Chapter 2

Geometry

Our tour of theoretical physics begins with geometry, and there are two reasons for this. One is that the framework of space and time provides, as it were, the stage upon which physical events are played out, and it will be helpful to gain a clear idea of what this stage looks like before introducing the cast. As a matter of fact, the geometry of space and time itself plays an active role in those physical processes that involve gravitation (and perhaps, according to some speculative theories, in other processes as well). Thus, our study of geometry will culminate, in chapter 4, in the account of gravity offered by Einstein's general theory of relativity. The other reason for beginning with geometry is that the mathematical notions we develop will reappear in later contexts.

To a large extent, the special and general theories of relativity are 'negative' theories. By this I mean that they consist more in relaxing incorrect, though plausible, assumptions that we are inclined to make about the nature of space and time than in introducing new ones. I propose to explain how this works in the following way. We shall start by introducing a prototype version of space and time, called a 'differentiable manifold', which possesses a bare minimum of geometrical properties—for example, the notion of length is not yet meaningful. (Actually, it may be necessary to abandon even these minimal properties if, for example, we want a geometry that is fully compatible with quantum theory and I shall touch briefly on this in chapter 15.) In order to arrive at a structure that more closely resembles space and time as we know them, we then have to endow the manifold with additional properties, known as an 'affine connection' and a 'metric'. Two points then emerge: first, the common-sense notions of Euclidean geometry correspond to very special choices for these affine and metric properties; second, other possible choices lead to geometrical states of affairs that have a natural interpretation in terms of gravitational effects. Stretching the point slightly, it may be said that, merely by *avoiding* unnecessary assumptions, we are able to see gravitation as something entirely to be expected, rather than as a phenomenon in need of explanation.

To me, this insight into the ways of nature is immensely satisfying, and it is in the hope of communicating this satisfaction to readers that I have chosen to approach the subject in this way. Unfortunately, the assumptions we are to avoid are, by and large, *simplifying* assumptions, so by avoiding them we let ourselves in for some degree of complication in the mathematical formalism. Therefore, to help readers preserve a sense of direction, I will, as promised in chapter 1, provide an introductory section outlining a more traditional approach to relativity and gravitation, in which we ask how our naïve geometrical ideas must be modified to embrace certain observed phenomena.

2.0 The Special and General Theories of Relativity

2.0.1 The special theory

The special theory of relativity is concerned in part with the relation between observations of some set of physical events in two inertial frames of reference that are in relative motion. By an inertial frame, we mean one in which Newton's first law of motion holds:

> Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed on it.
> (Newton 1686)

It is worth noting that this definition by itself is in danger of being a mere tautology, since a 'force' is in effect defined by Newton's second law in terms of the acceleration it produces:

> The change of motion is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed.
> (Newton 1686)

So, from these definitions alone, we have no way of deciding whether some observed acceleration of a body relative to a given frame should be attributed, on the one hand, to the action of a force or, on the other hand, to an acceleration of the frame of reference. Eddington has made this point by a facetious re-rendering of the first law:

> Every body tends to move in the track in which it actually does move, except insofar as it is compelled by material impacts to follow some other track than that in which it would otherwise move.
> (Eddington 1929)

The extra assumption we need, of course, is that forces can arise only from the influence of one body on another. An inertial frame is one relative to which any body sufficiently well isolated from all other matter for these influences to be negligible does not accelerate. In practice, needless to say, this isolation cannot be achieved. The successful application of Newtonian mechanics depends on our being able systematically to identify, and take proper account of, all those forces

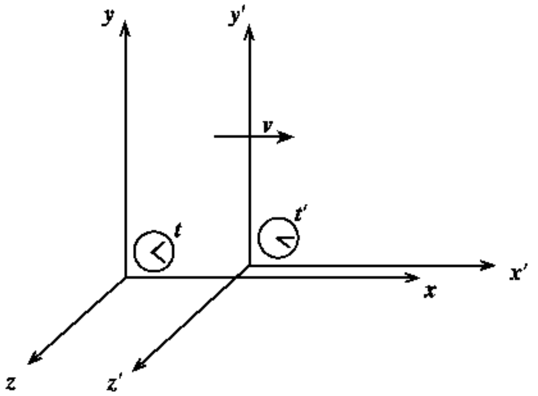


Figure 2.1. Two systems of Cartesian coordinates in relative motion.

that cannot be eliminated. To proceed, we must take it as established that, in principle, frames of reference can be constructed, relative to which any isolated body will, as a matter of fact, always refuse to accelerate. These frames we call inertial.

Obviously, any two inertial frames must either be relatively at rest or have a uniform relative velocity. Consider, then, two inertial frames, S and S' (standing for *S*ystems of coordinates) with Cartesian axes so arranged that the x and x' axes lie in the same line, and suppose that S' moves in the positive x direction with speed v relative to S. Taking y' parallel to y and z' parallel to z, we have the arrangement shown in figure 2.1. We assume that the sets of apparatus used to measure distances and times in the two systems are identical and, for simplicity, that both clocks are adjusted to read zero at the moment the two origins coincide.

Suppose that an event at the coordinates (x, y, z, t) relative to S is observed at (x', y', z', t') relative to S'. According to the Galilean, or common-sense, view of space and time, these two sets of coordinates must be related by

$$x' = x - vt \qquad y' = y \qquad z' = z \qquad t' = t. \tag{2.1}$$

Since the path of a moving particle is just a sequence of events, we easily find that its velocity relative to S, in vector notation $\boldsymbol{u} = \mathrm{d}\boldsymbol{x}/\mathrm{d}t$, is related to its velocity $\boldsymbol{u}' = \mathrm{d}\boldsymbol{x}'/\mathrm{d}t'$ relative to S' by $\boldsymbol{u}' = \boldsymbol{u} - \boldsymbol{v}$, with $\boldsymbol{v} = (v, 0, 0)$, and that its acceleration is the same in both frames, $\boldsymbol{a}' = \boldsymbol{a}$.

Despite its intuitive plausibility, the common-sense view turns out to be mistaken in several respects. The special theory of relativity hinges on the fact that the relation $\boldsymbol{u}' = \boldsymbol{u} - \boldsymbol{v}$ is not true. That is to say, this relation disagrees with experimental evidence, although discrepancies are detectable only when speeds are involved whose magnitudes are an appreciable fraction of a fundamental speed c, whose value is approximately $2.998 \times 10^8\,\mathrm{m\,s^{-1}}$. So far as is known, light travels through a vacuum at this speed, which is, of course, generally called the speed of light. Indeed, the speed of light is predicted by Maxwell's electromagnetic theory to be $(\epsilon_0\mu_0)^{-1/2}$ (in SI units, where ϵ_0 and μ_0 are called the permittivity and permeability of free space, respectively) but the theory does not single out any special frame relative to which this speed should be measured. For quite some time after the appearance of Maxwell's theory (published in its final form in 1864; see also Maxwell (1873)), it was thought that electromagnetic radiation consisted of vibrations of a medium, the 'luminiferous ether', and would travel at the speed c relative to the rest frame of the ether. However, a number of experiments cast doubt on this interpretation. The most celebrated, that of Michelson and Morley (1887), showed that the speed of the Earth relative to the ether must, at any time of year, be considerably smaller than that of its orbit round the Sun. Had the ether theory been correct, of course, the speed of the Earth relative to the ether should have changed by twice its orbital speed over a period of six months. The experiment seemed to imply, then, that light always travels at the same speed, c, relative to the apparatus used to observe it.

In his paper of 1905, Einstein makes the fundamental assumption (though he expresses things a little differently) that *light travels with exactly the same speed, c, relative to any inertial frame*. Since this is clearly incompatible with the Galilean transformation law given in (2.1), he takes the remarkable step of modifying this law to read

$$\begin{aligned} x' &= \frac{x - vt}{(1 - v^2/c^2)^{1/2}} & y' &= y \\ z' &= z & t' &= \frac{t - vx/c^2}{(1 - v^2/c^2)^{1/2}}. \end{aligned} \tag{2.2}$$

These equations are known as the *Lorentz transformation*, because a set of equations having essentially this form had been written down by H A Lorentz (1904) in the course of his attempt to explain the results of Michelson and Morley. However, Lorentz believed that his equations described a mechanical effect of the ether upon bodies moving through it, which he attributed to a modification of intermolecular forces. He does not appear to have interpreted them as Einstein did, namely as a general law relating coordinate systems in relative motion. The assumptions that lead to this transformation law are set out in exercise 2.1, where readers are invited to complete its derivation. Here, let us note that (2.2) does indeed embody the assumption that light travels with speed c relative to any inertial frame. For example, if a pulse of light is emitted from the common origin of S and S' at $t = t' = 0$, then the equation of the resulting spherical wavefront at time t relative to S is $x^2 + y^2 + z^2 = c^2t^2$. Using the transformation (2.2), we easily find that its equation at time t' relative to S' is $x'^2 + y'^2 + z'^2 = c^2t'^2$.

Many of the elementary consequences of special relativity follow directly from the Lorentz transformation, and we shall meet some of them in later chapters. What particularly concerns us at present—and what makes Einstein's interpretation of the transformation equations so remarkable—is the change that these equations require us to make in our view of space and time. On the face of it, equations (2.1) or (2.2) simply tell us how to relate observations made in two different frames of reference. At a deeper level, however, they contain information about the structure of space and time that is independent of any frame of reference. Consider two events with spacetime coordinates $(\boldsymbol{x}_1, t_1)$ and $(\boldsymbol{x}_2, t_2)$ relative to S. According to the Galilean transformation, the time interval $t_2 - t_1$ between them relative to S is equal to the interval $t'_2 - t'_1$ relative to S'. In particular, it may happen that these two events are simultaneous, so that $t_2 - t_1 = 0$, and this statement would be equally valid from the point of view of either frame of reference. For two simultaneous events, the spatial distances between them, $|\boldsymbol{x}_1 - \boldsymbol{x}_2|$ and $|\boldsymbol{x}'_1 - \boldsymbol{x}'_2|$ are also equal. Thus, the time interval between two events and the spatial distance between two simultaneous events have the same value in *every* inertial frame, and hence have real physical meanings that are independent of any system of coordinates. According to the Lorentz transformation (2.2), however, both the time interval and the distance have different values relative to different inertial frames. Since these frames are arbitrarily chosen by us, neither the time interval nor the distance has any definite, independent meaning. The one quantity that does have a definite, frame-independent meaning is the *proper time interval* $\Delta\tau$, defined by

$$c^2\Delta\tau^2 = c^2\Delta t^2 - \Delta x^2 \tag{2.3}$$

where $\Delta t = t_2 - t_1$ and $\Delta x = |\boldsymbol{x}_2 - \boldsymbol{x}_1|$. By using (2.2), it is easy to verify that $c^2\Delta t'^2 - \Delta x'^2$ is also equal to $c^2\Delta\tau^2$.

We see, therefore, that the Galilean transformation can be correct only in a *Galilean spacetime*; that is, a spacetime in which both time intervals and spatial distances have well-defined meanings. For the Lorentz transformation to be correct, the structure of space and time must be such that only proper-time intervals are well defined. There are, as we shall see, many such structures. The one in which the Lorentz transformation is valid is called *Minkowski spacetime* after Hermann Minkowski who first clearly described its geometrical properties (Minkowski, 1908). These properties are summarized by the definition (2.3) of proper time intervals. In this definition, the constant c does not refer to the speed of anything. Although it has the dimensions of velocity, its role is really no more than that of a conversion factor between units of length and time. Thus, although the special theory of relativity arose from attempts to understand the propagation of light, it has nothing to do with electromagnetic radiation as such. Indeed, it is not in essence about relativity either! Its essential feature is the structure of space and time expressed by (2.3), and the law for transforming between frames in relative motion serves only as a clue to what this structure is. With this in mind, Minkowski (1908) says of the name 'relativity' that it '... seems to me very feeble'.

The geometrical structure of space and time restricts the laws of motion that may govern the dynamical behaviour of objects that live there. This is true, at least, if one accepts the *principle of relativity*, expressed by Einstein as follows:

> The laws by which the states of physical systems undergo change are not affected, whether these changes of state be referred to the one or the other of two systems of coordinates in uniform translatory motion.
> (Einstein 1905)

Any inertial frame, that is to say, should be as good as any other as far as the laws of physics are concerned. Mathematically, this means that the equations expressing these laws should be *covariant*—they should have the same form in any inertial frame. Consider, for example, two objects, with masses m_1 and m_2, situated at x_1 and x_2 on the x axis of S. According to Newtonian mechanics and the Newtonian theory of gravity, the equation of motion for particle 1 is

$$m_1 \frac{\mathrm{d}^2 x_1}{\mathrm{d}t^2} = (G m_1 m_2) \frac{x_2 - x_1}{|x_2 - x_1|^3} \tag{2.4}$$

where $G \simeq 6.67 \times 10^{-11} \mathrm{N\,m^2\,kg^{-2}}$ is Newton's gravitational constant. If spacetime is Galilean and the transformation law (2.1) is valid, then $\mathrm{d}^2x'/\mathrm{d}t'^2 = \mathrm{d}^2x/\mathrm{d}t^2$ and $(x'_2 - x'_1) = (x_2 - x_1)$, so in S' the equation has exactly the same form and Einstein's principle is satisfied. In Minkowski spacetime, we must use the Lorentz transformation. The acceleration relative to S is not equal to the acceleration relative to S' (see exercise 2.2), but worse is to come! On the right-hand side, x_1 and x_2 refer to two events, namely the objects reaching these two positions, which occur simultaneously as viewed from S. As viewed from S', however, these two events are separated by a time interval $(t'_2 - t'_1) = (x'_1 - x'_2)v/c^2$, as readers may easily verify from (2.2). In Minkowski spacetime, therefore, (2.4) does not respect the principle of relativity. It is unsatisfactory as a law of motion because it implies that there is a preferred inertial frame, namely S, relative to which the force depends only on the instantaneous separation of the two objects; relative to any other frame, it depends on the distance between their positions at different times, and also on the velocity of the frame of reference relative to the preferred one. Actually, we do not know *a priori* that there is no such preferred frame. In the end, we trust the principle of relativity because the theories that stem from it explain a number of observed phenomena for which Newtonian mechanics cannot account.

We might imagine that electrical forces would present a similar problem, since we obtain Coulomb's law for particles with charges q_1 and q_2 merely by replacing the constant in parentheses in (2.4) with $-q_1q_2/4\pi\epsilon_0$. In fact, Maxwell's theory is not covariant under Galilean transformations, but can be made covariant under Lorentz transformations with only minor modifications. We shall deal with electromagnetism in some detail later on, and I do not want to enter into the technicalities at this point. We may note, however, the features that favour Lorentz covariance. In Maxwell's theory, the forces between charged particles are transmitted by electric and magnetic fields. We know that the fields due to a charged particle do indeed appear different in different inertial frames: in a frame in which the particle is at rest, we see only an electric field, while in a frame in which the particle is moving, we also see a magnetic field. Moreover, disturbances in these fields are transmitted at the speed of light. The problem of simultaneity is avoided because a second particle responds not directly to the first one, but rather to the electromagnetic field at its own position. The expression analogous to the right-hand side of (2.4) for the Coulomb force is valid only when there is a frame of reference in which particle 2 can be considered fixed, and then only as an approximation.

2.0.2 The general theory

The experimental fact that eventually led to the special theory was, as we have seen, the constancy of the speed of light. The general theory, and the account that it provides of gravitation, also spring from a crucial fact of observation, namely the equality of inertial and gravitational masses. In (2.4), the mass m_1 appears in two different guises. On the left-hand side, m_1 denotes the *inertial mass*, which governs the response of the body to a given force. On the right-hand side, it denotes the *gravitational mass*, which determines the strength of the gravitational force. The gravitational mass is analogous to the electric charge in Coulomb's law and, since the electrical charge on a body is not necessarily proportional to its mass, there is no obvious reason why the gravitational 'charge' should be determined by the mass either. The equality of gravitational and inertial masses is, of course, responsible for the fact that the acceleration of a body in the Earth's gravitational field is independent of its mass, and this has been familiar since the time of Galileo and Newton. It was checked in 1889 to an accuracy of about one part in 10^9 by Eötvös, whose method has been further refined more recently by R H Dicke and his collaborators.

It seemed to Einstein that this precise equality demanded some explanation, and he was struck by the fact that *inertial forces* such as centrifugal and Coriolis forces are proportional to the inertial mass of the body on which they act. These inertial forces are often regarded as 'fictitious', in the sense that they arise from the use of accelerating (and therefore non-inertial) frames of reference. Consider, for example, a spaceship far from any gravitating bodies such as stars or planets. When its motors are turned off, a frame of reference S fixed in the ship is inertial provided, as we assume, that it is not spinning relative to distant stars. Relative to this frame, the equation of motion of an object on which no forces act is $m\mathrm{d}^2\boldsymbol{x}/\mathrm{d}t^2 = 0$. Suppose the motors are started at time $t = 0$, giving the ship a constant acceleration a in the x direction. S is now not an inertial frame. If S' is the inertial frame that coincided with S for $t < 0$, then the equation of the object is still $m\mathrm{d}^2\boldsymbol{x}'/\mathrm{d}t'^2$=0, at least until the object collides with the cabin walls. Using Galilean relativity for simplicity, we have $x' = x + \frac{1}{2}at^2$ and $t' = t$, so relative to S the equation of motion is

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -ma. \tag{2.5}$$

The force on the right-hand side arises trivially from the coordinate transformation and is definitely proportional to the *inertial* mass.

Einstein's idea is that gravitational forces are of essentially the same kind as that appearing in (2.5), which means that the inertial and gravitational masses are necessarily identical. Suppose that the object in question is in fact a physicist, whose ship-board laboratory is completely soundproof and windowless. His sensation of weight, as expressed by (2.5), is equally consistent with the ship's being accelerated by its motors or with its having landed on a planet at whose surface the acceleration due to gravity is a. Conversely, when he was apparently weightless, he would be unable to tell whether his ship was actually in deep space or freely falling towards a nearby planet. This illustrates Einstein's *principle of equivalence*, according to which the effects of a gravitational field can locally be eliminated by using a freely-falling frame of reference. This frame is inertial and, relative to it, the laws of physics take the same form that they would have relative to any inertial frame in a region far removed from any gravitating bodies.

The word 'locally' indicates that the freely-falling inertial frame can usually extend only over a small region. Let us suppose that our spaceship is indeed falling freely towards a nearby planet. (Readers may rest assured that the pilot, unlike the physicist, is aware of this and will eventually act to avert the impending disaster.) If he has sufficiently accurate apparatus, the physicist can detect the presence of the planet in the following way. Knowing the standard landing procedure, he allows two small objects to float freely on either side of his laboratory, so that the line joining them is perpendicular to the direction in which he knows that the planet, if any, will lie. Each of these objects falls towards the centre of the planet, and therefore their paths slowly converge. As observed in the freely-falling laboratory, they do not accelerate in the direction of the planet, but they do accelerate towards each other, even though their mutual gravitational attraction is negligible. (The tendency of the cabin walls to converge in the same manner is, of course, counteracted by interatomic forces within them.) Strictly, then, the effects of gravity are eliminated in the freely-falling laboratory only to the extent that two straight lines passing through it, which meet at the centre of the planet, can be considered parallel. If the laboratory is small compared with its distance from the centre of the planet, then this will be true to a very good approximation, but the equivalence principle applies exactly only to an infinitesimal region.

The principle of equivalence as stated above is not as innocuous as it might appear. We illustrated it by considering the behaviour of freely-falling objects, and found that it followed in a more or less trivial manner from the equality of gravitational and inertial masses. A version restricted to such situations is sometimes called the *weak* principle of equivalence. The *strong* principle, applying to all the laws of physics, has much more profound implications. It led Einstein to the view that gravity is not a force of the usual kind. Rather, the effect of a massive body is to modify the geometry of space and time. Particles that are not acted on by any ordinary force do not accelerate; they merely appear to be accelerated by gravity if we make the false assumption that the geometry is that of Galilean or Minkowski spacetime and interpret our observations accordingly.

Consider again the expression for proper time intervals given in (2.3). It is valid when (x, y, z, t) refer to Cartesian coordinates in an inertial frame of reference. In the neighbourhood of a gravitating body, a freely-falling inertial frame can be defined only in a small region, so we write it as

$$c^2(\mathrm{d}\tau)^2 = c^2(\mathrm{d}t)^2 - (\mathrm{d}\boldsymbol{x})^2 \tag{2.6}$$

where $\mathrm{d}t$ and $\mathrm{d}\boldsymbol{x}$ are infinitesimal coordinate differences. Now let us make a transformation to an arbitrary system of coordinates (x^0, x^1, x^2, x^3), each new coordinate being expressible as some function of x, y, z and t. Using the chain rule, we find that (2.6) becomes

$$c^2(\mathrm{d}\tau)^2 = \sum_{\mu,\nu=0}^{3} g_{\mu\nu}(x)\mathrm{d}x^\mu \mathrm{d}x^\nu \tag{2.7}$$

where the functions $g_{\mu\nu}(x)$ are given in terms of the transformation functions. They are components of what is called the *metric tensor*. In the usual version of general relativity, it is the metric tensor that embodies all the geometrical structure of space and time. Suppose we are given a set of functions $g_{\mu\nu}(x)$ which describe this structure in terms of some system of coordinates $\{x^\mu\}$. According to the principle of equivalence, it is possible at any point (say X, with coordinates X^μ) to construct a freely falling inertial frame, valid in a small neighbourhood surrounding X, relative to which there are no gravitational effects and all other processes occur as in special relativity. This means that it is possible to find a set of coordinates (ct, x, y, z) such that the proper time interval (2.7) reverts to the form of (2.6). Using a matrix representation of the metric tensor, we can write

$$g_{\mu\nu}(X) = \eta_{\mu\nu} \equiv \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{2.8}$$

where $\eta_{\mu\nu}$ is the special metric tensor corresponding to (2.6).

If the geometry is that of Minkowski spacetime, then it will be possible to choose (ct, x, y, z) in such a way that $g_{\mu\nu} = \eta_{\mu\nu}$ everywhere. Otherwise, the best we can usually do is to make $g_{\mu\nu} = \eta_{\mu\nu}$ at a single point (though that point can be anywhere) or at every point along a curve, such as the path followed by an observer. Even when we do not have a Minkowski spacetime, it may be possible to set up an approximately inertial and approximately Cartesian coordinate system such that $g_{\mu\nu}$ differs only a little from $\eta_{\mu\nu}$ throughout a large region. In such a case, we can do much of our physics successfully by assuming that spacetime is exactly Minkowskian. If we do so, then, according to general relativity, we shall interpret the slight deviations from the true Minkowski metric as gravitational forces.

This concludes our introductory survey of the theories of relativity. We have concentrated on the ways in which our common-sense ideas of spacetime geometry must be modified in order to accommodate two key experimental observations: the constancy of the speed of light and the equality of gravitational and inertial masses. It is clear that the modified geometry leads to modifications in the laws that govern the behaviour of physical systems, but we have not discussed these laws in concrete terms. That we shall be better equipped to do after we have developed some mathematical tools in the remainder of this chapter. At that stage, we shall be able to see much more explicitly how gravity arises from geometry.

2.1 Spacetime as a Differentiable Manifold

Our aim is to construct a mathematical model of space and time that involves as few assumptions as possible, and to be explicitly aware of the assumptions we do make. In particular, we have seen that the theories of relativity call into question the meanings we attach to distances and time intervals, and we need to be clear about these. The mathematical structure that has proved to be a suitable starting point, at least for a non-quantum-mechanical model of space and time, is called a *differentiable manifold*. It is a collection of *points*, each of which will eventually correspond to a unique position in space and time, and the whole collection comprises the entire history of our model universe. It has two key features that represent familiar facts about our experience of space and time. The first is that any point can be uniquely specified by a set of four real numbers, so spacetime is four-dimensional. For the moment, the exact number of dimensions is not important. Later on, indeed, we shall encounter some recent theories which suggest that there may be more than four, the extra ones being invisible to us. Even in more conventional theories, we shall find that it is helpful to consider other numbers of dimensions as a purely mathematical device. The second feature is a kind of 'smoothness', meaning roughly that, given any two distinct points, there are more points in between them. This feature allows us to describe physical quantities such as particle trajectories or electromagnetic fields in terms of differentiable functions and hence to do theoretical physics of the usual kind. We do not know for certain that space and time are quite as smooth as this, but at least there is no evidence for any granularity down to the shortest distances we are able to probe experimentally.

Our first task is to express these properties in a more precise mathematical form. It is of fundamental importance that this can be done without recourse to any notion of length. The properties we require are *topological* ones, and we begin by introducing some elementary ideas of topology. Roughly speaking, we want to be able to say that some pairs of points are 'closer together' than others, without having any quantitative measure of distance. As an illustration, consider a sheet of rubber, marked off into different regions as in figure 2.2. For the purposes of this illustration, we shall say that there is no definite distance between two points

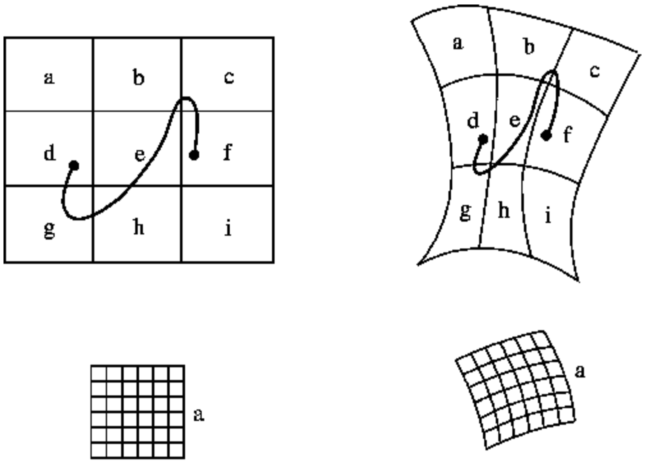
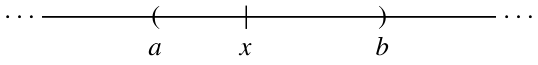


Figure 2.2. A deformable sheet of rubber, divided into several regions. Although there is no definite distance between the points indicated by ● , there are always other points between them, because any curve joining them must pass through at least one of the regions b, e and h.

on the sheet, because it can be deformed at will. No matter how it is deformed, however, any given region is still surrounded by the same neighbouring regions. Given a point in d and another in f, we can never draw a line between them that does not pass through at least one of regions b, e and h. The same holds, moreover, of more finely subdivided regions, as shown for subdivisions of a, each of which could be further subdivided, and so on. In this sense, points on the sheet are smoothly connected together. The smoothness would be lost if the rubber were vaporized, the individual molecules being considered as the collection of points. Mathematically, the kind of smoothness we want is a property of the real line (that is, the set of all real numbers, denoted by $\mathbb{R}$). So, as part of the definition of the manifold, we demand that it should be possible to set up correspondences (called 'maps') between points of the manifold and sets of real numbers. We shall next look at the topological properties of real numbers, and then see how we can ensure that the manifold shares them.

2.1.1 Topology of the real line $\mathbb{R}$ and of $\mathbb{R}^d$

The topological properties we are interested in are expressed in terms of 'open sets', which are defined in the following way. An *open interval* (a, b) is the set of all points (real numbers) x such that $a < x < b$:



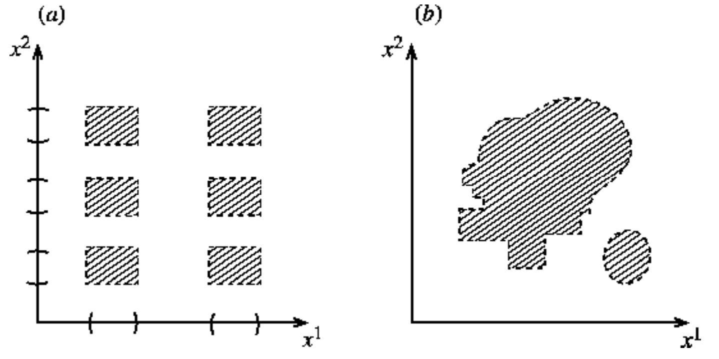
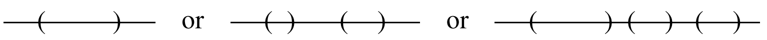


Figure 2.3. (a) An open set in $\mathbb{R}^2$. It is a union of open rectangles constructed from unions of open intervals in the two copies of $\mathbb{R}$ which form the x^1 and x^2 axes. (b) Another open set in $\mathbb{R}^2$, which can be constructed as a union of open rectangles.

The end points $x = a$ and $x = b$ are excluded. Consequently, *any* point x in (a, b) can be surrounded by another open interval $(x - \epsilon, x + \epsilon)$, all of whose points are also in (a, b). For example, however close x is to a, it cannot be equal to a. There are always points between a and x, and if x is closer to a than to b, we can take $\epsilon = (x - a)/2$. An *open set* of $\mathbb{R}$ is defined as any union of 1, 2, 3, ... open intervals:



etc. (The *union* $A \cup B \cup C \cdots$ of a number of sets is defined as the set of all points that belong to at least one of $A, B, C, \ldots$. The *intersection* $A \cap B \cap C \cdots$ is the set of all points that belong to all the sets $A, B, C, \ldots$.) In addition, the empty set, which contains no points, is defined to be an open set.

The space $\mathbb{R}^2$ is the set of all pairs of real numbers (x^1, x^2), which can be envisaged as an infinite plane. The definition of open sets is easily extended to $\mathbb{R}^2$, as illustrated in figure 2.3. If x^1 lies in a chosen open interval on the horizontal axis, and x^2 in a chosen open interval on the vertical axis, then (x^1, x^2) lies in an open rectangle corresponding to these two intervals. Any union of open rectangles is an open set. Since the rectangles can be arbitrarily small, we can say that any region bounded by a closed curve, but excluding points actually on the curve, is also an open set, and so is any union of such regions. Obviously, the same ideas can be further extended to $\mathbb{R}^d$, which is the set of all d-tuples of real numbers $(x^1, x^2, \ldots, x^d)$.

An important use of open sets is to define continuous functions. Consider, for instance, a function f which takes real numbers x as arguments and has real-number values $y = f(x)$. An example is shown in figure 2.4. The *inverse image* of a set of points on the y axis is the set of all those points on the x axis for which $f(x)$ belongs to the original set. Then we say that f is continuous if the

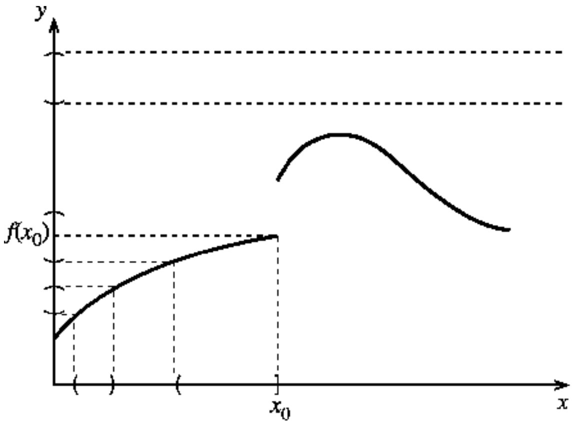


Figure 2.4. The graph $y = f(x)$ of a function which is discontinuous at x_0. Any open interval of y which includes $f(x_0)$ has an inverse image on the x axis which is not open. The inverse image of an interval in y which contains no values of $f(x)$ is the empty set.

inverse image of any open set on the y axis is an open set on the x axis. The example shown fails to be continuous because the inverse image of any open interval containing $f(x_0)$ contains an interval of the type $(x_1, x_0]$, which includes the end point x_0 and is therefore not open. (Readers who are not at home with this style of argument should spend a short while considering the implications of these definitions: why, for example, is it necessary to include not only open intervals but also their unions and the empty set as open sets?)

The open sets of $\mathbb{R}^d$ have two fairly obvious properties: (i) any union of open sets is itself an open set; (ii) any intersection of a finite number of open sets is itself an open set. Given any space (by which we mean a set of points), suppose that a collection of subsets of its points is specified, such that any union or finite intersection of them also belongs to the collection. We also specify that the entire space (which counts as a subset of itself) and the empty set belong to the collection. Then the subsets in this collection may, by analogy, be called *open sets*. The collection of open sets is called a *topology* and the space, together with its topology, is called a *topological space*. It is, of course, possible to endow a given space with many different topologies. For example, the collection of all subsets of the space clearly satisfies all the above conditions, and is called the *discrete topology*. By endowing the real line with this topology, we would obtain a new definition of continuity—it would not be a particularly useful definition, however, as any function at all would turn out to be continuous. The particular topology of $\mathbb{R}^d$ described above is called the *natural topology* and is the one we shall always use.

It is important to realize that a topology is quite independent of any notion of distance. For instance, a sheet of paper may be regarded as a part of $\mathbb{R}^2$.

The natural topology reflects the way in which its points fit together to form a coherent structure. If it is used to draw figures in Euclidean geometry, then the distance D between two points is defined by the Pythagoras rule as $D = \left[(\Delta x)^2 + (\Delta y)^2\right]^{1/2}$. But it might equally well be used to plot the mean atmospheric concentration of carbon monoxide in central London (represented by y) as a function of time (represented by x), in which case D would have no sensible meaning.

A topology imposes two kinds of structure on the space. The *local topology*—the way in which open sets fit inside one another over small regions—determines the way in which notions like continuity apply to the space. The *global topology*—the way in which the open sets can be made to cover the whole space—determines its overall structure. Thus, the plane, sphere and torus have the same local structure but different global structures. Physically, we have no definite information about the global topology of spacetime, but its local structure seems to be very similar to that of $\mathbb{R}^4$ (though we shall encounter speculative theories that call this apparently simple observation into question).

2.1.2 Differentiable spacetime manifold

In order that our model of space and time should be able to support continuous and differentiable functions of the sort that we rely on to do physics, we want it (for now) to have the same local topology as $\mathbb{R}^4$. First of all, then, it must be a topological space. That is, it must have a collection of open sets, in terms of which continuous functions can be defined. Second, the structure of these open sets must be similar, within small regions, to the natural topology of $\mathbb{R}^4$. To this end, we demand that every point of the space belong to at least one open set, all of whose points can be put into a one-to one correspondence with the points of some open set of $\mathbb{R}^4$. More technically, the correspondence is a one-to-one mapping of the open set of the space *onto* the open set of $\mathbb{R}^4$, which is to say that every point of the open set in the space has a unique image point in the open set of $\mathbb{R}^4$ and *vice versa*. We further demand that this mapping be continuous, according to our previous definition. When these conditions are met, the space is called a *manifold*. The existence of continuous mappings between the manifold and $\mathbb{R}^4$ implies that a function f defined on the manifold (that is, one that has a value $f(P)$ for each point P of the manifold) can be re-expressed as a function g defined on $\mathbb{R}^4$, so that $f(P) = g(x^0, \ldots, x^3)$, where $(x^0, \ldots, x^3)$ is the point of $\mathbb{R}^4$ corresponding to P. In this way, continuous functions defined on the manifold inherit the characteristics of those defined on $\mathbb{R}^4$.

This definition amounts to saying that the manifold can be covered by patches, in each of which a four-dimensional coordinate system can be set up, as illustrated in figure 2.5 for the more easily drawn case of a two-dimensional manifold. Normally, of course, many different coordinate systems can be set up on any part of the manifold. The definition also ensures that, within the range of coordinate values corresponding to a given patch, there exists a point of the

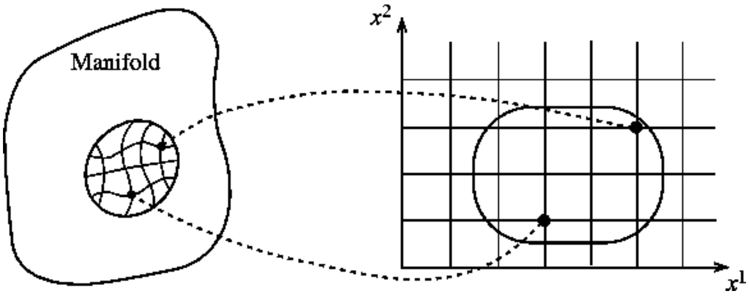


Figure 2.5. A coordinate patch on a two-dimensional manifold. Each point in the patch is mapped to a unique image point in a region of $\mathbb{R}^2$ and *vice versa*.

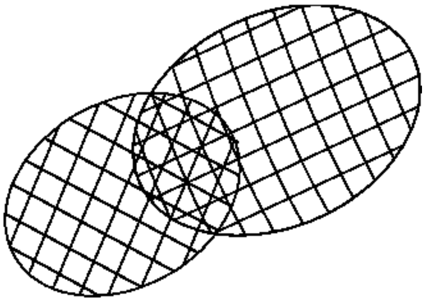


Figure 2.6. Two overlapping coordinate patches. A point in the overlap region can be identified using either set of coordinates.

manifold for each set of coordinate values—so there are no points 'missing' from the manifold, and also that there are no 'extra' points that cannot be assigned coordinates. Within a coordinate patch, a quantity such as an electric potential, which has a value at each point of the manifold, can be expressed as an ordinary function of the coordinates of the point. Often, we shall expect such functions to be *differentiable* (that is, to possess unique partial derivatives with respect to each coordinate at each point of the patch).

Suppose we have two patches, each with its own coordinate system, that partly or wholly overlap, as in figure 2.6. Each point in the overlap region has two sets of coordinates, say $(x^0, \ldots, x^3)$ and $(y^0, \ldots, y^3)$, and the y coordinates can be expressed as functions of the x coordinates: $y^0 = y^0(x^0, \ldots, x^3)$, etc. Given 'reasonable' coordinate systems, we might suppose that a function which is differentiable when expressed in terms of the x^μ ought also to be differentiable when expressed in terms of the y^μ. This will indeed be true if the transformation functions $y^\mu(x)$ are differentiable. If the manifold can be completely covered by a set of coordinate patches, in such a way that all of these transformation functions are differentiable, then we have a *differentiable manifold*. In order for a function

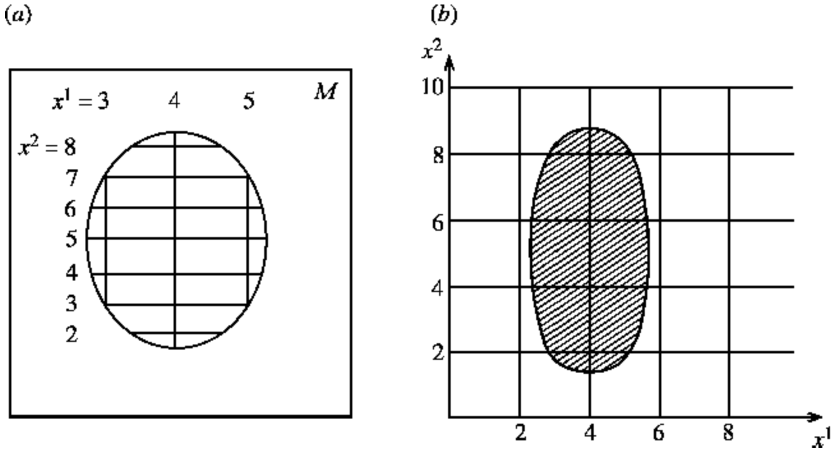


Figure 2.7. (*a*) A manifold M, part of the surface of this page, with a coordinate patch. (*b*) Part of $\mathbb{R}^2$, showing the coordinate values used in (*a*).

to remain differentiable at least n times after a change of coordinates, at least the first n derivatives of all the transformation functions must exist. If they do, then we have what is called a C^n manifold. Intuitively, we might think it possible to define functions of space and time that can be differentiated any number of times, for which we would need $n = \infty$. We shall indeed take a C^∞ manifold as the basis for our model spacetime. Mathematically, though, this is a rather strong assumption, and for many physical purposes it would be sufficient to take, say, $n = 4$.

2.1.3 Summary and examples

Our starting point for a model of space and time is a C^∞ manifold. The essence of the technical definition described above is, first, that it is possible to set up a local coordinate system covering any sufficiently 'small' region and, second, that it is possible to define functions on the manifold that are continuous and differentiable in the usual sense. It is, of course, perfectly possible to define functions that are neither continuous nor differentiable. The point is that, if a function fails to be continuous or differentiable, this will be the fault either of the function itself or of our choice of coordinates, but not the fault of the manifold. The word 'small' appears in inverted commas because, as I have emphasized, there is as yet no definite notion of length: it simply means that it may well not be possible to cover the entire manifold with a single coordinate system. The coordinate systems themselves are not part of the structure of the manifold. They serve merely as an aid to thought, providing a practical means of specifying properties of sets of points belonging to the manifold.

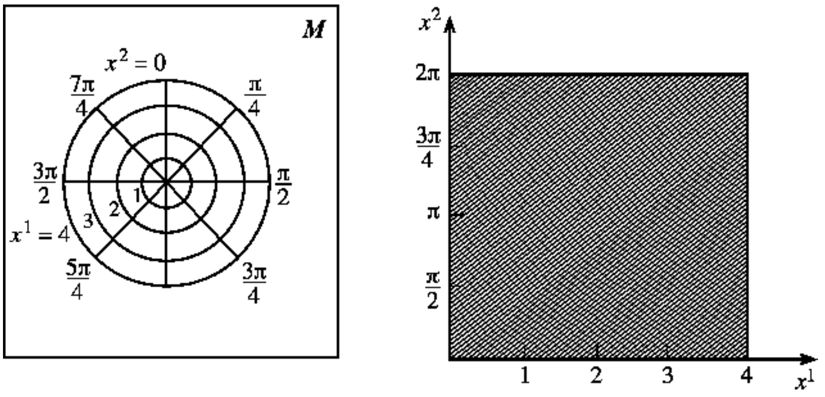


Figure 2.8. Same as figure 2.7, but using different coordinates.

The following examples illustrate, in terms of two-dimensional manifolds, some of the important ideas. Figure 2.7(a) shows a manifold, M, which is part of the surface of the paper on which it is printed. For the sake of argument, I am asking readers to suppose that this surface is perfectly smooth, rather than being composed of tiny fibres. For the definitions to work, we must take the manifold to be the interior of the rectangular region, excluding points *on* the boundary. The interior of the roughly circular region is a coordinate patch. Inside it are drawn some of the grid lines by means of which we assign coordinates x^1 and x^2 to each point. Figure 2.7(b) is a pictorial representation of part of the space $\mathbb{R}^2$ of pairs of coordinates. The interior of the shaded region represents the coordinates actually used. To every point of this region there corresponds a point of the coordinate patch in M and *vice versa*. Figure 2.8 shows a similar arrangement, using a different coordinate system. Here, again, the *interior* of the shaded region of $\mathbb{R}^2$ represents the open set of points that correspond uniquely to points of the coordinate patch. As before, the boundary of the coordinate patch and the corresponding line $x^1 = 4$ in $\mathbb{R}^2$ are excluded. Also excluded, however, are the boundary lines $x^1 = 0$, $x^2 = 0$ and $x^2 = 2\pi$ in $\mathbb{R}^2$, which means that points on the line labelled by $x^2 = 0$ in M do not, in fact, belong to the coordinate patch. Since the coordinate system is obviously usable, even when these points are included, their exclusion may seem like an annoying piece of bureaucracy: however, it is essential to apply the rules correctly if the definitions of continuity and differentiability are to work smoothly. For example, the function $g(x^1, x^2) = x^2$ is continuous throughout $\mathbb{R}^2$, but the corresponding function on M is discontinuous at $x^2 = 0$.

It should be clear that, whereas a single coordinate patch like that in figure 2.7 can be extended to cover the whole of M, at least two patches of the kind shown in figure 2.8 would be needed. Readers should also be able to convince themselves that, if M were the two-dimensional surface of a sphere, no single patch of any kind could cover all of it. These examples also illustrate the fact that, although the coordinates which label the points of M have definite numerical values, these values do not, in themselves, supply any notion of a distance between two points. The distance along some curve in M *may* be defined by some suitable rule, such as (i) 'use a ruler' or (ii) 'measure the volume of ink used by a standard pen to trace the curve' or, given a particular coordinate system, (iii) 'use the mathematical expression $D =$ (function of coordinates)'. Any such rule imposes an additional structure—called a *metric*—which is not inherent in the manifold. In particular, there is no naturally occurring function for use in (iii). Any specific function, such as the Pythagoras expression, would have quite different effects when applied to different coordinate systems, and the definition of the manifold certainly does not single out a special coordinate system to which that function would apply. We do have a more or less unambiguous means of determining distances on a sheet of paper, and this is because the paper, in addition to the topological properties it possesses as a manifold, has physical properties that enable us to apply a definite measuring procedure. The same is true of space and time and, although we have made some initial assumptions about their topological structure, we have yet to find out what physical properties determine their metrical structure.

2.2 Tensors

From our discussion so far, it is apparent that coordinate systems can be dangerous, even though they are often indispensable for giving concrete descriptions of a physical system. We have seen that the topology of a manifold such as that of space and time may permit the use of a particular coordinate system only within a small patch. Suppose, for the sake of argument, that the surface of the Earth is a smooth sphere. We encounter no difficulty in drawing, say, the street plan of a city on a flat sheet of paper using Cartesian coordinates, but we should obviously be misled if we assumed that this map could be extended straightforwardly to cover the whole globe. By assuming that two-dimensional Euclidean geometry was valid on the surface of the Earth, we should be making a mistake, owing to the curvature of the spherical surface, but the mistake would not become apparent as long as we made measurements only within a region the size of a city. Likewise, physicists before Einstein assumed that a frame of reference fixed on the Earth would be inertial, except for effects of the known orbital motion of the Earth around the Sun and its rotation about its own axis, which could be corrected for if necessary. According to Einstein, however, this assumption is also mistaken. It fails to take account of the true geometry of space and time in much the same way that, by treating a city plan as a Euclidean plane, we fail to take account of the true geometry of the Earth. The mistake only becomes apparent, however, when we make precise observations of gravitational phenomena.

The difficulty here is that we often express the laws of physics in the form which, we believe, applies to inertial frames. If we do not know, *a priori*, what the true geometry of space and time is, then we do not know whether any given frame is truly inertial. Therefore, we need to express our laws in a way that does not rely on our making any special assumption about the coordinate system. There are two ways of achieving this. The method adopted by Einstein himself is to write our equations in a form that applies to *any* coordinate system: the mathematical techniques for doing this constitute what is called *tensor analysis*. The other, more recent method is to write them in a manner that makes no reference to coordinate systems at all: this requires the techniques of *differential geometry*. For our purposes, these two approaches are entirely equivalent, but each has its own advantages and disadvantages in terms of conceptual and notational clarity. So far as I can, I will follow a middle course, which seems to me to maximize the advantages. Both techniques deal with objects called *tensors*. Tensor analysis, like elementary vector analysis, treats them as being defined by sets of components, referred to particular coordinate systems. Differential geometry treats them as entities in their own right, which may be described in terms of components, but need not be. When components are used, the two techniques become identical, so there is no difficulty in changing from one description to the other.

Many, though not all, of the physical objects that inhabit the spacetime manifold will be described by tensors. A *tensor* at a point P of the manifold refers only to that point. A *tensor field* assigns some property to every point of the manifold, and most physical quantities will be described by tensor fields. (For brevity, I shall often follow custom by referring to a tensor field simply as a 'tensor', when the meaning is obvious from the context.) Tensors and tensor fields are classified by their *rank*, a pair of numbers $\binom{a}{b}$.

Rank $\binom{0}{0}$ tensors, also called *scalars*, are simply real numbers. A *scalar field* is a real-valued function, say $f(P)$, which assigns a real number to each point of the manifold. If our manifold were just the three-dimensional space encountered in Newtonian physics, then at a particular instant in time, an electric potential $V(P)$ or the density of a fluid $\rho(P)$ would be examples of scalar fields. In relativistic physics, these and all other simple examples I can think of are not true scalars, because their definitions depend in one way or another on the use of specific coordinate systems or on metrical properties of the space that our manifold does not yet possess. For the time being, however, no great harm will be done if readers bear these examples in mind. If we introduce coordinates x^{μ}, then we can express $f(P)$ as an algebraic function $f(x^{\mu})$. (For convenience, I am using the same symbol f to denote two different, though related functions: we have $f(x^{\mu}) = f(P)$ when x^{μ} are the coordinates of the point P.) In a different coordinate system, where P has the coordinates $x^{\mu'}$, the same quantity will be described by a new algebraic function $f'(x^{\mu'})$, related to the old one by

$$f'(x^{\mu'}) = f(x^{\mu}) = f(P). \tag{2.9}$$

In tensor analysis, this transformation law is taken to *define* what is meant by a scalar field.

Rank $\binom{1}{0}$ tensors are called *vectors* in differential geometry. They correspond to what are called *contravariant vectors* in tensor analysis. The prototypical vector is the tangent vector to a curve. In ordinary Euclidean geometry, the equation of a curve may be expressed parametrically by giving three functions $x(\lambda)$, $y(\lambda)$ and $z(\lambda)$, so that each point of the curve is labelled by a value of λ and the functions give its coordinates. If λ is chosen to be the distance along the curve from a given starting point, then the tangent vector to the curve at the point labelled by λ has components $(\mathrm{d}x/\mathrm{d}\lambda, \mathrm{d}y/\mathrm{d}\lambda, \mathrm{d}z/\mathrm{d}\lambda)$. In our manifold, we have not yet given any meaning to 'distance along the curve', and we want to avoid defining vectors in terms of their components relative to a specific coordinate system. Differential geometry provides the following indirect method of generalizing the notion of a vector to any manifold. Consider, in Euclidean space, a differentiable function $f(x, y, z)$. This function has, in particular, a value $f(\lambda)$ at each point of the curve, which we obtain by substituting for x, y and z the appropriate functions of λ. The rate of change of f with respect to λ is

$$\frac{\mathrm{d}f}{\mathrm{d}\lambda} = \frac{\mathrm{d}x}{\mathrm{d}\lambda}\frac{\partial f}{\partial x} + \frac{\mathrm{d}y}{\mathrm{d}\lambda}\frac{\partial f}{\partial y} + \frac{\mathrm{d}z}{\mathrm{d}\lambda}\frac{\partial f}{\partial z} \tag{2.10}$$

so, by choosing $f = x$, $f = y$ or $f = z$, we can recover from this expression each component of the tangent vector. All the information about the tangent vector is contained in the differential operator $\mathrm{d}/\mathrm{d}\lambda$, and in differential geometry this operator is defined to *be* the tangent vector.

A little care is required when applying this definition to our manifold. We can certainly draw a continuous curve on the manifold and label its points continuously by a parameter λ. What we cannot yet do is select a special parameter that measures distance along it. Clearly, by choosing different parametrizations of the curve, we shall arrive at different definitions of its tangent vectors. It is convenient to refer to the one-dimensional set of points in the manifold as a *path*. Then each path may be parametrized in many different ways, and we regard each parametrization as a distinct *curve*. This has the advantage that each curve, with its parameter λ, has a unique tangent vector $\mathrm{d}/\mathrm{d}\lambda$ at every point. Suppose we have two curves, corresponding to the same path, but with parameters λ and μ that are related by $\mu = a\lambda + b$, a and b being constants. The difference is obviously a rather trivial one and the two parameters are said to be *affinely related*.

If we now introduce a coordinate system, we can resolve a vector into components, in much the same way as in Euclidean geometry. At this point, it is useful to introduce two abbreviations into our notation. First, we use the symbol ∂_μ to denote the partial derivative $\partial/\partial x^\mu$. Second, we shall use the *summation convention*, according to which, if an index such as μ appears in an expression twice, once in the upper position and once in the lower position, then a sum over the values $\mu = 0 \ldots 3$ is implied. (More generally, in a d-dimensional manifold, the sum is over the values $0 \ldots (d-1)$. In contexts other than spacetime geometry, there may be no useful distinction between upper and lower indices, and repeated indices implying a sum may both appear in the same position.) I shall use bold capital letters to denote vectors, such as $\boldsymbol{V} = \mathrm{d}/\mathrm{d}\lambda$. If, then, a curve is represented in a particular coordinate system by the functions $x^\mu(\lambda)$, we can write

$$\boldsymbol{V} \equiv \frac{\mathrm{d}}{\mathrm{d}\lambda} = \sum_{\mu=0}^{3} \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} \frac{\partial}{\partial x^\mu} \equiv V^\mu \partial_\mu \equiv V^\mu \boldsymbol{X}_\mu \tag{2.11}$$

where the partial derivatives $\boldsymbol{X}_\mu = \partial/\partial x^\mu$ are identified as the basis vectors in this system and V^μ are the corresponding components of $\boldsymbol{V}$. Note that components of a vector are labelled by upper indices and basis vectors by lower ones. In a new coordinate system, with coordinates $x^{\mu'}$, and basis vectors $\boldsymbol{X}_{\mu'} = \partial/\partial x^{\mu'}$, the chain rule $\partial_\mu = (\partial x^{\mu'}/\partial x^\mu)\partial_{\mu'}$ shows that the same vector has components

$$V^{\mu'} = \frac{\partial x^{\mu'}}{\partial x^\mu} V^\mu. \tag{2.12}$$

In tensor analysis, a contravariant vector is defined by specifying its components in some chosen coordinate system and requiring its components in any other system to be those given by the transformation law (2.12). It will be convenient to denote the transformation matrix by

$$\Lambda^{\mu'}{}_{\mu} = \frac{\partial x^{\mu'}}{\partial x^\mu}. \tag{2.13}$$

The convention of placing a prime on the index μ' to indicate that x^μ and $x^{\mu'}$ belong to different coordinate systems, rather than writing, say, x'^μ, is useful here in indicating to which system each index on Λ refers. Using the chain rule again, we find

$$\Lambda^{\mu}{}_{\nu'}\Lambda^{\nu'}{}_{\sigma} = \frac{\partial x^\mu}{\partial x^{\nu'}} \frac{\partial x^{\nu'}}{\partial x^\sigma} = \frac{\partial x^\mu}{\partial x^\sigma} = \delta^\mu_\sigma \tag{2.14}$$

so the matrix $\Lambda^{\mu}{}_{\nu'}$ is the inverse of the matrix $\Lambda^{\nu'}{}_{\mu}$.

Rank $\binom{0}{1}$ tensors are called *one-forms* in differential geometry or *covariant vectors* in tensor analysis. Consider the scalar product $\boldsymbol{u} \cdot \boldsymbol{v}$ of two Euclidean vectors. Normally, we regard this product as a rule that combines two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ to produce a real number. As we shall see, this scalar product involves metrical properties of Euclidean space that our manifold does not yet possess. There is, however, a different point of view that can be transferred to manifold. For a given vector $\boldsymbol{u}$, the symbol $\boldsymbol{u}\cdot$ can be regarded as defining a *function*, whose argument is a vector, say $\boldsymbol{v}$, and whose value is the real number $\boldsymbol{u} \cdot \boldsymbol{v}$. The function $\boldsymbol{u}\cdot$ is linear. That is to say, if we give it the argument $a\boldsymbol{v} + b\boldsymbol{w}$, where $\boldsymbol{v}$ and $\boldsymbol{w}$ are any two vectors, and a and b are any two real numbers, then $\boldsymbol{u} \cdot (a\boldsymbol{v} + b\boldsymbol{w}) = a\boldsymbol{u} \cdot \boldsymbol{v} + b\boldsymbol{u} \cdot \boldsymbol{w}$. This is, in fact, the definition of a one-form. In our manifold, a one-form, say ω, is a real-valued, linear function whose argument is a vector: $\omega(\boldsymbol{V}) =$ (real number). Because the one-form is a linear function, its value must be a linear combination of the components of the vector:

$$\omega(\boldsymbol{V}) = \omega_\mu V^\mu. \tag{2.15}$$

The coefficients ω_μ are the components of the one-form, relative to the coordinate system in which $\boldsymbol{V}$ has components V^μ. A *one-form field* is defined in the same way as a linear function of vector fields, whose value is a scalar field. In the definition of linearity, a and b may be any two scalar fields.

The expression (2.15) is, of course, similar to the rule for calculating the scalar product of two Euclidean vectors from their components. Nevertheless, it is clear from their definitions that vectors and one-forms are quite different things, and (2.15) does not allow us to form a scalar product of two vectors.

An example of a one-form field is the gradient of a scalar field f, whose components are $\partial_\mu f$. Notice the consistency of the convention for placing indices: the components of a one-form have indices that naturally appear in the lower position. Call this gradient one-form ω_f. If $\boldsymbol{V} = \mathrm{d}/\mathrm{d}\lambda$ is the tangent vector to a curve $x^\mu(\lambda)$, then the new scalar field $\omega_f(\boldsymbol{V})$ is the rate of change of f along the curve:

$$\omega_f(\boldsymbol{V}) = \frac{\partial f}{\partial x^\mu}\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} = \frac{\mathrm{d}f}{\mathrm{d}\lambda}. \tag{2.16}$$

Since vectors and one-forms exist independently of any coordinate system, the function $\omega(\boldsymbol{V})$ given in (2.15) must be a true scalar field—it must have the same value in any coordinate system. This means that the matrix which transforms the components of a one-form between two coordinate systems must be the inverse of that which transforms the components of a vector:

$$\omega_{\mu'} = \omega_\mu \Lambda^\mu{}_{\mu'} = \omega_\mu \frac{\partial x^\mu}{\partial x^{\mu'}}. \tag{2.17}$$

Then, on transforming (2.15), we get

$$\omega(\boldsymbol{V}) = \omega_{\mu'} V^{\mu'} = \omega_\mu \Lambda^\mu{}_{\mu'} \Lambda^{\mu'}{}_\nu V^\nu = \omega_\mu \delta^\mu{}_\nu V^\nu = \omega_\mu V^\mu. \tag{2.18}$$

In tensor analysis, a covariant vector is defined by requiring that its components obey the transformation law (2.17). Clearly, this is indeed the correct way of transforming a gradient.

Rank $\binom{a}{b}$ tensors and tensor fields can be defined in a coordinate-independent way, making use of the foregoing definitions of vectors and one-forms, and I shall say more about this in §3.7. For our present purposes, however, it becomes rather easier at this point to adopt the tensor analysis approach of defining higher-rank tensors in terms of their components. A tensor of *contravariant rank a* and *covariant rank b* has, in a d-dimensional manifold, d^{a+b} components, labelled by a upper indices and b lower ones. The tensor may be specified by giving all of its components relative to some chosen coordinate system. In any other system, the components are then given by a transformation law that generalizes those for vectors and one-forms in an obvious way:

$$T^{\alpha'\beta'\cdots}{}_{\mu'\nu'\cdots} = \Lambda^{\alpha'}{}_{\alpha}\Lambda^{\beta'}{}_{\beta}\cdots\Lambda^{\mu}{}_{\mu'}\Lambda^{\nu}{}_{\nu'}\cdots T^{\alpha\beta\cdots}{}_{\mu\nu\cdots}. \qquad (2.19)$$

From this we can see how to construct laws of physics in a way that will make them true in any coordinate system. Suppose that a fact about some physical system is expressed in the form $S = T$, where S and T are tensors of the same rank. On multiplying this equation on both sides by the appropriate product of Λ matrices, we obtain the equation $S' = T'$, which expresses the same fact, in an equation of the same form, but now applies to the new coordinate system. The point that may require some effort is to make sure that S and T really *are* tensors that transform in the appropriate way.

If ω is a one-form and $\boldsymbol{V}$ a vector, then the d^2 quantities $T_{\mu}^{\nu} = \omega_{\mu}V^{\nu}$ are the components of a rank $\binom{1}{1}$ tensor. As we saw in (2.15), by setting $\mu = \nu$ and carrying out the implied sum, we obtain a single number, which is a scalar (or a rank $\binom{0}{0}$ tensor). This process is called *contraction*. Given any tensor of rank $\binom{a}{b}$, with $a \geq 1$ and $b \geq 1$, we may contract an upper index with a lower one to obtain a new tensor of rank $\binom{a-1}{b-1}$. Readers should find it an easy matter to check from (2.19) that, for example, the object $S^{\alpha\gamma\cdots}{}_{\nu\cdots} = T^{\alpha\beta\gamma\cdots}{}_{\beta\nu\cdots}$ does indeed transform in the right way.

2.3 Extra Geometrical Structures

Two geometrical structures are needed to endow our manifold with the familiar properties of space and time: (i) the notion of *parallelism* is represented mathematically by an *affine connection*; (ii) the notions of *length* and *angle* are represented by a *metric*. In principle, these two structures are quite independent. In Euclidean geometry, of course, it is perfectly possible to define what we mean by parallel lines in terms of distances and angles, and this is also true of the structures that are most commonly used in general-relativistic geometry. Thus there is, as we shall see, a special kind of affine connection that can be deduced from a metric. It is called a *metric connection* (or sometimes, the *Levi-Civita connection*). We shall eventually assume that the actual geometry of space and time is indeed described by a metric connection. From a theoretical point of view, however, it is instructive to understand the distinction between those geometrical ideas that rely only on an affine connection and those that require a metric. Moreover, there are manifolds other than spacetime that play important roles in physics (in particular, those connected with the gauge theories of particle physics), which possess connections, but do not necessarily possess metrics. To emphasize this point, therefore, I shall deal first with the affine connection, then with the metric, and finally with the metric connection.

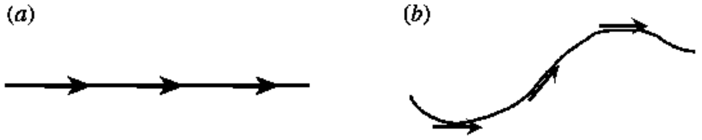


Figure 2.9. (*a*) A geodesic curve: successive tangent vectors are parallel to each other. (*b*) A non-geodesic curve: successive tangent vectors are not parallel.

2.3.1 The affine connection

There are four important geometrical tools provided by an affine connection: the notion of *parallelism*, the notion of *curvature*, the *covariant derivative* and the *geodesic*. Let us first understand what it is good for.

a) Newton's first law of motion claims that 'a body moves at constant speed in a straight line unless it is acted on by a force'. In general relativity, we shall replace this with the assertion that 'a test particle follows a geodesic curve unless it is acted on by a non-gravitational force'. As we saw earlier, gravitational forces are going to be interpreted in terms of spacetime geometry, which itself is modified by the presence of gravitating bodies. By a 'test particle', we mean one that responds to this geometry, but does not modify it significantly. A *geodesic* is a generalization of the straight line of Euclidean geometry. It is defined, roughly, as a curve whose tangent vectors at successive points are parallel, as illustrated in figure 2.9. Given a definition of 'parallel', as provided by the connection, this is perhaps intuitively recognizable as the natural state of motion for a particle that is not disturbed by external influences.

b) The equations of physics, which we wish to express entirely in terms of tensors, frequently involve the derivatives of vector or tensor fields. Now, the derivatives of a scalar field $\partial_\mu f$ are, as we have seen, the components of a one-form field. However, the derivatives of the components of a vector field, $\partial_\mu V^\nu$, are not the components of a tensor field, even though they are labelled by a contravariant and a covariant index. On transforming these derivatives to a new coordinate system, we find

$$\begin{aligned}\partial_{\mu'} V^{\nu'} &= \Lambda^{\mu}{}_{\mu'}\partial_\mu(\Lambda^{\nu'}{}_{\nu} V^\nu) \\ &= \Lambda^{\mu}{}_{\mu'}\Lambda^{\nu'}{}_{\nu}\partial_\mu V^\nu + \Lambda^{\mu}{}_{\mu'}(\partial_\mu \Lambda^{\nu'}{}_{\nu})V^\nu. \qquad (2.20)\end{aligned}$$

Because of the last term, this does not agree with the transformation law for a second-rank tensor. The affine connection will enable us to define what is called a *covariant derivative*, ∇_μ, whose action on a vector field is of the form $\nabla_\mu V^\nu = \partial_\mu V^\nu + (\text{connection term})$. The transformation of the extra term involving the affine connection will serve to cancel the unwanted part in (2.20), so that $\nabla_\mu V^\nu$ will be a tensor.

c) The fact that the functions $\partial_\mu V^\nu$ do not transform as the components of a tensor indicates that they have no coordinate-independent meaning. To see what

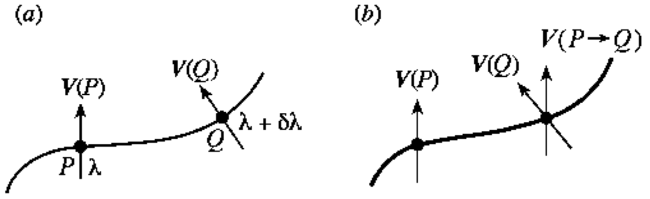


Figure 2.10. $\boldsymbol{V}(P)$ and $\boldsymbol{V}(Q)$ are the vectors at P and Q belonging to the vector field $\boldsymbol{V}$. $\boldsymbol{V}(P \to Q)$ is the vector at Q which results from parallelly transporting $\boldsymbol{V}(P)$ along the curve.

goes wrong, consider the derivative of a component of a vector field along a curve, as illustrated in figure 2.10(*a*), where P and Q are points on the curve with parameters λ and $\lambda + \delta\lambda$ respectively. The derivative at P is

$$\frac{\mathrm{d}V^\mu}{\mathrm{d}\lambda} = \frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}\frac{\partial V^\mu}{\partial x^\nu} = \lim_{\delta\lambda \to 0} \frac{V^\mu(Q) - V^\mu(P)}{\delta\lambda}. \tag{2.21}$$

For a scalar field, which has unique values at P and Q, such a derivative makes good sense. However, the values at P and Q of the components of a vector field depend on the coordinate system to which they are referred. It is easy to make a change of coordinates such that, for example, $V^\mu(Q)$ is changed while $V^\mu(P)$ is not, and so the difference of these two quantities has no coordinate-independent meaning. If we try to find the derivative of the vector field itself, we shall encounter the expression $\boldsymbol{V}(Q) - \boldsymbol{V}(P)$. Now, $\boldsymbol{V}(P)$ is the tangent vector to some curve passing through P (though not necessarily the one shown in figure 2.10(*a*)) and $\boldsymbol{V}(Q)$ is the tangent vector to some curve passing through Q. The difference of two vectors at P is another vector at P: each vector is tangent to some curve passing through P. However, $\boldsymbol{V}(Q) - \boldsymbol{V}(P)$ is not, in general, the tangent vector to a curve at a specific point. It is not, therefore, a vector and has, indeed, no obvious significance at all.

To define a meaningful derivative of a vector field, we need to compare two vectors at the same point, say Q. Therefore, we construct a new vector $\boldsymbol{V}(P \to Q)$, which exists at Q but represents $\boldsymbol{V}(P)$. Then a new vector, $\mathrm{D}\boldsymbol{V}/\mathrm{d}\lambda$, which will be regarded as the derivative of $\boldsymbol{V}$ along the curve, may be defined as

$$\left.\frac{\mathrm{D}\boldsymbol{V}}{\mathrm{d}\lambda}\right|_P = \lim_{\delta\lambda \to 0} \frac{\boldsymbol{V}(Q) - \boldsymbol{V}(P \to Q)}{\delta\lambda}. \tag{2.22}$$

In the limit, of course, Q coincides with P and this is where the new vector exists. There is no natural way in which a vector at Q corresponds to a vector at P, so we must provide a rule to define $\boldsymbol{V}(P \to Q)$ in terms of $\boldsymbol{V}(P)$. This rule is the affine connection. In figure 2.10(*b*), $\boldsymbol{V}(P \to Q)$ is shown as a vector at Q that is parallel to $\boldsymbol{V}(P)$. The figure looks this way because of the Euclidean properties of the paper on which it is printed. Mathematically, the affine

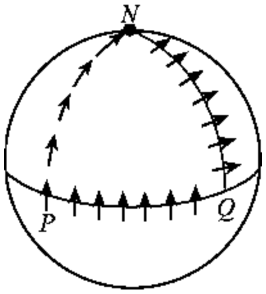


Figure 2.11. Parallel transport of a vector from P to Q on a spherical surface by two routes.

connection *defines* what it means for a vector at Q to be parallel to one at P: it is said to define *parallel transport* of a vector along the curve. From a mathematical point of view, we are free to specify the affine connection in any way we choose. Physically, on the other hand, we shall need to find out what the affine connection is, with which nature has actually provided us, and we shall address this problem in due course. It might be thought that a vector which represents $\boldsymbol{V}(P)$ should not only be parallel to it but also have the same length. In Euclidean geometry, the magnitude of a vector is $(\boldsymbol{v}\cdot\boldsymbol{v})^{1/2}$ and, as we have seen, the scalar product needs a metric for its definition. The metric connection, mentioned above, does indeed define parallel transport in a manner that preserves the magnitude of the transported vector.

The concrete definition of parallel transport is most clearly written down by choosing a coordinate system. If P and Q lie on a curve $x^\mu(\lambda)$ and are separated by an infinitesimal parameter distance $\delta\lambda$, then the components of $\boldsymbol{V}(P \to Q)$ are defined by

$$V^\mu(P \to Q) = V^\mu(P) - \delta\lambda\Gamma^\mu{}_{\nu\sigma}(P)V^\nu(P)\frac{\mathrm{d}x^\sigma}{\mathrm{d}\lambda} \tag{2.23}$$

and the functions $\Gamma^\mu{}_{\nu\sigma}$ are called the *affine connection coefficients*. These coefficients exist at each point of the manifold and are not associated with any particular curve. However, the rule (2.23) for parallel transport involves, in addition to the vector $\boldsymbol{V}$ itself, both the connection coefficients and the tangent vector $\mathrm{d}x^\sigma/\mathrm{d}\lambda$, so parallel transport is defined only along a curve. To transport $\boldsymbol{V}$ along a curve by a finite parameter distance, we have to integrate (2.23). If we wish to transport a vector from an initial point P to a final point Q, we must choose a curve, passing through both P and Q, along which to transport it. There will usually be many such curves and it is vital to realize that the vector which finally arrives at Q depends on the route taken: the functions $\Gamma^\mu{}_{\nu\sigma}$ will generally not take the same values along two different curves. This fact lies at the root of the idea of the *curvature* of a manifold, as we shall see shortly.

The idea of parallel transport is illustrated in figure 2.11, which shows the surface of a Euclidean sphere. For the purposes of this example, we assume the usual metrical properties of Euclidean space, so that distances and angles have their usual meanings. The manifold we consider is the two-dimensional surface of the sphere, so every vector is tangential to this surface. P and Q are points on the equator, separated by a quarter of its circumference, and N is the north pole. The equator and the curves PN and QN are parts of great circles on the sphere and are 'straight lines' as far as geometry on the spherical surface is concerned: one would follow such a path by walking straight ahead on the surface of a perfectly smooth Earth. Consider a vector $\boldsymbol{V}(P)$ that points due north—it is a tangent vector at P to the curve PN. We shall transport this vector to Q, first along the equator and second via the north pole. The rule for parallel transport of a vector along a straight line is particularly simple: the angle between the vector and the line remains constant. For transport along the equator, the vector clearly points north at each step and so $\boldsymbol{V}(P \to Q)$ also points north along QN. Along PN, the vector also points north, so on arrival at the pole it is perpendicular to QN. On its way south, it stays perpendicular to QN. Thus, the transported vector $\boldsymbol{V}(P \to Q)$ as defined by the polar route points along the equator.

At this point, readers should consider parallel transport along the sides of a plane equilateral triangle PNQ. It is easy to see that $\boldsymbol{V}(P \to Q)$ is independent of the route taken. Clearly, the difference between the two cases is that the spherical surface is curved while the plane surface is flat. The rule for parallel transport, embodied mathematically in the affine connection coefficients, evidently provides a measure of the curvature of a manifold, and we shall later formulate this precisely. It should be emphasized that a manifold possesses a curvature *only when* it has an affine connection. If it has no connection, then it is neither curved nor flat: the question just does not arise. Finally, returning to figure 2.11, suppose that we had chosen Q to lie close to P and considered only paths contained in a small neighbourhood of the two points. The surface would have been almost indistinguishable from a flat one and the transported vector would have been almost independent of the path. This is consistent with the mathematical expression (2.23). If P has coordinates x^μ and Q is infinitesimally close to P, with coordinates $x^\mu + \mathrm{d}x^\mu$, then we may substitute $\mathrm{d}x^\mu$ for $\delta\lambda \mathrm{d}x^\mu/\mathrm{d}\lambda$, and all reference to the path between P and Q disappears. The affine connection of two-dimensional Euclidean geometry is explored in exercise 2.10.

One of our motivations for introducing the affine connection was to be able to define a meaningful derivative of a vector field. The covariant derivative along a curve was to be defined, using the idea of parallel transport, by (2.22). As we have just seen, it is not actually necessary to specify a curve when P and Q are infinitesimally close. In terms of components, then, let us write $\mathrm{D}V^\mu/\mathrm{d}\lambda = (\mathrm{d}x^\sigma/\mathrm{d}\lambda)\nabla_\sigma V^\mu$ and calculate the covariant derivative $\nabla_\sigma V^\mu$ using (2.22) and (2.23). We find

$$\nabla_\sigma V^\mu = \partial_\sigma V^\mu + \Gamma^\mu{}_{\nu\sigma} V^\nu. \tag{2.24}$$

Notice that the three indices of the connection coefficient have different functions. There are, indeed, important situations in which the connection is *symmetric* in its two lower indices: $\Gamma^{\mu}{}_{\nu\sigma} = \Gamma^{\mu}{}_{\sigma\nu}$. In general, however, it is the last index that corresponds to that of ∇_σ. Since $\mathrm{D}V^\mu/\mathrm{d}\lambda$ and $\mathrm{d}x^\sigma/\mathrm{d}\lambda$ are both vectors, it follows from their transformation laws that the functions $\nabla_\sigma V^\mu$ are the components of a rank $\binom{1}{1}$ tensor, with the transformation law

$$\nabla_{\sigma'} V^{\mu'} = \Lambda^{\sigma}{}_{\sigma'} \Lambda^{\mu'}{}_{\mu} \nabla_\sigma V^\mu. \tag{2.25}$$

From this, we can deduce the transformation law for the connection coefficients themselves, which can be written as

$$\Gamma^{\mu'}{}_{\nu'\sigma'} = \left(\Lambda^{\mu'}{}_{\mu} \Lambda^{\nu}{}_{\nu'} \Lambda^{\sigma}{}_{\sigma'}\right) \Gamma^{\mu}{}_{\nu\sigma} + \Lambda^{\mu'}{}_{\nu} \left(\partial_{\sigma'} \Lambda^{\nu}{}_{\nu'}\right). \tag{2.26}$$

Readers are urged to verify this in detail, bearing in mind that $\partial_{\sigma'}(\Lambda^{\mu'}{}_{\nu}\Lambda^{\nu}{}_{\nu'}) = (\partial_{\sigma'}\Lambda^{\mu'}{}_{\nu})\Lambda^{\nu}{}_{\nu'} + \Lambda^{\mu'}{}_{\nu}(\partial_{\sigma'}\Lambda^{\nu}{}_{\nu'}) = \partial_{\sigma'}(\delta^{\mu'}_{\nu'}) = 0$.

Evidently, the affine connection is not itself a tensor. However, the covariant derivative that contains it acts on any tensor to produce another tensor of one higher covariant rank. So far, we have defined only the covariant derivative of a vector field, which was given in (2.24). The covariant derivative of a scalar field is just the partial derivative, $\nabla_\mu f = \partial_\mu f$, since this is already a vector field. In order for the covariant derivative of a one-form field to be a second-rank tensor field, we must have

$$\nabla_\sigma \omega_\mu = \partial_\sigma \omega_\mu - \Gamma^{\nu}{}_{\mu\sigma}\omega_\nu. \tag{2.27}$$

Notice that the roles of the upper and first lower indices have been reversed, compared with (2.24), and that the sign of the connection term has changed. It is straightforward to check that these changes are vital if this derivative is to transform as a rank $\binom{0}{2}$ tensor field. The covariant derivative of a tensor field of arbitrary rank is

$$\nabla_\sigma T^{\alpha\beta\ldots}{}_{\mu\nu\ldots} = \partial_\sigma T^{\alpha\beta\ldots}{}_{\mu\nu\ldots} + (\text{connection terms}). \tag{2.28}$$

There is one connection term for each index of the original tensor. For each upper index, it is a term like that in (2.24) and for each lower index it is like that in (2.27). Exercise 2.11 invites readers to consider in more detail how these definitions are arrived at.

There is a convenient notation that represents partial derivatives of tensor fields by a comma and covariant derivatives by a semicolon. That is:

$$\partial_\sigma T^{\alpha}{}_{\mu\nu} \equiv T^{\alpha}{}_{\mu\nu,\sigma} \qquad \text{and} \qquad \nabla_\sigma T^{\alpha}{}_{\mu\nu} \equiv T^{\alpha}{}_{\mu\nu;\sigma}. \tag{2.29}$$

2.3.2 Geodesics

As mentioned earlier, a geodesic is, in a sense, a generalization of the straight line of Euclidean geometry. Of course, we can reproduce only those properties of straight lines that make sense in our manifold with its affine connection. For example, the idea that a straight line is the shortest distance between two points will make sense only when we have a metric to measure distances. The idea of a geodesic is that, if we are to walk along a straight line, each step we take must be parallel to the last. Consider, then, the special case of the parallel transport equation (2.23) in which the vector transported from P to Q is the curve's own tangent vector at P: $V^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$. If the curve is a geodesic, the transported vector $\boldsymbol{V}(P \to Q)$ will be proportional to $\boldsymbol{V}(Q)$. Since the vectors have no definite length, the constant of proportionality may well depend on λ, but if P and Q are separated by an infinitesimal parameter distance, it will be only infinitesimally different from 1. So we may write

$$\left.\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\right|_{P\to Q} = [1 - f(\lambda)\delta\lambda]\left.\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\right|_{Q} \tag{2.30}$$

where $f(\lambda)$ is an unknown function. Using this in (2.23) and taking the limit $\delta\lambda \to 0$, we obtain the *geodesic equation*

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\lambda^2} + \Gamma^\mu_{\nu\sigma}\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\lambda} = f(\lambda)\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}. \tag{2.31}$$

A curve $x^\mu(\lambda)$ is a geodesic if and only if it satisfies an equation of this form, where $f(\lambda)$ can be any function.

Remember now that a given path through the manifold can be parametrized in many different ways, each one being regarded as a different curve. It is easy to see that if the curve given by one parametrization is a geodesic, then so is the curve that results from another parametrization of the same path. We need only express the new parameter, say μ, as a function of λ and use the chain rule in (2.31):

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\mu^2} + \Gamma^\mu_{\nu\sigma}\frac{\mathrm{d}x^\nu}{\mathrm{d}\mu}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\mu} = \left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)^{-2}\left[f(\lambda)\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} - \frac{\mathrm{d}^2\mu}{\mathrm{d}\lambda^2}\right]\frac{\mathrm{d}x^\mu}{\mathrm{d}\mu}. \tag{2.32}$$

This has the same form as (2.31) but involves a different function of μ on the right-hand side. In particular, it is always possible to find a parameter for which the right-hand side of (2.32) vanishes. Such a parameter is called an *affine parameter* for the path. It is left as a simple exercise for the reader to show that if λ is an affine parameter, then any parameter that is affinely related to it (that is, it is a linear function $\mu = a\lambda + b$) is also an affine parameter.

2.3.3 The Riemann curvature tensor

We saw in connection with figure 2.11 that parallel transport of a vector between two points along different curves can be used to detect curvature of the manifold. This is because both parallel transport and curvature are properties of the affine

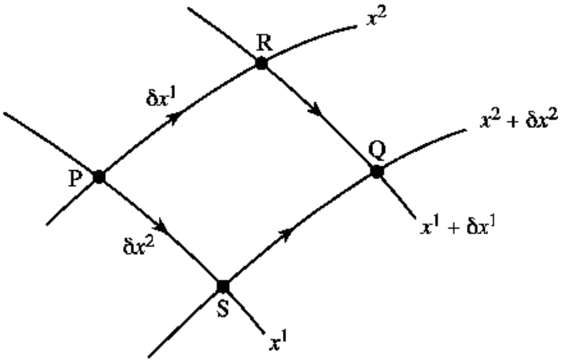


Figure 2.12. Two paths, PRQ and PSQ, for parallelly transporting a vector from P to Q.

connection. The definition of curvature is made precise by the Riemann curvature tensor. Consider two points P and Q with coordinates x^μ and $x^\mu + \delta x^\mu$ respectively, such that $\delta x^\mu = 0$, except for $\mu = 1$ or 2. A region of the (x^1, x^2) surface near these points is shown in figure 2.12. By transporting a vector $\boldsymbol{V}(P)$ to Q via R or S, we obtain at Q the two vectors $\boldsymbol{V}(P \to R \to Q)$ and $\boldsymbol{V}(P \to S \to Q)$. To first order in δx^μ these two vectors are the same, as we have seen. If we expand them to second order, however, they are different, and we obtain an expression of the form

$$V^\mu(P \to S \to Q) - V^\mu(P \to R \to Q) = R^\mu{}_{\nu 12} V^\nu \delta x^1 \delta x^2 + \dots \tag{2.33}$$

where the quantities $R^\mu{}_{\nu 12}$ depend on the connection coefficients and their derivatives. Readers are invited to verify that they are components of the Riemann tensor we are about to define.

It should be clear that the process of transporting the vector from P to Q along the two paths is related to that of taking two derivatives, with respect to x^1 and x^2, in either order. If we act on a vector field with the two covariant derivatives ∇_σ and ∇_τ in succession, the result depends on the order of the two operations; they do not commute. To work out the commutator, we use the definition (2.28), bearing in mind that $\nabla_\sigma V^\mu$ is itself a rank $\binom{1}{1}$ tensor. The result is

$$[\nabla_\sigma, \nabla_\tau] V^\mu \equiv \nabla_\sigma \left(\nabla_\tau V^\mu\right) - \nabla_\tau \left(\nabla_\sigma V^\mu\right) = R^\mu{}_{\nu\sigma\tau} V^\nu + \left(\Gamma^\lambda{}_{\sigma\tau} - \Gamma^\lambda{}_{\tau\sigma}\right) \nabla_\lambda V^\mu \tag{2.34}$$

where

$$R^\mu{}_{\nu\sigma\tau} = \Gamma^\mu{}_{\nu\tau,\sigma} - \Gamma^\mu{}_{\nu\sigma,\tau} + \Gamma^\mu{}_{\lambda\sigma}\Gamma^\lambda{}_{\nu\tau} - \Gamma^\mu{}_{\lambda\tau}\Gamma^\lambda{}_{\nu\sigma}. \tag{2.35}$$

This formidable expression defines the Riemann tensor. As a rank-4 tensor, it has $4^4 = 256$ components! Actually, owing to various symmetry properties, of which the most obvious is antisymmetry in the indices σ and τ, it can be shown that only 80 of these are independent. When $\Gamma^\mu{}_{\nu\sigma}$ is a *metric* connection of the kind to be described in §2.3.5, there is a further symmetry that reduces the number of independent components to 20. Even so, the Riemann tensor is clearly an inconvenient object to deal with. Readers should not panic yet, though. Many of the most important applications of general relativity (including all those to be discussed in this book) do not require the complete Riemann tensor. In practice, we shall need only a simpler tensor derived from it. This is the *Ricci tensor*, defined by contracting two indices of the Riemann tensor:

$$R_{\mu\nu} \equiv R^{\lambda}{}_{\mu\lambda\nu} = \Gamma^{\lambda}{}_{\mu\nu,\lambda} - \Gamma^{\lambda}{}_{\mu\lambda,\nu} + \Gamma^{\lambda}{}_{\sigma\lambda}\Gamma^{\sigma}{}_{\mu\nu} - \Gamma^{\lambda}{}_{\sigma\nu}\Gamma^{\sigma}{}_{\mu\lambda}. \qquad (2.36)$$

Although the definition still looks complicated, the components of this tensor can often be calculated with just a little patience, and it is relatively simple to use thereafter.

The second term on the right-hand side of (2.34) involves the antisymmetric part of the affine connection, $\Gamma^{\nu}{}_{\sigma\tau} - \Gamma^{\nu}{}_{\tau\sigma}$, which is called the *torsion* tensor. (Readers should find it instructive to verify, using (2.26) and (2.13) that this really is a tensor, even though $\Gamma^{\nu}{}_{\sigma\tau}$ itself is not.) In most versions of general relativity, it is assumed that spacetime has no torsion. We shall always assume this too, since it makes things much simpler. I do not know, however, of any direct method of testing this experimentally.

Some simple illustrations of the idea of curvature are given in the exercises. These make more obvious sense when we have a metric at our disposal, and we turn to that topic forthwith.

2.3.4 The metric

Yes, we are finally going to give our manifold a metrical structure that will make the notion of length meaningful. To define the infinitesimal distance $\mathrm{d}s$ between two points with coordinates x^{μ} and $x^{\mu} + \mathrm{d}x^{\mu}$, we use a generalization of the Pythagoras rule:

$$\mathrm{d}s^2 = g_{\mu\nu}(x)\mathrm{d}x^{\mu}\mathrm{d}x^{\nu}. \qquad (2.37)$$

Naturally, we want this distance to be a scalar quantity, independent of our choice of coordinate system, and it is easy to see that the coefficients $g_{\mu\nu}(x)$ must therefore be the components of a rank $\binom{0}{2}$ tensor field. It is called the *metric tensor field* or, for brevity, the 'metric tensor', or simply the 'metric'. Since an antisymmetric part would obviously make no contribution to $\mathrm{d}s$, it is taken to be symmetric in its indices μ and ν. Any finite distance between two points can be uniquely defined only as the length of a specified curve joining them. For the distance between P and Q on a curve $x^{\mu}(\lambda)$, we have the integral

$$s_{PQ} = \int_P^Q \frac{\mathrm{d}s}{\mathrm{d}\lambda}\mathrm{d}\lambda = \int_P^Q \left[g_{\mu\nu}\Big(x(\lambda)\Big)\frac{\mathrm{d}x^{\mu}}{\mathrm{d}\lambda}\frac{\mathrm{d}x^{\nu}}{\mathrm{d}\lambda}\right]^{1/2}\mathrm{d}\lambda. \qquad (2.38)$$

In the space of three-dimensional Euclidean geometry, the squared element of distance expressed in Cartesian coordinates is $\mathrm{d}s^2 = (\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2 + (\mathrm{d}x^3)^2$, so the components of the metric tensor in these coordinates are

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.39}$$

The metric tensor has several other geometrical uses, arising from the fact that it serves to define a scalar product of two vectors or vector fields:

$$\boldsymbol{U} \cdot \boldsymbol{V} = U^{\mu}(x) g_{\mu\nu}(x) V^{\nu}(x). \tag{2.40}$$

Clearly, this reduces to the usual 'dot product' in Euclidean space. Taking the two vectors to be the same, we get a definition of the magnitude or length of a vector,

$$|\boldsymbol{V}(x)|^2 = g_{\mu\nu}(x) V^{\mu}(x) V^{\nu}(x) \tag{2.41}$$

and we can then define the angle between two vectors by writing

$$g_{\mu\nu} U^{\mu} V^{\nu} = |\boldsymbol{U}||\boldsymbol{V}| \cos\theta. \tag{2.42}$$

A non-Euclidean metric does not necessarily give a positive value for the quantity $|\boldsymbol{V}(x)|^2$, so the lengths and angles defined in this way might turn out to be complex.

When introducing one-forms, I pointed out that the symbol $\boldsymbol{u}\cdot$, which appears in the Euclidean dot product, can be regarded as a linear function that takes a vector as its argument, and is, in fact, a one-form. From the scalar product (2.40), we see that $g_{\mu\nu}$ plays the role of the dot, and that the functions

$$U_{\nu} = U^{\mu} g_{\mu\nu} \tag{2.43}$$

are the components of a unique one-form corresponding to the vector $\boldsymbol{U}$. The metric tensor is said to *lower the index* of the vector to produce a one-form. In the same way, the metric associates a unique vector with each one-form ω: it is the vector whose corresponding one-form is ω. Actually, this assumes that the metric is non-singular. That is, it has an inverse matrix $g^{\mu\nu}$, whose elements are the components of a rank $\binom{2}{0}$ tensor field, such that

$$g_{\mu\sigma} g^{\sigma\nu} = \delta^{\nu}_{\mu}. \tag{2.44}$$

The geometrical properties of the metric would be rather peculiar if this were not so, and the existence of the inverse is sometimes included as part of the definition of a metric. So long as the inverse metric does exist, we can say that it *raises the index* of a one-form to produce a vector:

$$\omega^{\mu} = g^{\mu\nu} \omega_{\nu}. \tag{2.45}$$

In fact, any index of any tensor can be raised or lowered in this way. Since $g_{\mu\nu}$ is symmetric, it does not matter which of its indices is contracted.

Now that we have a metric tensor at our disposal, it is clearly possible in practice to regard vectors and one-forms as different versions of the same thing—hence the terms contravariant and covariant vector. In Euclidean geometry, we do not notice the difference, as long as we use Cartesian coordinates, because the metric tensor is just the unit matrix. In non-Cartesian coordinates, the metric tensor is not the unit matrix, and some consequences of this are explored in the exercises. Does this mean that there is, after all, no real distinction between vectors and one-forms, or between the contravariant and covariant versions of other tensors? This depends on our attitude towards the metric. In the relativistic theory of gravity, the metric embodies information about gravitational fields, and different metrics may represent different, but equally possible, physical situations. The relation between the contravariant and covariant versions of a given physical quantity depends on the metric, and it is legitimate to ask which version is intrinsic to the quantity itself and which is a compound of information about the quantity itself and about the metric. To decide this, we must ask what kind of tensor would be used to represent the quantity in question were a metric not available. For example, the Riemann tensor that appears in (2.34) has an index μ, which is in the upper position because it originates from parallel transport of a vector, and two indices σ and τ that must be in the lower position because they label directions along which the vector is being differentiated. Since metrical notions are taken for granted in much of our physical thinking, though, the answer to this may not always be obvious. If, as in Euclidean geometry, the metric is taken to be fixed and unalterable, then such questions need not arise.

2.3.5 The metric connection

Now that the magnitude of a vector and the angle between two vectors have acquired definite meanings, it is natural to demand that the rule for parallel transport should be consistent with them. Thus, if two vectors are transported along a curve, each one remaining parallel to itself, then the angle between them should remain constant. This requirement leads to a relation between the metric and the affine connection that we shall now derive. Consider a curve $x^\mu(\lambda)$ passing through the point P and two vectors $\boldsymbol{V}$ and $\boldsymbol{W}$ at P. We can define a vector field $\boldsymbol{V}(x)$ such that its value at any point Q on the curve is equal to the transported vector $\boldsymbol{V}(P \to Q)$, and a similar vector field $\boldsymbol{W}(x)$. If $\boldsymbol{U}$ is the tangent vector to the curve, then $U^\sigma \nabla_\sigma V^\mu$ is the covariant derivative of V^μ along the curve. It is given by the expression (2.22) and is clearly equal to zero, as is the corresponding derivative of $\boldsymbol{W}$. The consistency condition we want to impose is that the scalar product $g_{\mu\nu} V^\mu W^\nu$ has the same value everywhere along the curve. Recalling that the covariant derivative of a scalar field is equal to the ordinary derivative, we may express this condition as

$$U^\sigma \nabla_\sigma (g_{\mu\nu} V^\mu W^\nu) = 0. \tag{2.46}$$

Now, the covariant derivative of a product of tensors obeys the same Leibniz (or product) rule as an ordinary derivative:

$$\nabla_\sigma(g_{\mu\nu}V^\mu W^\nu) = (\nabla_\sigma g_{\mu\nu})V^\mu W^\nu + g_{\mu\nu}(\nabla_\sigma V^\mu)W^\nu + g_{\mu\nu}V^\mu(\nabla_\sigma W^\nu). \quad (2.47)$$

Readers may verify this explicitly or turn to exercise 2.11 for some further enlightenment. If we use this in (2.46), the last two terms vanish and our condition becomes $U^\sigma(\nabla_\sigma g_{\mu\nu})V^\mu W^\nu = 0$. This must hold for any three vectors $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{W}$, and therefore the covariant derivative of $g_{\mu\nu}$ must be zero:

$$\nabla_\sigma g_{\mu\nu} = g_{\mu\nu,\sigma} - \Gamma^\tau{}_{\mu\sigma}g_{\tau\nu} - \Gamma^\tau{}_{\nu\sigma}g_{\mu\tau} = 0. \quad (2.48)$$

This is sometimes expressed by saying that the metric is 'covariantly constant'. By combining this equation with two others obtained by renaming the indices, we get

$$g_{\sigma\mu,\nu} + g_{\sigma\nu,\mu} - g_{\mu\nu,\sigma} = (\Gamma^\tau{}_{\sigma\nu} - \Gamma^\tau{}_{\nu\sigma})g_{\tau\mu} + (\Gamma^\tau{}_{\sigma\mu} - \Gamma^\tau_{\mu\sigma})g_{\tau\nu} + (\Gamma^\tau_{\mu\nu} + \Gamma^\tau_{\nu\mu})g_{\tau\sigma}. \quad (2.49)$$

Assuming, as we discussed above, that the connection is symmetric in its lower indices, the first two terms on the right-hand side vanish. Then, on multiplying by $g^{\lambda\sigma}$, we find that this symmetric connection is completely determined by the metric:

$$\Gamma^\lambda{}_{\mu\nu} = \tfrac{1}{2}g^{\lambda\sigma}(g_{\sigma\mu,\nu} + g_{\sigma\nu,\mu} - g_{\mu\nu,\sigma}). \quad (2.50)$$

When $\Gamma^\lambda{}_{\mu\nu}$ is used to denote this expression, it is often called a *Christoffel symbol*. This metric connection expresses the definition of parallelism that is implied by the metric. In principle, there is no reason why a manifold should not possess one or more affine connections that would be quite independent of the metric. Indeed, it might also possess several different metrics. In such a case, there would exist several different kinds of 'distance' and several different meanings of 'parallel'. It appears, however, that a single metric and its associated connection given by (2.50) are sufficient to describe the properties of space and time as we know them.

Finally, we can now construct a scalar quantity that gives a measure of curvature (though it obviously contains much less information than the full Riemann tensor). The *Ricci curvature scalar* R is defined by

$$R = g^{\mu\nu}R_{\mu\nu} \quad (2.51)$$

and its interpretation in terms of a 'radius of curvature' is explored in exercise 2.15.

2.4 What is the Structure of Our Spacetime?

We have now invested considerable effort in understanding the mathematical nature of the affine and metrical structures that give precise meaning to our

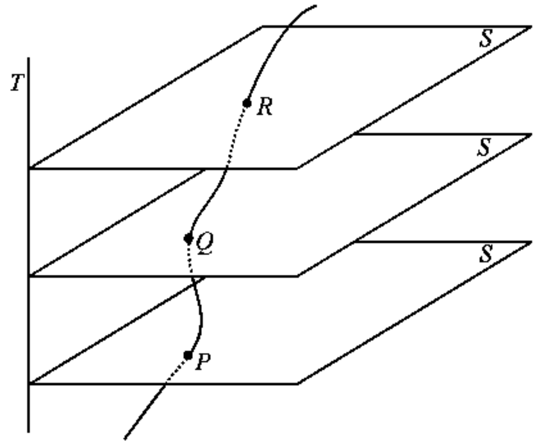


Figure 2.13. Fibre bundle structure of Galilean spacetime and the trajectory of a particle moving through it. Each fibre is a copy of three-dimensional Euclidean space S, which possesses a metric for measuring distances. The base manifold T has its own metric for measuring time intervals. There is no unique way of measuring the 'length' of the particle's trajectory.

intuitive geometrical ideas. The question naturally arises, what are the particular structures that occur in our real, physical space and time? Let us first consider what kind of an answer is needed.

Before Einstein's theories of relativity, it had seemed obvious that the geometry of space was that described by Euclid. (The logical possibility of non-Euclidean geometry had, however, been investigated rather earlier by Gauss, Bolyai, Lobachevski, Riemann and others. The history of this subject is nicely summarized by Weinberg (1972).) The Galilean spacetime that incorporates Euclidean geometry does not have exactly the kind of metrical structure we have been considering. It is a combination (in mathematical jargon, a *direct product*) of two manifolds T (time) and S (space), each of which has its own metric. This structure, illustrated in figure 2.13, is called a *fibre bundle*. It has a base manifold, T, to each point of which is attached a fibre. Each fibre is a copy of the three-dimensional Euclidean space S. A curve such as PQR passing through the spacetime has no well-defined length, although its projection onto one of the fibres does have a definite length l and its projection onto T spans a definite time interval t.

The big difference between Galilean spacetime and the spacetimes of Einstein's theories is that the latter are *metric spaces* (or, more accurately, *manifolds-with-metrics*). That is, the spacetime is a manifold in which a single metric tensor field defines, as we saw in our initial survey, the arc length of any curve. This 'length' is a combination of temporal and spatial intervals, but there is no unique way in which the two can be separated. There is, of course, a profound difference between space and time as we experience them, and we shall discuss in later chapters how this difference fits in with the mathematics.

An important similarity between Galilean spacetime and the Minkowski spacetime of special relativity is that their metrical properties are assumed to be known *a priori*, as specified either by (2.39) or by (2.8). Readers may be puzzled to see that the spatial components in (2.8) have changed sign relative to (2.39). This is purely a matter of convention: the squared proper time intervals in (2.3) or (2.6) are taken to be positive if the separation of two events in time is greater than $1/c$ times their spatial separation, and negative otherwise. (Since proper time intervals are scalar quantities, having the same values in all frames of reference, this distinction is also independent of the frame in which the time and distance measurements are made.) If we chose to think in terms of proper distance rather than proper time, the opposite convention would be more natural, and every component in (2.8) would have the opposite sign. In fact, both conventions are used in the literature, although the one we are using is somewhat more popular amongst high-energy physicists than amongst relativity theorists.

The crux of the general-relativistic theory of gravity is that neither of these simple assumptions about the metric tensor is in fact correct. Indeed, the most important conceptual step we have taken in this chapter is to recognize that the metric tensor is not an intrinsic part of the spacetime manifold, but rather an object that lives in the manifold. It is the same sort of thing as an electric or magnetic field. Electric and magnetic fields vary with position and time in accordance with definite physical laws, which relate them to distributions of charged particles and currents. In the same way, the metric tensor field can be expected to vary in accordance with its own laws of motion and to depend on the distribution of matter. So far, we have no idea what the laws of motion for the metric tensor field are. Electromagnetic fields are easy to produce and control under laboratory conditions, and the laws that govern them were, for the most part, inferred from comprehensive experimental investigations. In contrast, the gravitational forces that are the observable manifestations of the metric tensor field are extremely weak, unless they are produced by bodies of planetary size, and there is little hope of deducing the laws that govern them from a series of controlled experiments. What Einstein did was to guess at what these laws might be, assuming that they would be reasonably similar to other known laws of physics. After one or two false guesses, he arrived at a set of equations, the *field equations* of general relativity, which are consistent with the most precise astronomical observations that it has so far been possible to make.

With the benefit of hindsight, it is possible to see that these equations and all the other laws of classical (non-quantum-mechanical) physics can be deduced in exactly the same way from a single basic principle, called an *action principle*. This seems to me to be most satisfactory. I should be vastly more satisfied if I could explain why an action principle rather than something else is what actually works, but I cannot imagine how that would be done. (It *is* possible to derive the classical action principle from what amounts to a quantum-mechanical version of the same thing, but that is only to rephrase the question!) At this point, then, I propose to interrupt our study of geometry to examine how classical physics works in Galilean and Minkowski spacetimes. This is an important topic in its own right, because classical physics and the simple spacetimes often provide excellent approximations to the real world. In the course of understanding them, however, we shall also meet the action principle, whereupon we shall be equipped to embark upon general relativity and the theory of gravity.

Exercises

2.1. Consider two coordinate systems S and S' whose spatial Cartesian axes lie in the same three directions. The origin of S' moves with constant velocity $\boldsymbol{v}$ relative to S, and the origins of S and S' coincide at $t = t' = 0$. Assume that the relation between the two sets of coordinates is linear and that space is isotropic. The most general form of the transformation law can then be written as

$$\boldsymbol{x}' = \alpha\left[(1 - \lambda v^2)\boldsymbol{x} + (\lambda \boldsymbol{v}\cdot\boldsymbol{x} - \beta t)\boldsymbol{v}\right] \qquad t' = \gamma\left[t - (\delta/c^2)\boldsymbol{v}\cdot\boldsymbol{x}\right]$$

where α, β, γ, δ and λ are functions of v^2. For the case that $\boldsymbol{v}$ is in the positive x direction, write out the transformations for the four coordinates. Write down the trajectory of the S' origin as seen in S and that of the S origin as seen in S' and show that $\beta = 1$ and $\alpha = \gamma$. Write down the trajectories seen in S and S' of a light ray emitted from the origin at $t = t' = 0$ that travels in the positive x direction, assuming that it is observed to travel with speed c in each case. Show that $\delta = 1$. The transformation from S' to S should be the same as the transformation from S to S', except for the replacement of v by $-v$. Use this to find γ. By considering the equation of the spherical wavefront of a light wave emitted from the origin at $t = t' = 0$, complete the derivation of the Lorentz transformation (2.2).

2.2. Two coordinate frames are related by the Lorentz transformation (2.2). A particle moving in the x direction passes their common origin at $t = t' = 0$ with velocity u and acceleration a as measured in S. Show that its velocity and acceleration as measured in S' are

$$u' = \frac{u - v}{1 - uv/c^2} \qquad a' = \frac{(1 - v^2/c^2)^{3/2}}{(1 - uv/c^2)^3}\,a.$$

2.3. A rigid rod of length L is at rest in S', with one end at $x' = 0$ and the other at $x' = L$. Find the trajectories of the two ends of the rod as seen in S and show that the length of the rod as measured in S is L/γ, where $\gamma = (1 - v^2/c^2)^{-1/2}$. This is the *Fitzgerald contraction*. If the rod lies along the y' axis of S', what is its apparent length in S? A clock is at rest at the origin of S'. It ticks at $t' = 0$ and again at $t' = \tau$. Show that the interval between these ticks as measured in S is $\gamma\tau$. This is *time dilation*.

2.4. As seen in S, a signal is emitted from the origin at $t = 0$, travels along the x axis with speed u, and is received at time τ at $x = u\tau$. Show that, if $u > c^2/v$ then, as seen in S', the signal is received before being sent. Show that if such paradoxes are to be avoided, no signal can travel faster than light.

2.5. A wheel has a perfectly rigid circular rim connected by unbreakable joints to perfectly rigid spokes. When measured at rest, its radius is r and its circumference is $2\pi r$. When the wheel is set spinning with angular speed ω, what, according to exercise 2.3, is the apparent circumference of its rim and the apparent length of its spokes? What is the speed of sound in a solid material of density ρ whose Young's modulus is Y? Is the notion of a perfectly rigid material consistent with the conclusion of exercise 2.4?

2.6. Consider the following three curves in the Euclidean plane with Cartesian coordinates x and y: (i) $x = 2\sin\lambda$, $y = 2\cos\lambda$, $0 \leq \lambda < 2\pi$; (ii) $x = 2\cos(s/2)$, $y = 2\sin(s/2)$, $0 \leq s < 4\pi$; (iii) $x = 2\cos(\mathrm{e}^{\mu})$, $y = 2\sin(\mathrm{e}^{\mu})$, $-\infty < \mu \leq \ln(2\pi)$. Show that all three curves correspond to the same path, namely a circle of radius 2. Show that λ and s are affinely related. What is the special significance of s? Find the components of the tangent vectors to each curve. Compare the magnitudes and directions of the three tangent vectors at various points on the circle. What is special about the tangent vectors to curve (ii)?

2.7. Consider a four-dimensional manifold and a specific system of coordinates x^{μ}. You are given four functions, $a(x^{\mu})$, $b(x^{\mu})$, $c(x^{\mu})$ and $d(x^{\mu})$. Can you tell whether these are (i) four scalar fields, (ii) the components of a vector field, (iii) the components of a one-form field or (iv) none of these? If not, what further information would enable you to do so?

2.8. In the Euclidean plane, with Cartesian coordinates x and y, consider the vector field $\boldsymbol{V}$ whose components are $V^x = 2x$ and $V^y = y$, and the one-form field ω_f which is the gradient of the function $f = x^2 + y^2/2$. Show that in any system of Cartesian coordinates $x' = x\cos\alpha + y\sin\alpha$, $y' = y\cos\alpha - x\sin\alpha$, where α is a fixed angle, the components of ω_f are identical to those of $\boldsymbol{V}$. In polar coordinates (r, θ), such that $x = r\cos\theta$ and $y = r\sin\theta$, show that $\boldsymbol{V}$ has components $(r(1+\cos^2\theta), -\sin\theta\cos\theta)$ while ω_f has components $(r(1+\cos^2\theta), -r^2\sin\theta\cos\theta)$. Note that the 'gradient vector' defined in elementary vector calculus to have the components $(\partial f/\partial r, r^{-1}\partial f/\partial\theta)$ does not correspond to either $\boldsymbol{V}$ or ω_f.

2.9. Given a rank $\binom{a}{b}$ tensor, show that the result of contracting any upper index with any lower index is a rank $\binom{a-1}{b-1}$ tensor.

2.10. In the Euclidean plane, parallel transport is defined in the obvious way. If, in

Cartesian coordinates, the components of $\boldsymbol{V}(P)$ are (u, v), then the components of $\boldsymbol{V}(P \to Q)$ are also (u, v). Thus, the affine connection coefficients in Cartesian coordinates are all zero. Work out the matrices $\Lambda^{\mu'}{}_{\mu}$ for transforming between Cartesian and polar coordinates related by $x = r\cos\theta$ and $y = r\sin\theta$. Show that in polar coordinates, the only non-zero connection coefficients are $\Gamma^{r}{}_{\theta\theta} = -r$ and $\Gamma^{\theta}{}_{r\theta} = \Gamma^{\theta}{}_{\theta r} = 1/r$. Let P and Q be the points with Cartesian coordinates $(a, 0)$ and $(a\cos\alpha, a\sin\alpha)$ respectively, and let $\boldsymbol{V}(P)$ have Cartesian components $(1, 0)$. Using polar coordinates and parallel transport around the circle of radius a centred at the origin and parametrized by the polar angle θ, show that $\boldsymbol{V}(P \to Q)$ has polar components $(\cos\alpha, -a^{-1}\sin\alpha)$. By transforming this result, verify that $\boldsymbol{V}(P \to Q)$ has Cartesian components $(1, 0)$. [N.B. The notation here is intended to be friendly: if, say, $x^1 = r$ and $x^2 = \theta$, then $\Gamma^{r}{}_{\theta\theta}$ means $\Gamma^{1}{}_{22}$ and so on.]

2.11. The covariant derivatives of tensors of arbitrary rank can be defined recursively by the following rules: (i) for a scalar field f, we take $\nabla_\sigma f = \partial_\sigma f$; (ii) the covariant derivative of a vector field is given by (2.24); (iii) the covariant derivative of a rank $\binom{a}{b}$ tensor is a tensor of rank $\binom{a}{b+1}$; (iv) for any two tensors A and B, the Leibniz rule $\nabla_\sigma(AB) = (\nabla_\sigma A)B + A(\nabla_\sigma B)$ holds. By considering the fact that $\omega(\boldsymbol{V}) = \omega_\mu V^\mu$ is a scalar field, show that the covariant derivative of a one-form is given by (2.27). Convince yourself that the recursive definition leads to (2.28) for an arbitrary tensor field.

2.12. In the Euclidean plane, consider the straight line $x = a$. Using $\lambda = y$ as a parameter, show, in both Cartesian and polar coordinates, that the geodesic equation (2.31) is satisfied and that λ is an affine parameter. Repeat the exercise using both affine and non-affine parameters of your own invention.

2.13. Write down the components of the metric tensor field of the Euclidean plane in the polar coordinates of exercise 2.8. Show, using both Cartesian and polar coordinates, that the vector $\boldsymbol{V}$ is obtained by raising the indices of ω_f and *vice versa*. Show that $|\boldsymbol{V}|^2 = \omega_f(\boldsymbol{V})$. What is the magnitude of the 'gradient vector'? How does it involve the metric? Can a 'gradient vector' be defined in a manifold with a non-Euclidean metric, or in a manifold that possesses no metric?

2.14. Show that the affine connection of exercise 2.10 is the metric connection.

2.15. In three-dimensional Euclidean space, define polar coordinates in the usual way by $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$ and $z = r\cos\theta$. The spherical surface $r = a$ is called a 2-sphere, and the angles θ and ϕ can be used as coordinates for this two-dimensional curved surface. Show that the line element on the sphere is $\mathrm{d}s^2 = a^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$. Show that the only non-zero coefficients of the metric connection are $\Gamma^{\theta}{}_{\phi\phi} = -\sin\theta\cos\theta$ and $\Gamma^{\phi}{}_{\theta\phi} = \Gamma^{\phi}{}_{\phi\theta} = \cot\theta$. Show that the Ricci tensor is diagonal, with elements $R_{\theta\theta} = 1$ and $R_{\phi\phi} = \sin^2\theta$, and that the Ricci scalar is $R = 2/a^2$.