Part II Relativity

William Royce

July 8, 2024

Part II Physics, The University of Cambridge

Abstract

> 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 Tables

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*.

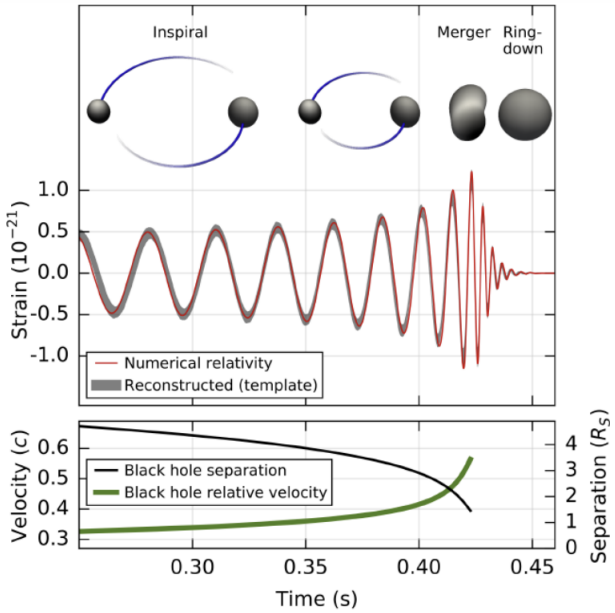

Fig. 1.1: Top: Estimated gravitational wave strain amplitude inferred from the LIGO data for their discovery event. The signal is generated from the inspiral, merger and ring-down of two massive black holes. The properties of the source can be estimated by comparing the measured waveform with detailed 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.

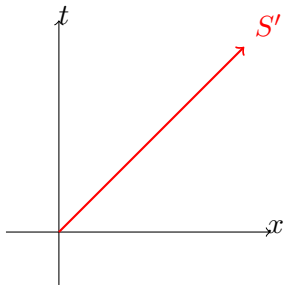

Fig. 2.1: Space-time diagram, representing the motion of a particle at the origin $x' = 0$ in S', which moves along the trajectory $x = vt$ in S.

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

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 $\mathbf{R}^4$. 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
- 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'.

[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.

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.

2.2.4 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:

$$\begin{aligned}\Delta s^2 &= c^2(\Delta t')^2 - (\Delta x')^2 - (\Delta y')^2 - (\Delta z')^2 \\ &= [(c\Delta t)\cosh\psi - (\Delta x)\sinh\psi]^2 - [-(c\Delta t)\sinh\psi + (\Delta x)\cosh\psi]^2 - \Delta y^2 - \Delta z^2 \\ &= c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2\end{aligned} \tag{2.20}$$

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

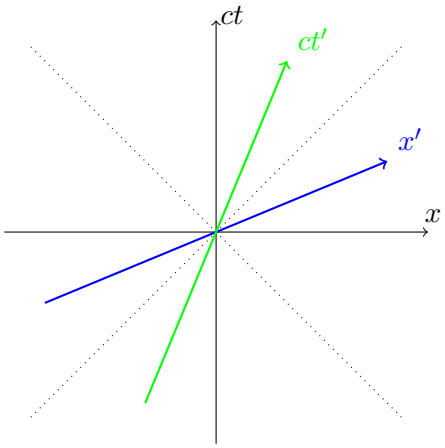

Fig. 2.2: Space-time diagram with axes corresponding to an inertial frame S' moving with a relative velocity. They can be the thought of as the x and ct axes, rotated by an equal amount towards the diagonal light ray. The fact the axes are symmetric about the light ray reflects the fact that the speed of light is equal to c in both frames.

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.

A particle moving in spacetime traces out a curve called a worldline as shown in the figure. 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 $x = vt$. Meanwhile, the x' axis is determined by $t' = 0$ which, from the Lorentz transformation (2.14), is given by the equation $ct = \frac{v}{c}x$.

These two axes are drawn on the Fig. (2.2). 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,

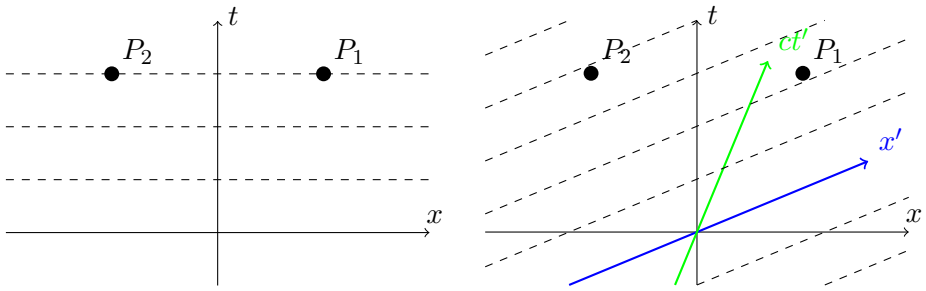

Fig. 2.3: Simultaneity is relative.

in the relativistic world, things are not so easy.

We start with an observer in inertial frame S, with time coordinate t. This observer sensibly decides that two events, P_1 and P_2, occur simultaneously if $t_1 = t_2$. In the spacetime diagram on the left of Fig. 2.3 we have drawn lines of simultaneity for this observer.

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 on the spacetime diagram on the right of Fig. 2.3.

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.

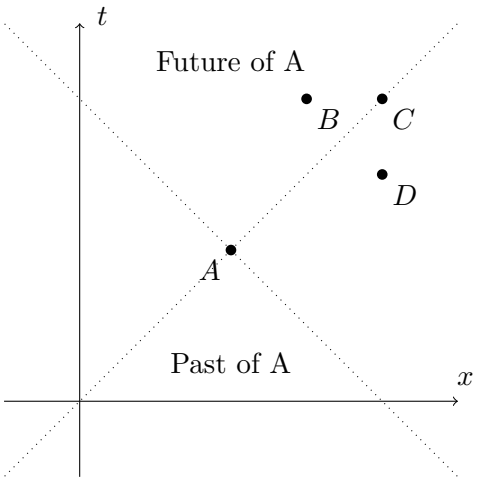

Fig. 2.4: Lightcone structure around the event A. Events B and A are separated by a timelike interval, and B lies in the forward lightcone of A. The events could be causally connected. Events C and A are separated by a null (or lightlike) interval and could be connected by a light signal. Events D and A are separated by a spacelike interval and cannot be causally connected.

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 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

$$t = \gamma\left(t' + \frac{vx'}{c^2}\right). \tag{2.21}$$

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

$$T = \gamma T' \tag{2.22}$$

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' sitting stationary in the frame S'. What is its length in frame S?

[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.

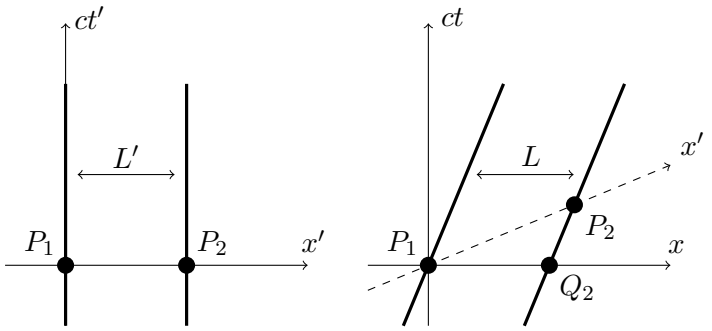

Fig. 2.5: The length L measured in frame S is $L = L'/\gamma$. It is shorter than the length of the rod in its rest frame by a factor of γ. This phenomenon is known as Lorentz contraction.

To begin, we should state more carefully something which seems obvious: when we say that a rod has length L', it means that the distance between the two end points at equal times is L'. So, drawing the axes for the frame S', the situation looks like the left diagram in Fig. 2.5. The two, simultaneous, end points in S' are P' and P'. Their coordinates in S' are $(ct', x') = (0, 0)$ and $(0, L')$ respectively.

Now let's look at this in frame S, illustrated in Fig. 2.5. Clearly P_1 sits at $(ct, x) = (0, 0)$. Meanwhile, the Lorentz transformation gives us the coordinate for P_2

$$x = \gamma L', \qquad \text{and,} \qquad t = \frac{\gamma v L'}{c^2}. \tag{2.23}$$

But to measure the rod in frame S, we want both ends to be at the same time. And the points P_1 and P_2 are not simultaneous in S. We can follow the point P_2 backwards along the trajectory of the end point to Q_2, which sits at

$$x = \gamma L' - vt. \tag{2.24}$$

We want Q_2 to be simultaneous with P_1 in frame S. This means we must move back a time $t = \gamma v L'/c^2$, giving

$$x = \gamma L' - \frac{\gamma v^2 L'}{c^2} = \frac{L'}{\gamma}. \tag{2.25}$$

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

$$L = \frac{L'}{\gamma} \tag{2.26}$$

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 rod suffers no contraction in the y- and z-directions (i.e., perpendicular to its velocity)

But hold on! If we look at the lengths of the rod, marked in the two diagrams in Fig. 2.5, it seems that we have got it the wrong way round: the length in frame S' definitely looks longer than in frame S. 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]

[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 $|(ct, x)| = (c^2t^2 - x^2)^{1/2}$.

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 $n = n'/\gamma$.

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 $\gamma \geq 2$.
- 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. 2.6) 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

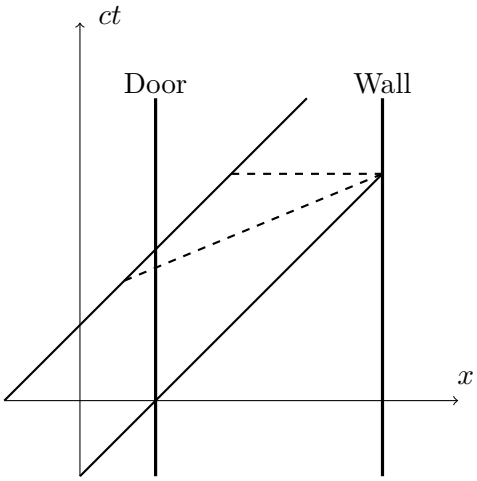

Fig. 2.6: The Ladder-and-Barn Non-Paradox: Regarded from the point of view of a space-time diagram, the paradox dissolves. One consequence of time not being invariant under Lorentz transformations is that the ladder 'fits in' the barn in one frame but does not 'fit in' in another.

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. 2.7, 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'.

[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.

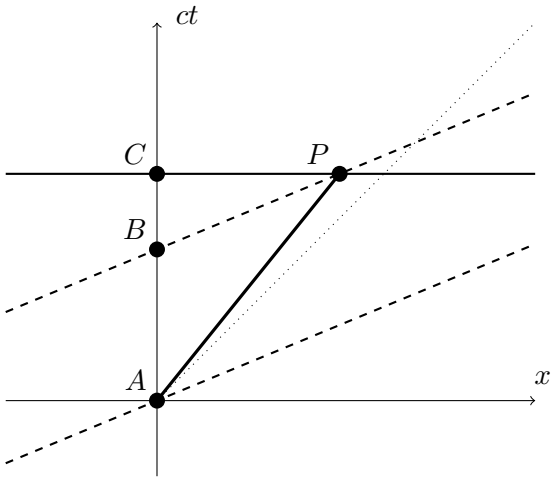

Fig. 2.7: The Twins Non-Paradox: Alice's world line is the ct (containing points A, B and C) axis and Bob's world line is the line containing A and P. P represents the event 'Bob arrives at Proxima Centauri'.

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:

$$T' = \gamma(T - v^2T/c^2) = T/\gamma = \tfrac{1}{2}T. \tag{2.27}$$

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)

$$\gamma(t + vx/c^2) = \tfrac{1}{2}T, \tag{2.28}$$

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

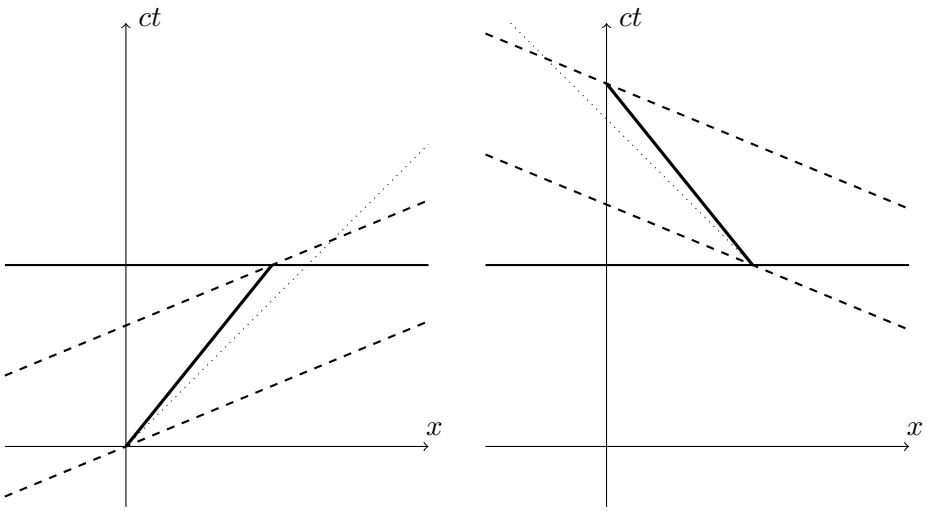

Fig. 2.8: Left: The outward journey. The heavy line is Bob's world line. The dotted line through the origin is the light cone. The dashed lines are the lines of simultaneity in Bob's frame. Right: The return journey. The heavy line is the world line of Bob′. The dotted line through the turn-round event is the light cone. The dashed lines are the lines of simultaneity in the frame of Bob′.

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 2.7–2.8.

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 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.

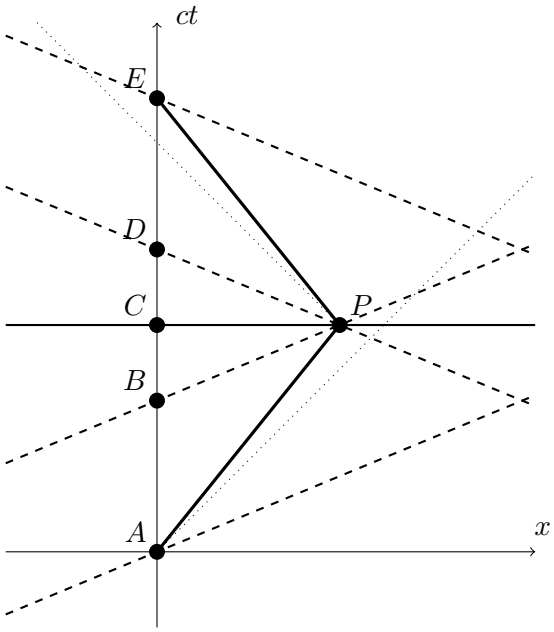

Fig. 2.9: The superposition of the previous two space-time diagrams in Fig. 2.8, representing together both the outward journey of Bob and the return journey of Bob′.

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:

$$\boxed{\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2\,.} \tag{2.29}$$

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

$$\Delta s = \int_A^B \mathrm{d}s\,. \tag{2.30}$$

The advantage of the invariant interval is that it is something all observers agree upon.

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,

$$c^2\,\mathrm{d}\tau^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \tag{2.31}$$

$$\implies \mathrm{d}\tau = \mathrm{d}t\,/\gamma, \tag{2.32}$$

i.e.

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \gamma. \tag{2.33}$$

Note that here in (2.33), 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

$$\Delta\tau = \int \mathrm{d}\tau = \int \gamma^{-1}\,\mathrm{d}t\,; \tag{2.34}$$

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

We can use proper time to derive the velocity addition formula (2.46) for an observer moving with non-constant velocity. We parameterise the observer's world line by τ:

$$x = x(\tau), \quad t = t(\tau), \qquad \text{in } S \tag{2.35}$$

$$x' = x'(\tau), \quad t' = t'(\tau), \qquad \text{in } S' \tag{2.36}$$

and

$$u = \frac{\mathrm{d}x}{\mathrm{d}\tau}\Big/\frac{\mathrm{d}t}{\mathrm{d}\tau}, \qquad u' = \frac{\mathrm{d}x'}{\mathrm{d}\tau}\Big/\frac{\mathrm{d}t'}{\mathrm{d}\tau}. \tag{2.37}$$

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

$$\frac{\mathrm{d}x'}{\mathrm{d}\tau} = \gamma\left(\frac{\mathrm{d}x}{\mathrm{d}\tau} - v\frac{\mathrm{d}t}{\mathrm{d}\tau}\right) = \gamma(u-v)\frac{\mathrm{d}t}{\mathrm{d}\tau} \tag{2.38}$$

$$\frac{\mathrm{d}t'}{\mathrm{d}\tau} = \gamma\left(\frac{\mathrm{d}t}{\mathrm{d}\tau} - \frac{v}{c^2}\frac{\mathrm{d}x}{\mathrm{d}\tau}\right) = \gamma(1-\frac{uv}{c^2})\frac{\mathrm{d}t}{\mathrm{d}\tau} \tag{2.39}$$

and dividing these expressions gives

$$u' = \frac{u-v}{1-uv/c^2}. \tag{2.40}$$

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. 2.10). 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.

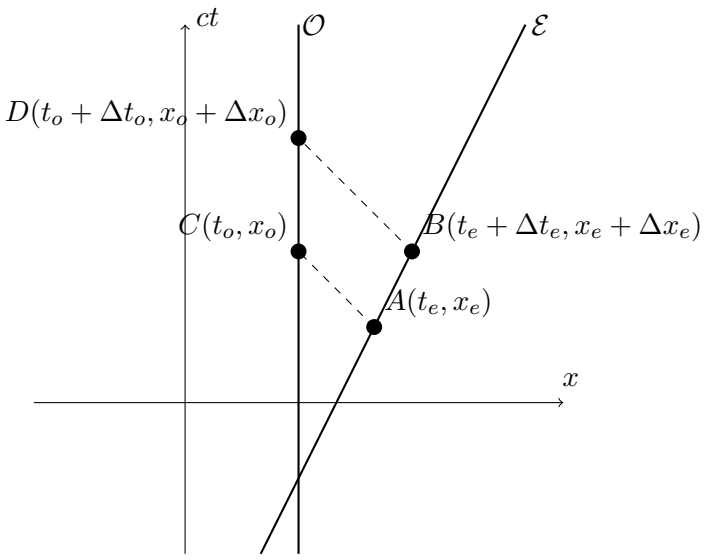

Fig. 2.10: Spacetime diagram of the Doppler effect. An observer $\mathcal{E}$ moves at speed v along the x-axis of an inertial frame S in which an observer $\mathcal{O}$ is at rest at position x_o. A wavecrest is emitted by $\mathcal{E}$ at the event A with coordinates (t_e, x_e) in S and is received 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 relation between $\Delta\tau_{AB}$ and the time Δt_e between the emission events in S is

$$\Delta\tau_{AB} = \Delta t_e/\gamma. \tag{2.41}$$

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

$$\Delta t_o = \left(1 + \frac{v}{c}\right)\Delta t_e, \tag{2.42}$$

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

$$\frac{\Delta\tau_{AB}}{\Delta\tau_{CD}} = \frac{\sqrt{1 - \frac{v^2}{c^2}}\Delta t_e}{(1 + \frac{v}{c})\Delta t_e} = \sqrt{\frac{1 - \frac{v}{c}}{1 + \frac{v}{c}}}. \tag{2.43}$$

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

$$x' = ut'. \tag{2.44}$$

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

$$u = \frac{x}{t} = \frac{\gamma(x' + vt')}{\gamma(t' + vx'/c^2)}, \tag{2.45}$$

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 (2.44) into the expression above, and performing a little algebra, gives us the result we want:

$$u = \frac{u' + v}{1 + u'v/c^2}. \tag{2.46}$$

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

$$c - u = c - \frac{u' + v}{1 + u'v/c^2} = \frac{c(c - u')(c - v)}{c^2 + u'v} > 0, \tag{2.47}$$

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

$$u_x = \frac{\mathrm{d}x'}{\mathrm{d}t'} = \frac{u'_x + v}{1 + u'_x v/c^2}, \tag{2.48}$$

$$u_y = \frac{\mathrm{d}y'}{\mathrm{d}t'} = \frac{u'_y}{\gamma(1 + u'_x v/c^2)}, \tag{2.49}$$

$$u_z = \frac{\mathrm{d}z'}{\mathrm{d}t'} = \frac{u'_z}{\gamma(1 + u'_x v/c^2)}. \tag{2.50}$$

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:

$$\frac{u}{c} = \tanh\beta, \qquad \frac{u'}{c} = \tanh\beta', \qquad \frac{v}{c} = \tanh\alpha. \tag{2.51}$$

We can then find the composition of the two Lorentz transforms in terms of the rapidities by substitution into (2.40) to give

$$\tanh\beta' = \frac{\tanh\beta - \tanh\alpha}{1 - \tanh\alpha\tanh\beta} = \tanh(\beta - \alpha). \tag{2.52}$$

so an alternative form of the transformation law is

$$\beta' = \beta - \alpha \tag{2.53}$$

i.e.

$$\tanh^{-1}\left(\frac{u'}{c}\right) = \tanh^{-1}\left(\frac{u}{c}\right) + \tanh^{-1}\left(\frac{v}{c}\right). \tag{2.54}$$

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).

2.5 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.

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.

For reasons that will later become clear (see section **??**), we need to 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 (2.46) 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

$$u' = \frac{u - v}{1 - uv/c^2} \equiv \frac{(c^2/v)(1 - v^2/c^2)}{1 - uv/c^2} - \frac{c^2}{v}. \tag{2.55}$$

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

$$\frac{\mathrm{d}u'}{\mathrm{d}\tau} = \frac{1 - v^2/c^2}{(1 - uv/c^2)^2}\frac{\mathrm{d}u}{\mathrm{d}\tau}. \tag{2.56}$$

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

$$\begin{aligned} a' &= \frac{\mathrm{d}u'}{\mathrm{d}t'} \\ &= \frac{\mathrm{d}u'}{\mathrm{d}\tau}\Big/\frac{\mathrm{d}t'}{\mathrm{d}\tau} \\ &= \frac{1-v^2/c^2}{(1-uv/c^2)^2}\frac{\mathrm{d}u}{\mathrm{d}\tau}\Big/\frac{\mathrm{d}t'}{\mathrm{d}\tau} && \text{(using (2.56))} \\ &= \frac{1-v^2/c^2}{(1-uv/c^2)^2}\frac{\mathrm{d}u}{\mathrm{d}\tau}\Big/\gamma(1-uv/c^2)\frac{\mathrm{d}t}{\mathrm{d}\tau} && \text{(using (2.38))} \\ &= \frac{(1-v^2/c^2)^{3/2}}{(1-uv/c^2)^3}a \\ &= \frac{1}{\gamma^3(1-uv/c^2)^3}a. && (2.57) \end{aligned}$$

As mentioned above there are other ways of obtaining this result; for example, more elegantly using four-vectors (see section **??**).

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 (2.57) becomes

$$a = a'/\gamma^3. \tag{2.58}$$

Now

$$a = \frac{\mathrm{d}u}{\mathrm{d}\tau}\Big/\frac{\mathrm{d}t}{\mathrm{d}\tau}, \tag{2.59}$$

and using (2.33), so we can find the parameterised equation of the world line by integrating

$$\frac{\mathrm{d}u}{\mathrm{d}\tau} = a\frac{\mathrm{d}t}{\mathrm{d}\tau} = a'/\gamma^2. \tag{2.60}$$

This gives

$$u = c\tanh\left(a'\tau/c\right), \tag{2.61}$$

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

$$\gamma = \cosh\left(a'\tau/c\right). \tag{2.62}$$

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

$$t = c/a'\sinh\left(a'\tau/c\right), \tag{2.63}$$

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

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}\tau} &= \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau} && (2.64) \\ &= u\gamma && (2.65) \\ &= c\sinh\left(a'\tau/c\right), && (2.66) \end{aligned}$$

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

$$x = c^2/a'\cosh\left(a'\tau/c\right). \tag{2.67}$$

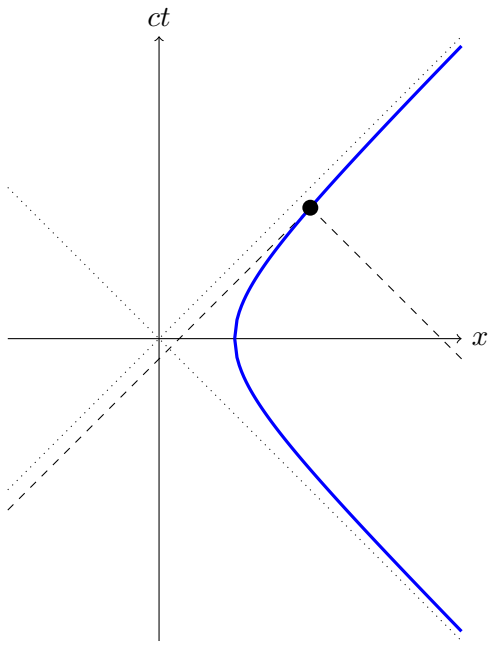

Fig. 2.11: The space-time diagram for an accelerated observer. The thick hyperbola is the observer's world line. An observer 'below' the dashed lines could in principle send a message to the observer marked as a heavy dot; other observers could not.

Uniformly accelerated particles therefore move on rectangular hyperbolas of the form

$$x^2 - (ct)^2 = (c^2/a')^2. \tag{2.68}$$

The diagram in Fig 2.11 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 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 $\tau \to \infty$.

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 section 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 section 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 $\mathbf{R}^n$. 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 $\mathbf{R}^n$, the sphere $\mathbf{S}^n$, and the torus $\mathbf{T}^n = \mathbf{S}^1 \times \cdots \times \mathbf{S}^1$. 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 $\mathbf{R}^n$. Globally, the manifold may be more interesting than $\mathbf{R}^n$, 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 $\mathbf{R}^n$. 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 $\mathbf{R}^n$ 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 $\mathbf{R}^n$ 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 $\mathbf{R}^n$.

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 $\mathbf{R}^2$: The Euclidean plane $\mathbf{R}^2$ 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 $\mathbf{R}^n$ 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 $\mathbf{S}^2$: The 2-sphere is the set of points in $\mathbf{R}^3$ 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 $\mathbf{S}^2$, 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:

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

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.

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 $\mathbf{R}^n$).

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,

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

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.,

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

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

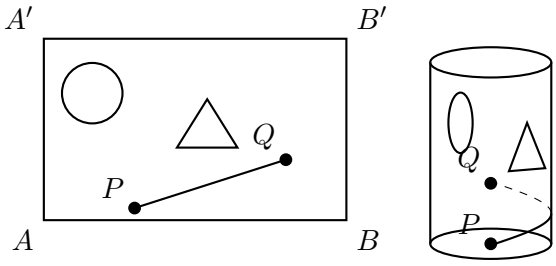

Fig. 3.1: The Euclidean plane $\mathbf{R}^2$ can be rolled up into a cylindrical surface without distortion. The intrinsic geometry of the cylindrical surface is therefore the same as the plane. In particular, a bug confined to the surface would measure the sum of the angles of a triangle to be 180°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

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 $\mathbf{R}^3$ (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 $\mathbf{R}^2$ 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 $\mathbf{R}^3$ is clearly curved (non-Euclidean). We can contrast the cylinder to a 2-sphere of radius a embedded in $\mathbf{R}^3$. 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 $\mathbf{R}^3$: 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 $\mathbf{R}^3$; 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 ($\mathbf{R}^3$ 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 $\mathbf{R}^4$: Now consider the 3-sphere, defined by $x^2 + y^2 + z^2 + w^2 = a^2$, embedded in 4D Euclidean space $\mathbf{R}^4$ 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

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

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 $\mathbf{R}^3$ 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 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 $\mathbf{R}^3$

Consider again the 2-sphere of radius a embedded in $\mathbf{R}^3$. 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.

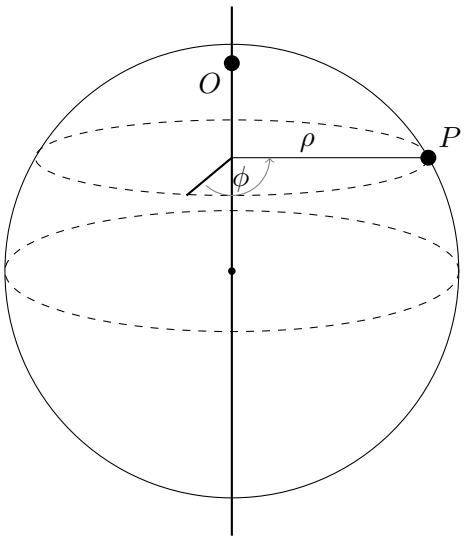

Fig. 3.2: Surface of the 2-sphere in $\mathbf{R}^3$, with centre O.

- 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 $\mathbf{R}^3$ has same local intrinsic geometry as the Euclidean plane $\mathbf{R}^2$, 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 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 $\mathbf{R}^2$ and $\mathbf{R}^3$, 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, $\phi'(x'^a)$, 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

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

at P, where x^a is some coordinate chart.

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)

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

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

$$\mathbf{v} \to v'^a \frac{\partial}{\partial x'^a} \tag{4.5}$$

$$= \underbrace{\frac{\partial x'^a}{\partial x^b}\frac{\partial x^c}{\partial x'^a}}_{\delta^b_c} v^b \frac{\partial}{\partial x^c} \tag{4.6}$$

$$= v^b \frac{\partial}{\partial x^b} \tag{4.7}$$

$$= \mathbf{v} \tag{4.8}$$

Note that the components v^a and the basis vectors transform inversely under changes of coordinates. 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 as

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

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

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

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

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

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 and type-$(0, 1)$ tensors are dual vectors. 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 $\mathbf{S}\otimes\mathbf{T}$.

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

$$\begin{aligned} T'^{ab} &= \frac{\partial x'^a}{\partial x^c}u^c\frac{\partial x'^b}{\partial x^d}v^d \\ &= \frac{\partial x'^a}{\partial x^c}\frac{\partial x'^b}{\partial x^d}T^{cd}, \end{aligned} \tag{4.18}$$

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

$$T'^{ab}{}_c = \frac{\partial x'^a}{\partial x^p}\frac{\partial x'^b}{\partial x^q}\frac{\partial x^r}{\partial x'^c}T^{pq}{}_r\,, \tag{4.19}$$

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

$$\begin{aligned} S'^a &= \frac{\partial x'^a}{\partial x^p} \underbrace{\frac{\partial x'^b}{\partial x^q}\frac{\partial x^r}{\partial x'^c}}_{\delta^q_r} T^{pq}{}_r \\ &= \frac{\partial x'^a}{\partial x^p} T^{pq}{}_q \\ &= \frac{\partial x'^a}{\partial x^p} S^p. \end{aligned} \tag{4.20}$$

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^aX_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

$$S_{ab} = \frac{1}{2}(S_{ab} + S_{ba}) + \frac{1}{2}(S_{ab} - S_{ba}). \tag{4.21}$$

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

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

Antisymmetrisation is usually denoted by square brackets:

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

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

$$\begin{aligned} S_{(ab\cdots c)} &= \frac{1}{n!}(\text{sum over all perms of } a, b, \cdots, c), \\ S_{[ab\cdots c]} &= \frac{1}{n!}(\text{alternating sum over all perms}), \end{aligned} \tag{4.24}$$

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

$$S_{[abc]} = \frac{1}{6}(S_{abc} - S_{acb} + S_{cab} - S_{abc} + S_{bca} - S_{bac}). \tag{4.25}$$

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

$$S_{(ab)c} = \frac{1}{2}(S_{abc} - S_{bac}). \tag{4.26}$$

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

$$T'^a{}_{bc}\,v^c = \frac{\partial x'^a}{\partial x^d}\frac{\partial x^e}{\partial x'^b}T^d{}_{ef}\,v^f. \tag{4.27}$$

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

$$\begin{aligned} T'^a{}_{bc}\frac{\partial x'^c}{\partial x^f} &= \frac{\partial x'^a}{\partial x^d}\frac{\partial x^e}{\partial x'^b}T^d{}_{ef} \\ \implies T'^a{}_{bc} &= \frac{\partial x'^a}{\partial x^d}\frac{\partial x^e}{\partial x'^b}\frac{\partial x^f}{\partial x'^c}T^d{}_{ef}\,. \end{aligned} \tag{4.28}$$

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

$$\mathrm{d}s^2 = g_{ab}\,\mathrm{d}x^a\,\mathrm{d}x^b\,. \tag{4.29}$$

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

$$g'_{ab} = \frac{\partial x^c}{\partial x'^a}\frac{\partial x^d}{\partial x'^b} g_{cd} \tag{4.30}$$

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

$$\mathbf{g}(\mathbf{u},\mathbf{v}) = g_{ab}u^a v^b. \tag{4.31}$$

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

$$\boxed{v_a \equiv g_{ab}v^b.} \tag{4.32}$$

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 $T^a{}_b$,

$$T_{ab} \equiv g_{ac}T^c{}_b \tag{4.33}$$

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

$$T_{abc} \equiv g_{ap}g_{bq}T^{pq}{}_c\,. \tag{4.34}$$

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

$$\left(g^{-1}\right)^{ab} g_{bc} = \delta^a_c. \tag{4.35}$$

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

$$\left(g'^{-1}\right)^{ab} = \frac{\partial x'^a}{\partial x^c}\frac{\partial x'^b}{\partial x^d}\left(g^{-1}\right)^{cd}, \tag{4.36}$$

since then

$$\begin{aligned}\left(g'^{-1}\right)^{ab} g'_{bc} &= \frac{\partial x'^a}{\partial x^p}\left(g^{-1}\right)^{pq}\underbrace{\frac{\partial x'^b}{\partial x^q}\frac{\partial x^r}{\partial x'^b}}_{\delta^r_q}\frac{\partial x^s}{\partial x'^c}g_{rs}\\ &= \frac{\partial x'^a}{\partial x^p}\frac{\partial x^s}{\partial x'^c}\underbrace{\left(g^{-1}\right)^{pq}g_{qs}}_{\delta^p_s}\\ &= \frac{\partial x'^a}{\partial x^p}\frac{\partial x^p}{\partial x'^c}\\ &= \delta^a_c,\end{aligned} \tag{4.37}$$

as required.

However, Eq. (4.36) 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

$$g_{ac}g_{bd}g^{cd} = g_{ac}\delta^b_c = g_{ab}. \tag{4.38}$$

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

$$X^a \equiv g^{ab}X_b, \tag{4.39}$$

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[2]:

$$v^a \xrightarrow{\mathbf{g}} g_{ab}v^b \xrightarrow{\mathbf{g}^{-1}} g^{ac}g_{cb}v^b = v^a. \tag{4.40}$$

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

$$T_a{}^{bc} \equiv g_{ad}g^{ce}T^{db}{}_e\,. \tag{4.41}$$

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. 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).

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

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:

$$g^{ab}g_{bc} = \delta^a_c. \tag{4.42}$$

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,

$$\begin{aligned} g'^a{}_b &= \frac{\partial x'^a}{\partial x^c}\frac{\partial x^d}{\partial x'^b}g^c{}_d \\ &= \frac{\partial x'^a}{\partial x^c}\frac{\partial x^c}{\partial x'^b} \\ &= \delta^a_b = g^a{}_b, \end{aligned} \tag{4.43}$$

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:

$$g_{ab}u^a v^b = g^{ab}u_a v_b = u^a v_a = u_a v^a. \tag{4.44}$$

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 non-zero vectors (*null vectors*) v^a with $g_{ab}v^a v^b = 0$.

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

$$|\mathbf{v}| \equiv \left|g_{ab}v^a v^b\right|^{1/2}; \tag{4.45}$$

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

$$\cos\theta \equiv \frac{u_a v^a}{|u_b u^b|^{1/2}|v_c v^c|^{1/2}}. \tag{4.46}$$

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 $\boldsymbol{\nabla}_c T^{a_1\cdots a_k}{}_{b_1\cdots b_l}$, 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.,

$$\boldsymbol{\nabla}_a \phi = \frac{\partial \phi}{\partial x^a}. \tag{5.4}$$

- 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.,

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

- 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

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

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

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

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 $\boldsymbol{\nabla}_a v^b$ transforms as a type-$(1,1)$ tensor.

Forming the covariant derivative in the new coordinates, we have

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

If $\boldsymbol{\nabla}_a v^b$ 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

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

so that, generally,

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

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

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

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

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

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

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

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

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

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

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

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:

$$\boxed{\boldsymbol{\nabla}_a g_{bc} = 0.} \tag{5.17}$$

- Commutative action on scalar fields, so that

$$\boxed{\boldsymbol{\nabla}_a \boldsymbol{\nabla}_b \phi = \boldsymbol{\nabla}_b \boldsymbol{\nabla}_a \phi.} \tag{5.18}$$

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 $\boldsymbol{\nabla}_a \boldsymbol{\nabla}_b \phi$ as

$$\boldsymbol{\nabla}_a \boldsymbol{\nabla}_b \phi = \partial_a \partial_b \phi - \Gamma^c_{ab} \partial_c \phi. \tag{5.20}$$

The first term on the right is symmetric in a and b, so if $\boldsymbol{\nabla}_{[a} \boldsymbol{\nabla}_{b]} \phi = 0$ 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 $\boldsymbol{\nabla}_{[a} \boldsymbol{\nabla}_{b]} \phi = 0$. We now turn to metric compatibility:

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

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,

$$\boldsymbol{\nabla}_a g^{bc} = 0, \tag{5.26}$$

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

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

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.,

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

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 $\boldsymbol{\nabla}_c g_{ab} = 0$, 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

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

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

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

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 $\boldsymbol{\nabla}_a v^a$. It follows from Eq. (4.36) that

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

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

$$(\mathrm{curl}\mathbf{X})_{ab} \equiv \boldsymbol{\nabla}_a X_b - \boldsymbol{\nabla}_b X_a. \tag{5.43}$$

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

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

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

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

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

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

The Laplacian generalises to tensor fields, e.g.

$$\nabla^2 T^{ab} = g^{cd}\boldsymbol{\nabla}_c\boldsymbol{\nabla}_d T^{ab}. \tag{5.46}$$

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

$$\frac{Dv^a}{Du} \equiv \frac{\mathrm{d}x^b}{\mathrm{d}u}\boldsymbol{\nabla}_b v^a = \frac{\mathrm{d}x^b}{\mathrm{d}u}(\partial_b v^a + \Gamma^a_{bc}v^c). \tag{5.47}$$

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:

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

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

- $\mathrm{d}v^a/\mathrm{d}u$ 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

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

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

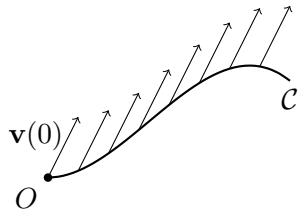

Fig. 5.1: The vector field $\mathbf{v}(u)$ defined by the parallel transport of a vector $\mathbf{v}(0)$ along a curve $\mathcal{C}$ defined in 2D Euclidean space in Cartesian coordinates by $x^a(u)$.

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, $Dv^a/Du = 0$, 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

$$\boxed{\frac{Dv^a}{Du} = 0.} \tag{5.51}$$

This definition easily extends to parallel transport of other tensors, e.g., $DT^{ab}/Du = 0$.

5.3.1 Properties of Parallel Transport

Note the following properties of parallel transport:

- The equation $Dv^a/Du = 0$ 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[3]

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

- 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

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

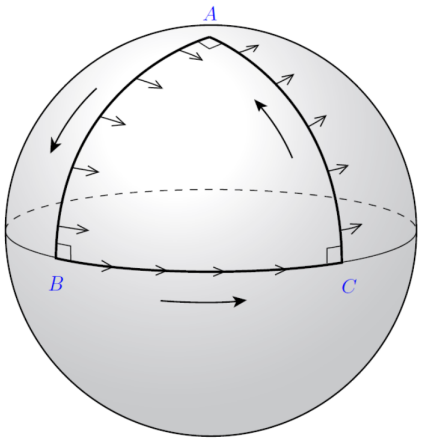

Fig. 5.2: Parallel transport around a closed path on the surface of the 2-sphere. The path consists of a great circle through the north pole (A) down to the equator at B, a length of the equator from B to C, and the great circle through C and A. The vector indicated by the small arrows is parallel transported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

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[4], including massless particles, follow geodesic curves in spacetime.

[4]In this context, a test particle is supposed to have sufficiently small mass that its motion does not affect the spacetime geometry

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

$$L = \int_{\kappa(0)}^{\kappa(1)} \left| g_{ab} \frac{\mathrm{d}x^a}{\mathrm{d}\kappa} \frac{\mathrm{d}x^b}{\mathrm{d}\kappa} \right|^{1/2} \mathrm{d}\kappa\,. \tag{5.58}$$

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.1. 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

$$\frac{Dt^a}{Du} = \left(\frac{\ddot{s}}{\dot{s}}\right)t^a. \tag{5.65}$$

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

$$\boxed{\frac{Dt^a}{Du} = 0.} \tag{5.67}$$

Indeed, an equivalent definition of a non-null geodesic with affine parameterisation is that it is a curve whose tangent vector is parallel transported.[5] 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 $Dt^a/Du = 0$ 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 $Dt^a/Du = 0$ 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:

$$g_{ab}\frac{Dt^a}{Du} = 0 \implies \frac{Dt_a}{Du} = \frac{\mathrm{d}t_a}{\mathrm{d}u} - \Gamma^c_{ba}t^bt_c = 0. \tag{5.68}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(g_{ab}\frac{\mathrm{d}x^b}{\mathrm{d}u}\right) = \frac{1}{2}\frac{\partial g_{bc}}{\partial x^a}\frac{\mathrm{d}x^b}{\mathrm{d}u}\frac{\mathrm{d}x^c}{\mathrm{d}u}. \tag{5.71}$$

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"

$$L = g_{ab}\frac{\mathrm{d}x^a}{\mathrm{d}u}\frac{\mathrm{d}x^b}{\mathrm{d}u}. \tag{5.73}$$

[5] For connections more general than the metric connection, such *auto-parallel curves* are generally non-geodesic.

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

$$\mathbf{g}(\mathbf{e}_\mu, \mathbf{e}_\nu) = \eta_{\mu\nu}. \tag{6.16}$$

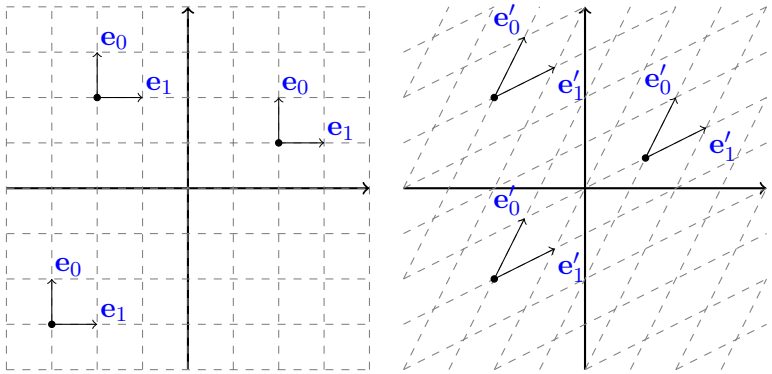

Fig. 6.1: Coordinate curves for two systems of coordinates x^μ and x'^μ, corresponding to Cartesian inertial frames S and S' in standard configuration. The coordinate basis vectors for each system are also shown, indicated as arrows tangent to the coordinate curves. The 2− and 3− directions are suppressed 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

$$\begin{array}{ll} \eta_{\mu\nu}v^\mu v^\nu > 0 & \text{timelike,} \\ \eta_{\mu\nu}v^\mu v^\nu < 0 & \text{spacelike,} \\ \eta_{\mu\nu}v^\mu v^\nu = 0 & \text{null.} \end{array} \tag{6.21}$$

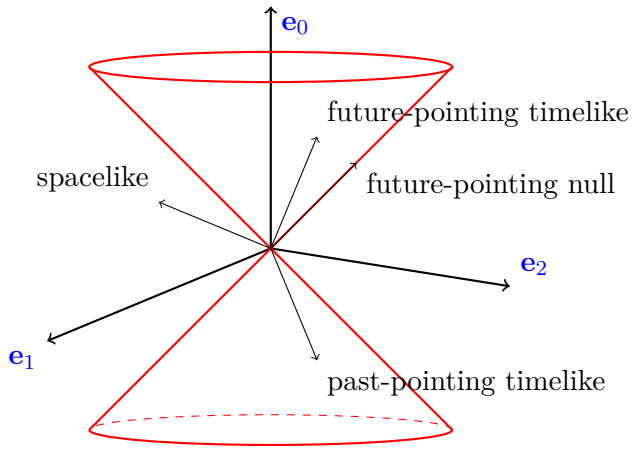

Fig. 6.2: A vector $\mathbf{v}$ is timelike, spacelike, or null according to the character of $\mathbf{g}(\mathbf{v}, \mathbf{v})$; in Cartesian coordinates $\mathbf{v}$ is timelike for $\eta_{\mu\nu}v^\mu v^\nu > 0$; spacelike for $\eta_{\mu\nu}v^\mu v^\nu < 0$; and null for $\eta_{\mu\nu}v^\mu v^\nu = 0$. 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 Section 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 Section 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:

$$\frac{Du^\mu}{D\tau} = 0. \tag{6.35}$$

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

$$a^\mu = \frac{Du^\mu}{D\tau}. \tag{6.36}$$

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 **v** 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):

$$E = \gamma_u mc^2. \tag{6.46}$$

To see this, consider the rate of working $\vec{f} \cdot \vec{u}$ of the force accelerating a particle:

$$\begin{aligned}
\vec{u} \cdot \vec{f} &= \vec{u} \cdot \frac{\mathrm{d}\vec{p}}{\mathrm{d}t} \\
&= \vec{u} \cdot \frac{\mathrm{d}}{\mathrm{d}t}(\gamma_u m\vec{u}) \\
&= \gamma_u m\left(\vec{u} \cdot \vec{a} + \gamma_u^2 \vec{u} \cdot \vec{a}\frac{|\vec{u}|^2}{c^2}\right) \\
&= \gamma_u^3 m\vec{u} \cdot \vec{a} \\
&= mc^2\frac{\mathrm{d}\gamma_u}{\mathrm{d}t}.
\end{aligned} \tag{6.47}$$

With $E = \gamma_u mc^2$, 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*

$$E^2 - |\vec{p}|^2c^2 = m^2c^4. \tag{6.49}$$

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,

$$\frac{Dp^\mu}{D\tau} = 0. \tag{6.50}$$

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

$$\frac{Dp^\mu}{D\tau} = f^\mu. \tag{6.51}$$

[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

$$\frac{Dp^\mu}{D\lambda} = 0. \tag{6.56}$$

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 Section 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

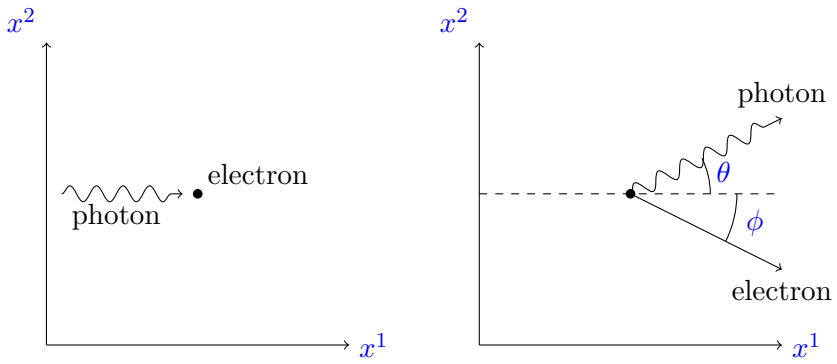

Fig. 6.3: The Compton effect showing a photon initially propagating along the x-axis scattering off an electron at rest (left). After the collision (right), the photon propagates at an angle θ to the x-axis, and the electron recoils.

process the electron recoils (see Fig. 6.3). The components of the relevant 4-momenta are

$$\begin{aligned} p^\mu &= (h\nu/c, h\nu/c, 0, 0) \\ q^\mu &= (m_e c, 0, 0, 0) \\ \bar{p}^\mu &= (h\bar{\nu}/c, (h\bar{\nu}/c)\cos\theta, (h\bar{\nu}/c)\sin\theta, 0, \end{aligned} \tag{6.64}$$

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.

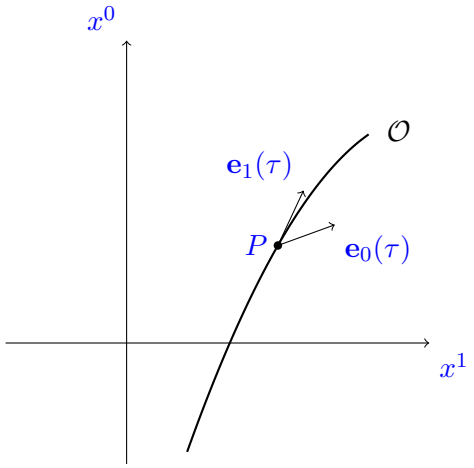

Fig. 6.4: For a general observer $\mathcal{O}$ following a worldline $x^\mu(\tau)$, we can define the instantaneous rest-frame of the particle as the inertial frame in which the particle is instantaneously at rest. At proper time τ, the coordinate basis vectors of the instantaneous rest-frame at the observer's position, P, constitute an orthonormal set of basis vectors $\mathbf{e}_\mu(\tau)$.

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 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.

[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

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:

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

Chapter 7

Chapter

Chapter 8

Chapter

Chapter 9

Chapter

Chapter 10

Chapter

Chapter 11

Chapter

Chapter 12

Chapter

Chapter 13

Chapter

Appendix A

Appendix

A.1 Euler-Lagrange Equations