(t_R) = 2\mu + a \exp\{(-ct_R/4\mu)\}, \tag{12.34}$$

where a is an unimportant constant. We see that the radius the cloud has when observed at t_R approaches $r = 2\mu$ exponentially with characteristic time $4\mu/c$. For a Solar-mass cloud, the characteristic time is very short: $4\mu/c = 2 \times 10^{-5}$s.

We can also compute the redshift of the light from the surface. We have

$$\frac{\nu_R}{\nu_E} = \frac{\mathbf{g}(\mathbf{p}_R, \mathbf{u}_R)}{\mathbf{g}(\mathbf{p}_E, \mathbf{u}_E)}, \tag{12.35}$$

where $\mathbf{p}_E$ is the 4-momentum of a radial outgoing photon at emission at r_E, and similarly for $\mathbf{p}_R$, and $\mathbf{u}_E$ is the 4-velocity of the edge of the dust cloud at r_E and $\mathbf{u}_R$ is the 4-velocity of the distant observer. We have

$$u^\mu_E = \left((1 - 2\mu/r_E)^{-1}, -\sqrt{2\mu c^2/r_E}, 0, 0\right), \tag{12.36}$$

for the cloud starting from rest at infinity. For the static observer, taking $r \to \infty$, we have

$$u^\mu_R = (1, 0, 0, 0). \tag{12.37}$$

The 4-momentum of a radial outgoing photon is

$$p^\mu = \left(\frac{\mathrm{d}t}{\mathrm{d}\lambda}, \frac{\mathrm{d}r}{\mathrm{d}\lambda}, 0, 0\right) = \frac{\mathrm{d}t}{\mathrm{d}\lambda}\left(1, c\left(1 - \frac{2\mu}{r}\right), 0, 0\right), \tag{12.38}$$

where we have used the null condition $g_{\mu\nu}p^\mu p^\nu = 0$. As usual,

$$\frac{\mathrm{d}t}{\mathrm{d}\lambda}\left(1, c\left(1 - \frac{2\mu}{r}\right)\right) = k, \tag{12.39}$$

where k is a constant. It follows that

$$\nu_R = kc^2, \tag{12.40}$$

$$\nu_E = kc^2\left(1 - \sqrt{\frac{2\mu}{r_E}}\right)^{-1}, \tag{12.41}$$

so

$$\frac{\nu_R}{\nu_E} = \left(1 - \sqrt{\frac{2\mu}{r_E}}\right). \tag{12.42}$$

Note that this tends to zero as $r_E \to 2\mu$. Finally, using Eq. (12.34), we can write

$$\begin{aligned}\frac{\nu_R}{\nu_E} &\sim 1 - \left[1 + a\exp\left(-\frac{ct_R}{4\mu}\right)\right]^{-1/2} \\ &\sim \frac{1}{2}a\exp\left(-\frac{ct_R}{4\mu}\right),\end{aligned} \tag{12.43}$$

so the observed frequency decreases exponentially with characteristic decay time $4\mu/c$.

Chapter 13

Cosmology

In this chapter, we shall apply General Relativity to model the Universe as a whole. The late-time Universe is clearly very complicated and we cannot hope to find exact, analytic solutions of the Einstein field equations in this case. However, if we smooth on sufficiently large spatial scales (of the order of 100Mpc $\sim 10^7$ light years), the Universe looks remarkably symmetric in space. This symmetry allows us to find analytic solutions for the spacetime of a smoothed-out universe. These solution are the starting point for all cosmological studies, with the structures that we observe on smaller scales treated as a (not necessarily small amplitude) perturbation to the highly symmetric background.

13.1 Homogeneity and Isotropy

Cosmological observations, most notably of the cosmic microwave background (CMB) radiation, suggest that at any given time the Universe looks very nearly the same in all directions, i.e., is *isotropic* about us. Of course, we can only make observations on our past lightcone and along one particular wordline through spacetime. However, if we assume that we are not in a privileged location, then it should be possible to construct a whole class of observers, filling space, who all observe the Universe to be isotropic. Moreover, with appropriate synchronisation of their clocks, the fundamental observers must agree on what they observe at any given proper time, i.e., the Universe is spatially *homogeneous*. We shall call these preferred observers *fundamental observers*.

From these symmetries alone, it follows that the fundamental observers must have the following properties

- They must comove with the matter in the Universe, since if they did not the 3-velocity of the matter they measure locally would break isotropy.
- They must be free-falling, since acceleration (i.e., departure from motion along a geodesic) would also break isotropy.
- If we consider the (smoothed out) matter in the Universe as a fluid, the hypersurfaces of constant proper density must be orthogonal to the worldlines of the fundamental observers. Local measurements of the density in the instantaneous rest-space of each fundamental observer must reveal no spatial gradients (or else these would break isotropy), so the local rest-spaces must be tangent to the hypersurfaces of constant density (see Fig. 13.1).

This last point also ensures that there are no pressure gradients in the instantaneous rest-space of the fluid, which would otherwise cause acceleration.

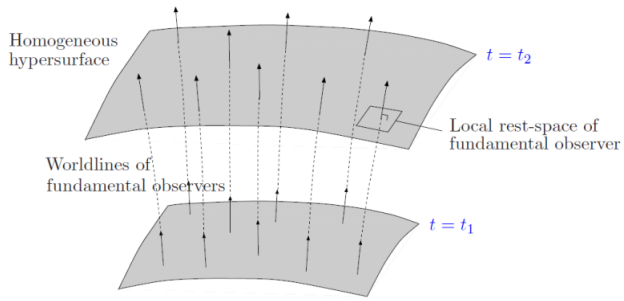

Fig. 13.1: Worldlines of fundamental observers, who move with the matter in the Universe, are orthogonal to the spacelike hypersurfaces of homogeneity. This ensures that the instantaneous rest-spaces of each observer lie in the homogeneous hypersurfaces.

13.1.1 Synchronous Coordinates

We adopt a coordinate system, called *synchronous coordinates*, adapted to the fundamental observers as follows. We assign fixed spatial coordinates x^i $(i = 1, 2, 3)$ to each fundamental observer. We label the surfaces of homogeneity, i.e., the hypersurfaces of constant proper density, with proper time as measured by one of the fundamental observers. We call this *synchronous* or *cosmic time* and denote it by $x^0 = t$. By homogeneity, all fundamental observers will see cosmic time pass at the same rate as their proper time.

The line element in these synchronous coordinates must take the following form:

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 + g_{ij}(t, \vec{x})\,\mathrm{d}x^i\,\mathrm{d}x^j\,, \tag{13.1}$$

where there is an implicit summation over repeated spatial indices. To see this, first note that the wordlines of fundamental observers are of the form $x^\mu = (t, x^i)$, where the x^i are constant for each observer. It follows that $\mathrm{d}x^\mu = \delta^\mu_0\,\mathrm{d}t$ along the wordline so that

$$\mathrm{d}s^2 = g_{00}\,\mathrm{d}t^2 = c^2\,\mathrm{d}t^2\,. \tag{13.2}$$

This must equal $c^2\,\mathrm{d}\tau^2$, where τ is proper time for the observer, so that we can take $t = \tau$. Second, note that any infinitesimal displacement in the hypersurface $t = \text{const.}$ is of the form $\mathrm{d}x^\mu = (0, \mathrm{d}x^i)$, and this must be orthogonal everywhere to the 4-velocity of the fundamental observers, $u^\mu = \delta^\mu_0$. This requires $g_{0i} = 0$. Finally, we must check that the worldlines $x^\mu = (t, x^i)$, for fixed x^i, are geodesics. As t is an affine parameter, we require

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}t^2} + \Gamma^\mu_{\nu\rho}\frac{\mathrm{d}x^\nu}{\mathrm{d}t}\frac{\mathrm{d}x^\rho}{\mathrm{d}t} = 0 \quad \Longrightarrow \quad \Gamma^\mu_{00} = 0. \tag{13.3}$$

This is satisfied for the metric in Eq. (13.1) since

$$\Gamma^\mu_{00} = \frac{1}{2}g^{\mu\nu}(2\partial_0 g_{\nu 0} - \partial_\nu g_{00}) = 0, \tag{13.4}$$

as $g_{00} = c^2$ and $g_{0i} = 0$.

13.2 The Robertson–Walker Metric

The intrinsic geometry of the hypersurfaces $t = \text{const.}$ is determined by the spatial components of the metric $g_{ij}(t, \vec{x})$. This intrinsic geometry must be consistent with isotropy

and homogeneity. This will only be the case for all t if every component of g_{ij} evolves in the same way with t, so that we can write

$$g_{ij}(t,\vec{x}) = -a^2(t)\gamma_{ij}(\vec{x}). \tag{13.5}$$

Here, $a(t)$ is the *scale factor*, which determines the overall scale of the intrinsic geometry of the $t = \text{const.}$ surfaces.

The $\gamma_{ij}(\vec{x})$ play the role (up to scaling) of the 3D metric in the surfaces $t = \text{const.}$; under time-independent coordinate transformations $x^i \to x'^i(\vec{x})$ in spacetime, γ_{ij} transform as a 3D type-$(0,2)$ tensor. We require γ_{ij} to describe a 3D space that is homogeneous and isotropic. Isotropy implies that γ_{ij} must be spherically symmetric so we can always write the 3D line element in spherical-polar coordinates as

$$\begin{aligned} \mathrm{d}\sigma^2 &= \gamma_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j \\ &= B(r)\,\mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2\,, \end{aligned} \tag{13.6}$$

where $\mathrm{d}\Omega^2$ is the metric on the unit 2-sphere. Here, we have used the residual freedom in our coordinates to adopt an area-like radial coordinate r. It follows that the components of the metric and its inverse are

$$\begin{aligned} \gamma_{rr} &= B(r), & \gamma^{rr} &= \frac{1}{B(r)}, \\ \gamma_{\theta\theta} &= r^2, & \gamma^{\theta\theta} &= \frac{1}{r^2}, \\ \gamma_{\phi\phi} &= r^2\sin^2\theta, & \gamma^{\phi\phi} &= \frac{1}{r^2\sin^2\theta}. \end{aligned} \tag{13.7}$$

We can construct a 3D metric connection, ${}^{(3)}\Gamma^i_{jk}$, from γ_{ij} with (independent) non-zero components

$$\begin{aligned} {}^{(3)}\Gamma^r_{rr} &= \frac{1}{2B}\frac{\mathrm{d}B}{\mathrm{d}r}, & {}^{(3)}\Gamma^r_{\theta\theta} &= -\frac{r}{B}, \\ {}^{(3)}\Gamma^r_{\phi\phi} &= -\frac{r\sin^2\theta}{B}, & {}^{(3)}\Gamma^\theta_{r\theta} &= \frac{1}{r}, \\ {}^{(3)}\Gamma^\theta_{\phi\phi} &= -\sin\theta\cos\theta, & {}^{(3)}\Gamma^\phi_{r\phi} &= \frac{1}{r}, \\ {}^{(3)}\Gamma^\phi_{\theta\phi} &= \cot\theta, \end{aligned} \tag{13.8}$$

Recall from Eq. (8.20), that the 3D Riemann curvature tensor is given by

$${}^{(3)}R_{ijk}{}^{l} = -\partial_i {}^{(3)}\Gamma^l_{jk} + \partial_j {}^{(3)}\Gamma^l_{ik} + {}^{(3)}\Gamma^m_{ik}\,{}^{(3)}\Gamma^l_{jm} - {}^{(3)}\Gamma^m_{jk}\,{}^{(3)}\Gamma^l_{im}. \tag{13.9}$$

It has six independent components, but only three of these are non-zero for the spherically-symmetric metric:

$${}^{(3)}R_{r\theta r\theta} = \frac{r}{2B}\frac{\mathrm{d}B}{\mathrm{d}r}, \tag{13.10}$$

$${}^{(3)}R_{r\phi r\phi} = \frac{r}{2B}\frac{\mathrm{d}B}{\mathrm{d}r}\sin^2\theta, \tag{13.11}$$

$${}^{(3)}R_{\theta\phi\theta\phi} = \left(1 - \frac{1}{B}\right) r^2\sin^2\theta. \tag{13.12}$$

Contracting with γ_{ij}, we can form the 3D Ricci tensor:

$$ {}^{(3)}R_{ij} = \gamma^{kl\,(3)}R_{kijl}, \tag{13.13} $$

which has non-zero (independent) components

$$ {}^{(3)}R_{rr} = -\frac{1}{rB}\frac{\mathrm{d}B}{\mathrm{d}r}, \tag{13.14} $$

$$ {}^{(3)}R_{\theta\theta} = -1 + \frac{1}{B} - \frac{r}{2B^2}\frac{\mathrm{d}B}{\mathrm{d}r}, \tag{13.15} $$

$$ {}^{(3)}R_{\phi\phi} = \sin^2\theta\, {}^{(3)}R_{\theta\theta}. \tag{13.16} $$

Finally, we can form the 3D Ricci scalar

$$ \begin{aligned} {}^{(3)}R = \gamma^{ij\,(3)}R_{ij} &= -\frac{2}{r^2}\left(1 - \frac{1}{B} + \frac{r}{B^2}\frac{\mathrm{d}B}{\mathrm{d}r}\right) \\ &= -\frac{2}{r^2}\left[1 - \frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r}{B}\right)\right]. \end{aligned} \tag{13.17} $$

If the 3D space is homogeneous, the Ricci scalar cannot depend on position. Writing

$$ {}^{(3)}R = -6K, \tag{13.18} $$

where K is a constant (and the prefactor is for later convenience), we require

$$ \begin{aligned} 1 - \frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r}{B}\right) &= 3Kr^2 \\ \implies \frac{r}{B} &= A + r - Kr^3 \\ \implies B &= \frac{1}{A/r + (1 - Kr^2)}. \end{aligned} \tag{13.19} $$

Here, A is an integration constant, but this must vanish for the space to be homogeneous. To see this, note that we require all curvature invariants to be constant; for example, considering

$$ R_{ij}R^{ij} = 12K^2 + \frac{3A^2}{2r^6}, \tag{13.20} $$

we find that this is only homogeneous for $A = 0$. The final form of the 3D line element for isotropic and homogeneous spaces is thus

$$ \boxed{\mathrm{d}\sigma^2 = \frac{\mathrm{d}r^2}{(1 - Kr^2)} + r^2\,\mathrm{d}\Omega^2\,.} \tag{13.21} $$

The spacetime line element is then of *Robertson–Walker* form

$$ \boxed{\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - a^2(t)\left[\frac{\mathrm{d}r^2}{(1 - Kr^2)} + r^2\,\mathrm{d}\Omega^2\right].} \tag{13.22} $$

13.2.0.1 Maximally-Symmetric Spaces

Substituting for $B(r) = (1 - Kr^2)^{-1}$ in the components of the 3D Ricci tensor, we have

$$ {}^{(3)}R_{rr} = \frac{-2K}{(1 - Kr^2)} = -2K\gamma_{rr}, \tag{13.23} $$

$$ {}^{(3)}R_{\theta\theta} = -2Kr^2 = -2K\gamma_{\theta\theta}, \tag{13.24} $$

$$ {}^{(3)}R_{\phi\phi} = -2Kr^2 \sin^2\theta = -2K\gamma_{\phi\phi}. \tag{13.25} $$

It follows that

$$ {}^{(3)}R_{ij} = -2K\gamma_{ij}, \tag{13.26} $$

which is a tensor equation valid in any coordinates x^i. Moreover, evaluating the components of the 3D Riemann tensor, we find

$$ {}^{(3)}R_{ijkl} = K(\gamma_{ik}\gamma_{jl} - \gamma_{il}\gamma_{jk}), \tag{13.27} $$

It is easy to check that the tensor on the right has all the symmetries required of the Riemann tensor.

Generally, manifolds that have a Riemann tensor of the form

$$ R_{abcd} = K(g_{ac}g_{bd} - g_{ad}g_{bc}), \tag{13.28} $$

are known as *maximally-symmetric spaces* and they have the same number of symmetries as Euclidean space of the same dimension. The quantity K has to be constant by virtue of the Bianchi identity. For example, in ND the Ricci tensor and scalar in a maximally-symmetric space are

$$ R_{ab} = K(N - 1)g_{ab}, \tag{13.29} $$

$$ R = KN(N - 1) \tag{13.30} $$

so the contracted Bianchi identity gives

$$ \nabla^a\left(R_{ab} - \frac{1}{2}g_{ab}R\right) = 0 $$

$$ \implies \frac{1}{2}(N - 1)(N - 2)\nabla^a(Kg_{ab}) = 0, \tag{13.31} $$

which implies $\nabla_b K = 0$ (so K is constant) for $N > 2$. The space with the line element in Eq. (13.21) is therefore a 3D maximally-symmetric space. Isotropy is manifest, but homogeneity is rather hidden by having made an explicit choice of origin.

13.2.1 Geometry of the 3D Spaces

The properties of the 3D maximally-symmetric spaces with line element

$$ \mathrm{d}\sigma^2 = \frac{\mathrm{d}r^2}{(1 - Kr^2)} + r^2\,\mathrm{d}\Omega^2 \tag{13.32} $$

depend on whether K is positive, negative or zero.

- $\boldsymbol{K = 0}$: With $K = 0$, we have Euclidean space in spherical-polar coordinates:

$$\mathrm{d}\sigma^2 = \mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2\,. \tag{13.33}$$

- $\boldsymbol{K > 0}$: For $K > 0$, it is convenient to switch from the dimensionfull r coordinate to another dimensionfull coordinate χ, where

$$r = \frac{1}{\sqrt{K}}\sin\Big(\sqrt{K}\chi\Big) \equiv S_K(\chi). \tag{13.34}$$

Using $\mathrm{d}r = \cos\Big(\sqrt{K}\chi\Big)\,\mathrm{d}\chi$, the line element becomes

$$\mathrm{d}\sigma^2 = \mathrm{d}\chi^2 + S_K^2(\chi)\,\mathrm{d}\Omega^2\,. \tag{13.35}$$

This is the same as the line element on the surface of a 3-sphere of radius $1/\sqrt{K}$ embedded in 4D Euclidean space, which we can see as follows.

Let (w, x, y, z) be Cartesian coordinates in $\mathbb{R}^4$, so that

$$\mathrm{d}s^2 = \mathrm{d}w^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2\,. \tag{13.36}$$

The 3-sphere has

$$w^2 + x^2 + y^2 + z^2 = \frac{1}{K}, \tag{13.37}$$

and can be parameterised by

$$w = \frac{1}{\sqrt{K}}\cos\Big(\sqrt{K}\chi\Big), \tag{13.38}$$
$$x = S_K(\chi)\sin\theta\cos\phi, \tag{13.39}$$
$$y = S_K(\chi)\sin\theta\sin\phi, \tag{13.40}$$
$$z = S_K(\chi)\cos\theta. \tag{13.41}$$

Here, θ and ϕ are the usual angular coordinates and $0 \le \sqrt{K}\chi \le \pi$. The induced metric on the 3-sphere follows from

$$\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 = \cos^2(\sqrt{K}\chi)\,\mathrm{d}\chi^2 + S_K^2(\chi)\,\mathrm{d}\Omega^2\,, \tag{13.42}$$

and

$$\mathrm{d}w^2 = \sin^2(\sqrt{K}\chi)\,\mathrm{d}\chi^2\,; \tag{13.43}$$

adding these together gives Eq. (13.35). It follows that 3D maximally-symmetric space with $K > 0$ is compact (or *closed*). The area of a 2-sphere $\chi = \text{const.}$ is

$$A = 4\pi S_K^2(\chi), \tag{13.44}$$

and grows from zero at $\chi = 0$ to a maximum of $4\pi/K$ at $\sqrt{K}\chi = \pi/2$, before shrinking back to zero as $\sqrt{K}\chi \to \pi$. The space has finite volume

$$V = 4\pi\int_0^{\pi/\sqrt{K}} S_K^2(\chi)\,\mathrm{d}\chi = \frac{2\pi^2}{K^{3/2}}. \tag{13.45}$$

- $\boldsymbol{K < 0}$: For $K < 0$, we write

$$r = \frac{1}{\sqrt{|K|}} \sinh\Big(\sqrt{|K|}\chi\Big) \equiv S_K(\chi), \tag{13.46}$$

which defines $S_K(\chi)$ for $K < 0$. Now, $\mathrm{d}r = \cosh\Big(\sqrt{|K|}\chi\Big)\,\mathrm{d}\chi$, and the line element becomes

$$\mathrm{d}\sigma^2 = \mathrm{d}\chi^2 + S_K^2(\chi)\,\mathrm{d}\Omega^2\,. \tag{13.47}$$

Up to a sign, this is the same as the line element on the surface of spacelike hyperboloid,

$$w^2 - x^2 - y^2 - z^2 = \frac{1}{|K|}, \tag{13.48}$$

embedded in Minkowski space

$$\mathrm{d}s^2 = \mathrm{d}w^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2\,. \tag{13.49}$$

The hyperboloid can be parameterised by

$$w = \frac{1}{\sqrt{|K|}} \cosh\Big(\sqrt{K}\chi\Big), \tag{13.50}$$
$$x = S_K(\chi)\sin\theta\cos\phi, \tag{13.51}$$
$$y = S_K(\chi)\sin\theta\sin\phi, \tag{13.52}$$
$$z = S_K(\chi)\cos\theta, \tag{13.53}$$

where, now, $0 \le \sqrt{|K|}\chi < \infty$. The induced metric on the hyperboloid follows from

$$\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 = \cosh^2(\sqrt{|K|}\chi)\,\mathrm{d}\chi^2 + S_K^2(\chi)\,\mathrm{d}\Omega^2\,, \tag{13.54}$$

and

$$\mathrm{d}w^2 = \sinh^2(\sqrt{|K|}\chi)\,\mathrm{d}\chi^2\,, \tag{13.55}$$

which combine to give Eq. (13.47). The 3D maximally-symmetric space with $K < 0$ is infinite (or *open*) and has infinite volume.

In terms of the radial coordinate χ, we can write the spacetime line element for all three cases as

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - a^2(t)\Big[\mathrm{d}\chi^2 + S_K^2(\chi)\,\mathrm{d}\Omega^2\Big], \tag{13.56}$$

where

$$S_K(\chi) \equiv \begin{cases} \sin\Big(\sqrt{K}\chi\Big) & \text{for} K > 0, \\ \chi & \text{for} K = 0, \\ \sinh\Big(\sqrt{|K|}\chi\Big)/\sqrt{|K|} & \text{for} K = 0. \end{cases} \tag{13.57}$$

Note that $S_K(\chi)$ is continuous in K at $K = 0$ as, e.g., $\lim_{K\to 0} S_K(\chi) = \chi$ from above and below. Note further that (χ, θ, ϕ) are still comoving coordinates as χ is related to r through a time-independent transformation.

13.3 An Expanding Universe

Consider the fundamental observer at $\chi = 0$ and a neighbouring one at $\Delta\chi$. The proper distance between these observers is $l(t) = a(t)\Delta\chi$. If the scale factor depends on time, this distance will also change as

$$\frac{1}{l(t)}\frac{\mathrm{d}l(t)}{\mathrm{d}t} = \frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t} \equiv H(t) \tag{13.58}$$

which defines the *Hubble parameter* $H(t)$. For $H > 0$, the observer at the origin sees all fundamental observers (and hence the matter in the Universe) moving away isotropically with the same fractional rate. The same is true for any other fundamental observer by isotropy.[1] We know that our Universe is expanding as the light from distant galaxies is observed to be redshifted (see Subsection 13.3.1). The current value of the Hubble parameter is denoted by H_0; its value is around $68\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

13.3.1 Cosmological Redshift

Consider a photon emitted by a fundamental observer at coordinates $(t_E, 0, 0, 0)$, which is received later by a fundamental observer at $(t_R, \chi_R, \theta_R, \phi_R)$. The radial (symmetry!) path of the photon is

$$x^\mu(\lambda) = (t(\lambda), \chi(\lambda), \theta_R, \phi_R), \tag{13.59}$$

where λ is an affine parameter. The 4-momentum of the photon is

$$p^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} = \Big(p^0, p^1, 0, 0\Big), \tag{13.60}$$

and lowering the index with the (diagonal) metric gives

$$p_\mu = (p_0, p_1, 0, 0), \tag{13.61}$$

where, numerically, $p_0 = c^2 p^0$ and $p_1 = -a^2 p^1$. The photon travels on a null geodesic, parallel transporting p^μ. It is convenient to consider the geodesic equation (5.70) in the form (see Section 5.4)

$$\frac{\mathrm{d}p_\mu}{\mathrm{d}\lambda} = \frac{1}{2}\frac{\partial g_{\nu\rho}}{\partial x^\mu} p^\nu p^\rho. \tag{13.62}$$

We have

$$\begin{aligned} \frac{\mathrm{d}p_1}{\mathrm{d}\lambda} &= \frac{1}{2}\left(\frac{\partial g_{00}}{\partial \chi} p^0 p^0 + \frac{\partial g_{11}}{\partial \chi} p^1 p^1\right) \\ \implies p_1 &= \text{const.} \end{aligned} \tag{13.63}$$

where we have used $g_{00} = c^2$ and $g_{11} = -a^2$, both of which are independent of χ.

The energy of a photon of 4-momentum p^μ as measured by a fundamental observer with 4-velocity u^μ is

$$E = g_{\mu\nu} = p^\mu p^\nu. \tag{13.64}$$

[1] An often-quoted analogy is the surface of a balloon covered in spots that is being inflated. Every spot sees all others moving away isotropically.

As discussed earlier, $u^\mu = \delta^\mu_0$, so that

$$E = p_\mu u^\mu = p_0. \tag{13.65}$$

We can relate p_0 to the conserved p_1 using the null condition

$$g^{\mu\nu} p_\mu p_\nu = 0 \quad \Longrightarrow \quad c^{-2}(p_0)^2 - a^{-2}(p_1)^2 = 0. \tag{13.66}$$

It follows that $p_0 = cp_1/a$ so

$$Ea(t) = \text{const.} \tag{13.67}$$

Finally, we can compute the redshift of the photon:

$$1 + z \equiv \frac{\lambda_R}{\lambda_E} = \frac{E_E}{E_R} = \frac{a(t_R)}{a(t_E)}, \tag{13.68}$$

where λ_E and λ_R are the emitted and received wavelengths, respectively, and E_E and E_R are the associated energies.

We see that in an expanding universe, light from distant galaxies (assumed to be moving almost as fundamental observers) will be redshifted. The first evidence for this effect was obtained by Edwin Hubble in 1929. Note that we performed the calculation of the redshift for the simple case that one of the observers was at the origin. However, by homogeneity, the same result will hold for any pair of fundamental observers separated by the same (shortest) proper distance χ_R.

13.4 Cosmological Field Equations

The Robertson–Walker metric contains a single function of time $a(t)$ whose evolution is determined by Eq. (9.49) with the Einstein field equations:

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu} = -\kappa T_{\mu\nu}, \tag{13.69}$$

where, recall, $\kappa = 8\pi G/c^4$. We have included the cosmological constant term $\Lambda g_{\mu\nu}$, introduced in Section 9.4, as this can be important on cosmological scales. For the energy–momentum tensor, isotropy demands that we consider the ideal fluid form from Eq. (9.8)

$$T^{\mu\nu} = \left(\rho + \frac{p}{c^2}\right) u^\mu u^\nu - p g^{\mu\nu}, \tag{13.70}$$

since this is isotropic in the instantaneous rest-frame defined by u^μ. We must take u^μ to coincide with the 4-velocity of the fundamental observers. Homogeneity demands that the proper energy density ρc^2 and pressure p are functions of time only.

13.4.1 Friedmann Equations

Our starting point is the spacetime line element, which we write in the form

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - a^2(t)\gamma_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j\,, \tag{13.71}$$

where γ_{ij} is the metric of a 3D maximally-symmetric space and depends only on the comoving coordinates x^i. The components of the spacetime metric and its inverse are

$$\begin{aligned} g_{00} &= c^2, & g^{00} &= \frac{1}{c^2}, \\ g_{ij} &= -a^2\gamma_{ij}, & g^{ij} &= -\tfrac{1}{a^2}\gamma^{ij}, \end{aligned} \tag{13.72}$$

where γ^{ij} is the inverse of γ_{ij}. The metric connection has the following (independent) non-zero components:

$$\Gamma^0_{ij} = \frac{\dot{a}a}{c^2}\gamma_{ij}, \quad \Gamma^i_{0j} = \frac{\dot{a}}{a}\delta^i_j, \quad \Gamma^i_{jk} = {}^{(3)}\Gamma^i_{jk}, \tag{13.73}$$

where overdots denote differentiation with respect to cosmic time t and ${}^{(3)}\Gamma^i_{jk}$ are the metric connection coefficients of the 3D metric γ_{ij}. (The ${}^{(3)}\Gamma^i_{jk}$ were given in Eq. (13.8) in the (r, θ, ϕ) coordinates; however, we shall not require their specific form here.)

Connection Coefficients for the Robertson–Walker Metric

- The calculation of the connection coefficients follows from

$$\Gamma^\mu_{\nu\rho} = \frac{1}{2}g^{\mu\sigma}(\partial_\nu g_{\sigma\rho} + \partial_\rho g_{\sigma\nu} - \partial_\sigma g_{\nu\rho}). \tag{13.74}$$

- For example,

$$\begin{aligned} \Gamma^0_{ij} &= \frac{1}{2}g^{0\sigma}(\partial_i g_{\sigma j} + \partial_j g_{\sigma i} - \partial_\sigma g_{ij}) \\ &= \frac{1}{2}g^{00}(\partial_i g_{0j} + \partial_j g_{0i} - \partial_0 g_{ij}) \\ &= \frac{1}{2c^2}\frac{\mathrm{d}a^2}{\mathrm{d}t}\gamma_{ij} \\ &= \frac{\dot{a}a}{c^2}\gamma_{ij}. \end{aligned} \tag{13.75}$$

- We also have

$$\begin{aligned} \Gamma^i_{0j} &= \frac{1}{2}g^{i\sigma}(\partial_0 g_{\sigma j} + \partial_j g_{\sigma 0} - \partial_\sigma g_{0j}) \\ &= \frac{1}{2}g^{ik}(\partial_0 g_{kj} + \partial_j g_{k0} - \partial_k g_{0j}) \\ &= \frac{1}{2a^2}\frac{\mathrm{d}a^2}{\mathrm{d}t}\gamma_{kj} \\ &= \frac{\dot{a}}{a}\delta^i_j. \end{aligned} \tag{13.76}$$

- The only other non-zero connection coefficient is

$$\begin{aligned} \Gamma^i_{jk} &= \frac{1}{2}g^{i\sigma}(\partial_j g_{\sigma k} + \partial_k g_{\sigma j} - \partial_\sigma g_{jk}) \\ &= \frac{1}{2}g^{il}(\partial_j g_{lk} + \partial_k g_{lj} - \partial_l g_{jk}) \\ &= \frac{1}{2}\gamma^{il}(\partial_j \gamma_{lk} + \partial_k \gamma_{lj} - \partial_l \gamma_{jk}) \\ &= {}^{(3)}\Gamma^i_{jk}. \end{aligned} \tag{13.77}$$

We also require the spacetime Ricci tensor, given by

$$R_{\mu\nu} = -\partial_\rho \Gamma^\rho_{\mu\nu} + \partial_\mu \Gamma^\rho_{\rho\nu} + \Gamma^\sigma_{\rho\nu}\Gamma^\rho_{\mu\sigma} - \Gamma^\sigma_{\mu\nu}\Gamma^\rho_{\rho\sigma}. \tag{13.78}$$

A useful intermediate result, which follows from $\Gamma^\rho_{\rho\mu} = \Gamma^0_{0\mu} + \Gamma^i_{i\mu}$, is

$$\Gamma^\rho_{\rho 0} = 3\dot{a}/a \quad \text{and} \quad \Gamma^\rho_{\rho i} = {}^{(3)}\Gamma^j_{ji}. \tag{13.79}$$

By isotropy, $R_{0i} = 0$; the remaining components of the Ricci tensor are

$$R_{00} = 3\frac{\ddot{a}}{a}, \tag{13.80}$$

$$R_{ij} = -\frac{1}{c^2}\Big(\ddot{a}a + 2\dot{a}^2 + 2Kc^2\Big)\gamma_{ij}. \tag{13.81}$$

Ricci Tensor for the Robertson–Walker Metric

- For R_{00} we have

$$\begin{aligned} R_{00} &= -\partial_\rho \underbrace{\Gamma^\rho_{00}}_{0} + \partial_0 \underbrace{\Gamma^\rho_{\rho 0}}_{3\dot{a}/a} + \Gamma^\sigma_{\rho 0}\Gamma^\rho_{0\sigma} - \underbrace{\Gamma^\sigma_{00}}_{0}\Gamma^\rho_{\rho\sigma} \\ &= 3\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\dot{a}}{a}\right) + \Gamma^i_{j0}\Gamma^j_{0i} \\ &= 3\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\dot{a}}{a}\right) + \left(\frac{\dot{a}}{a}\right)^2 \delta^i_j\delta^j_i \\ &= 3\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\dot{a}}{a}\right) + \left(\frac{\dot{a}}{a}\right)^2\right] \\ &= 3\frac{\ddot{a}}{a}. \end{aligned} \tag{13.82}$$

- For R_{ij}, we have

$$R_{ij} = -\partial_\rho \Gamma^\rho_{ij} + \partial_i \Gamma^\rho_{\rho j} + \Gamma^\sigma_{\rho j}\Gamma^\rho_{i\sigma} - \Gamma^\sigma_{ij}\Gamma^\rho_{\rho\sigma}. \tag{13.83}$$

  - The first term is

$$\begin{aligned} -\partial_\rho \Gamma^\rho_{ij} &= -\partial_0 \Gamma^0_{ij} - \partial_k \Gamma^k_{ij} \\ &= -\frac{1}{c^2}\frac{\mathrm{d}(\dot{a}a}{\mathrm{d}t}\gamma_{ij} - \partial_k {}^{(3)}\Gamma^k_{ij}. \end{aligned} \tag{13.84}$$

  - The second term is

$$\partial_i \Gamma^\rho_{\rho j} = \partial_i {}^{(3)}\Gamma^k_{kj}. \tag{13.85}$$

  - For the third, we have

$$\begin{aligned} \Gamma^\sigma_{\rho j}\Gamma^\rho_{i\sigma} &= \Gamma^0_{\rho j}\Gamma^\rho_{i0} + \Gamma^k_{\rho j}\Gamma^\rho_{ik} \\ &= \Gamma^0_{kj}\Gamma^k_{i0} + \Gamma^k_{0j}\Gamma^0_{ik} + \Gamma^k_{lj}\Gamma^l_{ik} \\ &= \frac{\dot{a}a}{c^2}\gamma_{kj}\frac{\dot{a}}{a}\delta^k_i + \frac{\dot{a}}{a}\delta^k_j\frac{\dot{a}a}{c^2}\gamma_{ik} + {}^{(3)}\Gamma^k_{lj}\,{}^{(3)}\Gamma^l_{ik} \\ &= 2\frac{\dot{a}^2}{c^2}\gamma_{ij} + {}^{(3)}\Gamma^k_{lj}\,{}^{(3)}\Gamma^l_{ik}. \end{aligned} \tag{13.86}$$

– The final term is

$$-\Gamma^{\sigma}_{ij}\Gamma^{\rho}_{\rho\sigma} = -\Gamma^{0}_{ij}\Gamma^{\rho}_{\rho 0} - \Gamma^{k}_{ij}\Gamma^{\rho}_{\rho k}$$
$$= -3\frac{\dot{a}^2}{c^2}\gamma_{ij} - {}^{(3)}\Gamma^{k}_{ij}\,{}^{(3)}\Gamma^{l}_{lk}. \tag{13.87}$$

- Adding these four terms, the parts involving ${}^{(3)}\Gamma^{i}_{jk}$ assemble to give the 3D Ricci tensor, leaving

$$R_{ij} = -\frac{1}{c^2}\Big(\ddot{a}a + 2\dot{a}^2\Big)\gamma_{ij} + {}^{(3)}R_{ij}$$
$$= -\frac{1}{c^2}\Big(\ddot{a}a + 2\dot{a}^2 + 2Kc^2\Big)\gamma_{ij}, \tag{13.88}$$

where we used the result ${}^{(3)}R_{ij} = -2Kc^2\gamma_{ij}$ for the maximally-symmetric 3D space.

We now write the Einstein field equations from Eq. (13.69) as

$$R_{\mu\nu} = -\kappa\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T\right) + \Lambda g_{\mu\nu}, \tag{13.89}$$

where $T \equiv g_{\mu\nu}T^{\mu\nu}$ is the trace of the energy–momentum tensor. For the ideal fluid, we have

$$T = g_{\mu\nu}\left(\rho + \frac{p}{c^2}\right)u^{\mu}u^{\nu} - pg_{\mu\nu}g^{\mu\nu}$$
$$= \rho c^2 - 3p, \tag{13.90}$$

so that

$$T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T = \left(\rho + \frac{p}{c^2}\right)u_{\mu}u_{\nu} - \frac{1}{2}\Big(\rho c^2 - p\Big)g_{\mu\nu}. \tag{13.91}$$

The 4-velocity has components $u^{\mu} = \delta^{\mu}_{0}$, so that

$$u_{\mu} = g_{\mu 0} = c^2\delta_{\mu 0} \tag{13.92}$$

It follows that the non-zero Einstein equations reduce to

$$R_{00} = -4\pi G\left(\rho + \frac{3p}{c^2}\right) + \Lambda c^2, \tag{13.93}$$

$$R_{ij} = -\frac{4\pi G}{c^2}\left(\rho - \frac{p}{c^2}\right)a^2\gamma_{ij} - \Lambda a^2\gamma_{ij}. \tag{13.94}$$

Using the explicit form for the components of the Ricci tensor that we worked out above, the first equation reduces to

$$\boxed{\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{1}{3}\Lambda c^2,} \tag{13.95}$$

and the second to

$$\frac{\ddot{a}}{a} + 2\left(\frac{\dot{a}}{a}\right)^2 + \frac{2Kc^2}{a^2} = 4\pi G\left(\rho - \frac{p}{c^2}\right) + \Lambda c^2. \tag{13.96}$$

Eliminating $\ddot{a}/a$, this reduces to

$$\boxed{\left(\frac{\dot{a}}{a}\right)^2 + \frac{Kc^2}{a^2} = \frac{8\pi G}{3}\rho + \frac{1}{3}\Lambda c^2.} \tag{13.97}$$

Equations (13.95) and (13.97) are known as the *Friedmann equations*.

13.4.2 Conservation of the Energy–Momentum Tensor

Recall from Eq. (9.9) that conservation of energy and momentum is described by

$$T_\mu T^{\mu\nu} = 0. \tag{13.98}$$

For the energy–momentum tensor

$$T_{\mu\nu} = \left(\rho + \frac{p}{c^2}\right) u^\mu u^\nu - p g^{\mu\nu}, \tag{13.99}$$

we have

$$0 = \nabla_\mu T^{\mu\nu} = u^\nu u^\mu \nabla_\mu \left(\rho + \frac{p}{c^2}\right) + \left(\rho + \frac{p}{c^2}\right)(u^\nu \nabla_\mu u^\mu + u^\mu \nabla_\mu u^\nu) - \nabla^\nu p. \tag{13.100}$$

Since $u^\mu = \delta^\mu_0$, we have $u^\mu \nabla_\mu \rho = \dot{\rho}$. Furthermore, with $u^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$, we have

$$u^\mu \nabla_\mu u^\nu = \frac{\mathrm{D}u^\nu}{\mathrm{D}\tau}, \tag{13.101}$$

i.e., the 4-acceleration of the fluid. However, as the fluid moves along geodesics, this vanishes and $u^\mu \nabla_\mu u^\nu = 0$. Finally, we have

$$\begin{aligned} \nabla_\mu u^\mu &= \partial_\mu u^\mu + \Gamma^\mu_{\mu\nu} u^\nu \\ &= \Gamma^\mu_{\mu 0} = 3\dot{a}/a. \end{aligned} \tag{13.102}$$

It follows that Eq. (13.100) becomes

$$u^\nu \left[\dot{\rho} + \frac{\dot{p}}{c^2} + 3\frac{\dot{a}}{a}\left(\rho + \frac{p}{c^2}\right)\right] - \nabla^\nu p = 0. \tag{13.103}$$

The component along u^ν is obtained by contracting with u_ν, which gives

$$\boxed{\dot{\rho} + 3\frac{\dot{a}}{a}\left(\rho + \frac{p}{c^2}\right) = 0;} \tag{13.104}$$

the projection perpendicular to u^ν vanishes identically since p is homogeneous. Eq. (13.104) expresses conservation of energy. Note that Eq. (13.104) is implied by the two Friedmann equations (because of the contracted Bianchi identity), and so is not independent.

For example, for dust ($p = 0$), we have

$$\frac{\dot{\rho}}{\rho} = -3\frac{\dot{a}}{a} \quad \Longrightarrow \quad \rho a^3 = \text{const.} \tag{13.105}$$

Since the proper volume of a given set of fluid particles scales as a^3, the proper number density of particles and hence energy density goes as a^{-3}.

In contrast, for radiation ($p = \rho c^2/3$), we have

$$\frac{\dot{\rho}}{\rho} = -4\frac{\dot{a}}{a} \quad \Longrightarrow \quad \rho a^4 = \text{const.} \tag{13.106}$$

In this case, the energy density falls more quickly as the Universe expands due to the pV work done by the fluid.

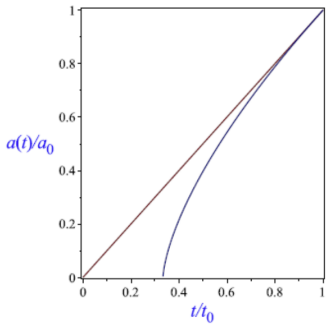

Fig. 13.2: $H \equiv \dot{a}/a$ is the Hubble parameter with value H_0 today. If we take $t = t_0$ today, in an expanding universe $\ddot{a} < 0$ implies age of universe $< a(t_0)/\dot{a}(t_0) = 1/H_0$. We see that the age of the Universe is less than the *Hubble time* $1/H_0$.

13.5 Cosmological Models

We now briefly consider some properties of the evolution of the scale factor $a(t)$ that is implied by the Friedmann equations. We have

$$H^2 + \frac{Kc^2}{a^2} = \frac{8\pi G}{3}\rho + \frac{1}{3}\Lambda c^2, \tag{13.107}$$

where, recall, $H \equiv \dot{a}/a$ is the Hubble parameter with value H_0 today. If we define a *critical density*

$$\rho_{\rm crit} \equiv \frac{3H^2}{8\pi G}, \tag{13.108}$$

then, for $\Lambda = 0$,

$$\begin{array}{lcllcll} \rho & > & \rho_{\rm crit} & \Longrightarrow & K & > & 0 \quad \text{(closed)} \\ \rho & = & \rho_{\rm crit} & \Longrightarrow & K & = & 0 \quad \text{(flat)} \\ \rho & < & \rho_{\rm crit} & \Longrightarrow & K & < & 0 \quad \text{(open)}. \end{array} \tag{13.109}$$

$\Lambda = 0$ Consider $\Lambda = 0$, and "ordinary" matter with $\rho > 0$ and $p \geq 0$. Then, since

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{1}{3}\Lambda c^2, \tag{13.110}$$

we have $\ddot{a} < 0$. This implies $a = 0$ at some time in the past, i.e., the Universe emerged from a singularity in the past (the *big bang*). If we take $t = t_0$ today, in an expanding universe $\ddot{a} < 0$ implies

$$\text{age of universe} < \frac{a(t_0)}{\dot{a}(t_0)} = \frac{1}{H_0}. \tag{13.111}$$

We see that the age of the Universe is less than the *Hubble time* $1/H_0$ (see Fig. 13.2). For a current value of $H_0 = 68 \text{km}\,\text{s}^{-1}\,\text{Mpc}^{-1} = (14\text{Gyr})^{-1}$, the age of the Universe would be less than 14Gyr if Λ were zero.[2]

[2]The age of the Universe is known actually to be very close to 14Gyr; this is due to the late-time acceleration of our Universe (which increases the age for a given Hubble parameter today) compensating for $\ddot{a} < 0$ at earlier times.

In a flat or open universe ($K \leq 0$), we have

$$H^2 = \frac{8\pi G}{3}\rho - \frac{Kc^2}{a^2} > 0, \tag{13.112}$$

so that the expansion never stops. In contrast, in a closed universe there exists a maximum scale factor $a_{\max}$ where

$$\frac{Kc^2}{a_{\max}^2} = \frac{8\pi G}{3}\rho \tag{13.113}$$

and $\dot{a} = 0$. (Here, we have used that ρ falls at least as fast as a^{-3} in an expanding universe with $p \geq 0$.) Since $\ddot{a} < 0$, the universe subsequently contracts back to $a = 0$ (a future singularity). Important special cases for $\Lambda = 0$ include models with $K = 0$ and $p = 0$ (the *Einstein–de Sitter* model), which is a good model of our Universe for most of its history. In this case, $\rho \propto a^{-3}$ and so

$$\left(\frac{\dot{a}}{a}\right)^2 \propto \frac{1}{a^3} \quad \Longrightarrow \quad a \propto t^{2/3} \quad \text{and} \quad H = \frac{2}{3t}, \tag{13.114}$$

where we have taken $t = 0$ at $a = 0$.

As $\rho \propto a^{-4}$ for radiation, at sufficiently early times the energy density of our Universe is dominated by radiation rather than pressure-free (dark) matter. For radiation domination,

$$\left(\frac{\dot{a}}{a}\right)^2 \propto \frac{1}{a^4} \quad \Longrightarrow \quad a \propto t^{1/2} \quad \text{and} \quad H = \frac{1}{2t}. \tag{13.115}$$

$\Lambda > 0$ Models with sufficiently large cosmological constant can undergo accelerated expansion, $\ddot{a} > 0$ (see Eq. (13.110)). A range of cosmological observations show that our Universe is accelerating at the current time, presumably under the action of something like a cosmological constant. Consider a flat universe ($K = 0$) with $\Lambda > 0$. This will expand forever, and at late times

$$H^2 = \frac{8\pi G}{3}\rho + \frac{1}{3}\Lambda c^2 \to \frac{1}{3}\Lambda c^2. \tag{13.116}$$

The Hubble parameter therefore tends to a constant value and, asymptotically,

$$a \propto \exp\left(\sqrt{\frac{\Lambda c^2}{3}}t\right) \tag{13.117}$$

The resulting solution is known as the *de Sitter* universe.

Appendix A

Appendix: Euler-Lagrange Equations

A.1 Functionals

Consider the following definite integral involving a real function $y(x)$,

$$G = \int_\alpha^\beta \left[(y'(x))^2 - (y(x))^2 \right] \mathrm{d}x \,. \tag{A.1}$$

The real number G is independent of x but depends on $y(x)$. This is a simple example of a *functional* and we denote it by $G[y]$. A real function of many variables $\{y_k; k = 1, 2, \ldots, N\}$ takes $\{y_k\}$ and gives a real number as output,

$$f : \{y_k\} \to f(\{y_k\}) \in \mathbb{R}. \tag{A.2}$$

A real functional is a generalization to a continuous infinity of variables $\{y(x); x \in \mathbb{R}\}$. It takes a function $y(x)$ and gives a real number as output,

$$G : y(x) \to G[y] \in \mathbb{R}. \tag{A.3}$$

In Eq. (A.1) above the integrand of $G[y]$ *implicitly* depends on x through y and its derivatives, but it may also *explicitly* depend on x,

$$G[y] = \int_\alpha^\beta f(y, y', y'', \cdots; x) \, \mathrm{d}x \,. \tag{A.4}$$

In general, there may be more dependent variables $\{y_i\}$ and a multiple integral over a number of independent variables $\{y_i\}$. We shall usually be concerned with functionals of the form,

$$G[y] = \int_\alpha^\beta f(y, y'; x) \, \mathrm{d}x \,. \tag{A.5}$$

The calculus of variations extends the calculus of functions to functionals It aims to answer questions such as: what functions $y(x)$ extremise the functional $G[y]$?

- It will usually be obvious from the problem whether a given extremum is a maximum, a minimum or something else – there's no equivalent of the usual criteria for functions (or at least one that's practical to use).
- We must also keep in mind that, as with ordinary calculus, an extremum we find may be only a *local* extremum and not a *global* extremum.

Functionals are useful because many problems can be formulated as a *variational principle*, the extremisation of some functional. E.g. a chain suspended between two fixed points hangs in equilibrium such that its total potential energy is minimized and an extension of this idea (incorporating chemical potential energy) can be applied to chemical reactions. An important example is *Hamilton's principle of least action* in mechanics.

A.2 Functional Derivatives

Consider the effect of changing a function $y(x)$ to a nearby function $y(x) + \delta y(x)$. The variation of G is defined by

$$\begin{aligned}\delta G &= g[y+\delta y] - G[y] \\ &= \int_\alpha^\beta f(y+\delta y, y'+(\delta y)'; x)\,\mathrm{d}x - \int_\alpha^\beta f(y, y'; x) \\ &= \int_\alpha^\beta \left[\delta y \frac{\partial f}{\partial y} + (\delta y)' \frac{\partial f}{\partial y'}\right]\mathrm{d}x + \cdots \\ &= \left[\delta y \frac{\partial f}{\partial y'}\right] + \int_\alpha^\beta \delta y \left[\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\right]\mathrm{d}x + \cdots . \end{aligned} \tag{A.6}$$

where we have omitted terms of order $(\delta y)^2$. If the boundary term is zero (e.g. if y is fixed on the boundary) then

$$\delta G = \int_\alpha^\beta \delta y(x) \frac{\delta G}{\delta y(x)}\,\mathrm{d}x + \cdots . \tag{A.7}$$

Here we have *defined* the *functional derivative* of G with respect to the function y to be,

$$\frac{\delta G}{\delta y(x)} \equiv \frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right). \tag{A.8}$$

Note that this depends on x. (Compare with the variation of a function $f(\{y_i\})$: $\delta f = \sum_i \delta y_i \;\partial f/\partial y_i$.) The functional is stationary when $\delta G/\delta y(x) = 0$, i.e.

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) = \frac{\partial f}{\partial y}.} \tag{A.9}$$

This is the **Euler-Lagrange (EL) equation**.

Note on notation: $\partial f/\partial y'$ may look strange - it seems impossible for y' to change if y doesn't. Here $\partial/\partial y$ and $\partial/\partial y'$ are just formal derivatives: pretend that y and y' are unconnected. By contrast, $\mathrm{d}/\mathrm{d}x$ is the usual full derivative with respect to x.

A.3 First Integral

In the example above we reduced the second-order EL equation to a first-order equation, $y' = const.$, a "*first integral*" of the EL equation. We found $\partial f/\partial y = 0$ and so the EL equation gave $\partial f/\partial y' = const.$ We can also reduce the EL equation to a first integral if $\partial f/\partial x = 0$, i.e. $f(y, y'; x)$ has no explicit dependence on x. Indeed, from the chain rule we have,

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\partial f}{\partial x} + y'\frac{\partial f}{\partial y} + y''\frac{\partial f}{\partial y'}. \tag{A.10}$$

Using the EL equation gives,

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}x} &= \frac{\partial f}{\partial x} + y'\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) + y''\frac{\partial f}{\partial y'} \\ &= \frac{\partial f}{\partial x} + \frac{\mathrm{d}}{\mathrm{d}x}\left(y'\frac{\partial f}{\partial y'}\right), \end{aligned} \tag{A.11}$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(f - y'\frac{\partial f}{\partial y'}\right) = \frac{\partial f}{\partial x}. \tag{A.12}$$

When $\partial f/\partial x = 0$,

$$\boxed{y'\frac{\partial f}{\partial y'} - f = \text{const.}} \tag{A.13}$$

A.4 Hamilton's Principle

Lagrange and Hamilton developed a powerful reformulation of Newtonian mechanics in terms of a "principle of least action" based on energy rather than force. The time evolution of a system is viewed as the motion of a point in a multi-dimensional *configuration space* described by some *generalised coordinates* $\{q_i\}$. Examples:

- A system of n particles ($3n$-dimensional coordinate space) can be described by the 3 coordinates for each of n positions.
- A rigid pendulum swinging in a vertical plane requires one generalised coordinate, the angle to the vertical.
- A top spinning on its axis on a smooth plane requires five generalised coordinates: two to describe the position of the point of contact, one for the angle of the axis to the vertical, one for the rotation of the axis about the vertical, and one for the rotation of the top about its axis.

Problems can often be simplified by a convenient choice of generalised coordinates – this is part of the power of these methods.

The *Lagrangian* is defined as

$$\boxed{\mathcal{L} = T - V} \tag{A.14}$$

where T is the kinetic energy and V is the potential energy. The action of a path, starting at time t_i and ending at t_f, is given by,

$$\boxed{S[\{q_i\}] = \int_{t_i}^{t_f} \mathcal{L}(\{q_i\}, \{\dot{q}_i\}, \cdots; t)\,\mathrm{d}t\,.} \tag{A.15}$$

Hamilton's principle states that the motion in configuration space extremises the action functional S. For $\mathcal{L}(\{q_i\}, \{\dot{q}_i\}; t)$, with N generalised coordinates and fixed start and end points,

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right) - \frac{\partial \mathcal{L}}{\partial q_i} = 0, \quad i = 1, \cdots, N.} \tag{A.16}$$

These are Lagrange's equations. If $\mathcal{L}(\{q_i\}, \{\dot{q}_i\}; t)$ has *no explicity dependence on* t, generalising the derivation of the first integral, we can find a constant of the motion. The chain rule and Lagrange's equations give,

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}t} = \frac{\partial \mathcal{L}}{\partial t} + \sum_{i=1}^{N}\left\{\dot{q}_i\frac{\partial \mathcal{L}}{\partial q_i} + \ddot{q}_i\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right\} = \frac{\partial \mathcal{L}}{\partial t} + \frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{i=1}^{N}\dot{q}_i\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right), \tag{A.17}$$

and hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathcal{L} - \sum_{i=1}^{N} \dot{q}_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right] = \frac{\partial \mathcal{L}}{\partial t}. \tag{A.18}$$

Given $\partial \mathcal{L}/\partial t = 0$ we have,

$$\boxed{\sum_{i=1}^{N} \dot{q}_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i} - \mathcal{L} = \text{const.}} \tag{A.19}$$

In general, if $\mathcal{L}$ does not explicitly depend on time, T is a homogeneous quadratic in the generalised velocities $\{\dot{q}_i\}$, i.e. $T \sim \sum_i \sum_j a_{ij}(q_1, \cdots, q_N)\dot{q}_i\dot{q}_j$, and V does not depend on $\{\dot{q}_i\}$, then it can be shown that,

$$\sum_{i=1}^{N} \dot{q}_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i} - \mathcal{L} = T + V = \text{const.} \tag{A.20}$$

i.e. the total energy $E = T + V$ is conserved.