1

The Principle of Equivalence

1.1 Are Smooth Curves Natural?

When you first encounter calculus, the very first thing you are usually taught to do is to measure the gradient of a curve. Usually, the functional form of that curve is some polynomial, for example, $y = 3x^2$. You are shown how to multiply the coefficient of x^2 by the power and then decrease the power by one unit, so that the slope in this case is $dy/dx = 6x^1$, and probably asked to evaluate the slope at some value of x, for example, the slope is 12 where $x = 2$. This machinery was invented by Newton (Newton et al. 1999), although others, notably Leibniz (Roy 2021), had similar ideas around the same time, and there was closely related work outside the west, for example, at the Kerala school in India (Katz 1995), that predated Newton's renowned contributions.

I think it is fair to say that most introductory calculus courses spend little time worrying about the soundness of the assumptions underlying these mathematical methods. That is because you meet calculus first in mathematics and there are lots of abstract functions for which those assumptions seem to be safe. As physicists, however, our job is to draw parallels between the behaviour of nature, which we measure, and the properties of abstract mathematics. Does nature adhere to the fundamental assumptions underlying calculus?

So, let us look at the central assumption that a curve is differentiable. What does this mean? It means that if I observe a portion of a curve with ever-greater magnification, then the ever-smaller sections of the curve look straighter and straighter. If this is true, then calculus will work, because it gets easier and easier to estimate the slope of a curve as it looks straighter and straighter, and the estimates of the slope based on examining ever-smaller portions of a curve agree with each other better and better as the magnification increases. This means that the concept of a limit is useful – the gradient of a curve is the estimated slope based on extrapolation of the estimates of the slope based on looking at portions of the curve where the

magnification tends to infinity. Mathematically, the gradient of a function $y(x)$ of one variable x is

$$\frac{dy}{dx} = \lim_{\varepsilon \to 0} \frac{y(x+\varepsilon) - y(x)}{\varepsilon}. \tag{1.1}$$

You might think that this is fine for abstract polynomials, although the idea of the ratio of two quantities that are both approaching zero being finite and well behaved is not without its problems, and mathematicians have had to place stronger foundations under the concept of a derivative. Putting these concerns aside, is the concept of a derivative valid in nature? Are nature's curves differentiable?

Some are not. Material objects are ultimately made up of atoms, which do not lend themselves to the formation of mathematical smooth curves. A metal straight-edge from a machine shop certainly looks like a mathematical straight line. However, under a powerful microscope our 'straight-edge' actually looks less and less smooth as the magnification increases. Although this might worry us from a philosophical standpoint, the straight-edge is still very useful for checking the flatness and smoothness of pieces you are fabricating in a machine shop. This works because when you engineer something, you are working to some agreed tolerance, say a micrometer. As long as your straight edge is flat to within one micrometre, it is good enough to check the straightness of other macroscopic objects, also to within similar tolerances. So, our mathematical abstract concept of straightness is bought into approximate correspondence with the realities of a machine shop by agreeing that we will not worry about the breakdown of nature's adherence to the mathematical idea of straightness, as long as the departures between nature and mathematics occur at distance scales that are less than a micrometre. A micrometre is, as a scientist might say, microscopically large but macroscopically small. We can also machine curved objects that look like mathematical curves on a distance scale above a micron, although again when examined more closely, the resemblance disappears.

You might also ponder other curves in nature. When I throw a ball into the air, it follows a curve because of the pull of gravity on the ball. You can imagine this curve in your minds eye, and making the assumption that the acceleration due to gravity close to Earth's surface is of constant magnitude g directed downwards, the form of the curve is

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} u^x t \\ u^y t - \frac{g}{2} t^2 \end{pmatrix}, \tag{1.2}$$

where u^x and u^y are the components of the velocity with which the ball was thrown parallel to and perpendicular to the ground, and t is time. This curve is a parabola from the abstract world of mathematics. How good a model is it for the actual path of the ball? Actually, it is very good but not perfect. For a start, there is air

resistance and turbulence and wind, and of course the acceleration due to gravity is not exactly the same everywhere, then there is the rotation of the Earth, so that the observer standing on the ground is not actually at rest or even moving at a constant velocity, so that the ball is subject to pseudoforces in the accelerating frame of reference of the person who threw it. We could take away many of these approximations in theory by repeating the experiment in an evacuated enclosure, and we could use a mathematical model that accounted for pseudoforces and model the Earth's gravitational field more carefully in the evacuated box. We also have to note that the ball is of finite extent, so that it is the centre of mass of the ball that is meant to follow a smooth mathematical curve. To observe the ball, it has to be under bombardment by the quanta that make up light, which impart some momentum to the ball and introduce random fluctuations to the pathway that the ball actually takes about the smooth abstract curve. Finally, there is the idea that the very nature of the matter making up the ball is due to its coupling to the Higgs boson, so that anything moving at a velocity less than that of light is retarded to its pedestrian pace by constant interactions with Higgs bosons pulled out of the vacuum of standard model particle physics. You can see that the more troubling of these effects involve the underlying quantum nature of reality. Again, at small enough scales in distance or at high enough precision, nature appears to behave in a way that is very hard to model with calculus.

1.2 The Birth of General Relativity

General relativity was invented before quantum mechanics, before we had realised the deep ramifications of the microscopic world and its strange properties. This was actually a great blessing if it meant that Einstein was not overly distracted by worries about whether the smooth geometric world he imagined corresponded to reality. He could proceed to work out the consequences of an idea that he had, which was actually quite simple. The idea can be uncovered by going back to our ball flying through the air. We notice that in Equation (1.2), the mass m of the ball does not appear. This led Einstein to think of the pathway taken by an object freely falling in a gravitational field as being a property of space through which the object is moving. You could take objects of different masses and throw them all into the air with the same initial velocity, and according to Equation (1.2), their centres of mass would all follow the same path. This path is curved, as we can all see when we watch a ball trace it out. Therefore, in some sense, the presence of the large mass of the Earth is causing space and time to become locally curved. This was nicely summarised by John Wheeler:

> Space-time tells matter how to move; matter tells space-time how to curve
>
> (John Archibald Wheeler, 2000)

If we were only interested in ethereal ideas, this fascinating statement might be all we needed. The curving of space-time causes bodies in free fall to follow curved arcs, called geodesics. Equally, the curvature of those geodesics is due to the presence of massive bodies. The two concepts are actually equivalent. Curvature of space-time arises due to matter, matter gives rise to space-time curvature.

However, we are physicists, so we must learn the mathematics that give precise meaning to these statements. What is a curved space, which we know from special relativity is part of a broader concept called space-time? How do you quantify the presence of mass, which we know from special relativity is also the presence of energy? The definition of curvature will be made using the machinery of calculus, which when conducted in higher dimensions than two is called differential geometry. In using calculus to describe the curvature of the space-time of nature, we are assuming that space-time can be considered smooth at all. We know from the earlier discussion that this is a bold assumption.

Ultimately, the smooth space-time that general relativity requires must interface with the microscopic world. Perhaps this indicates that general relativity is an effective theory based on some underlying more fundamental formalism that is unified with the quantum field theory of the standard model. Or, perhaps, gravity is an emergent phenomenon that appears at distance scales that are microscopically large. The smoothness necessary to model it with general relativity could be a sufficiently good approximation at the scales where gravity is observed. At microscopic scales where the smoothness of space-time might cease to be valid, gravity could cease to be a force of nature. After all, it is famously difficult to study gravity at small distance scales; current best efforts probe gravity down to a distance scale of approximately 10 micrometers, a similar scale to the tolerances to which high-quality machine shop straight edges are made. Since we currently do not have a unified theory of gravity and quantum mechanics, we do not know which of these possibilities corresponds to reality. For the remainder of this book, we will ignore these questions and treat space-time as if it is differentiable and free of the complexities of quantum mechanics and quantum fields.

1.3 Tangents to Curves in One Dimension

Suppose we take the curve $y = 3x^2$, which has a gradient of 12 at the point $x = 2$, $y = 12$. The equation of a straight line with gradient $m = 12$ at point $(x_1, y_1) = (2, 12)$ is obtained by substituting it into $y_T - y_1 = m(x - x_1)$ and is $y - 12 = 12(x - 2)$ or $y_T = 12(x - 1)$. Both the curve and the straight line are plotted in Figure 1.1.

We can see that if we define $y' = y - 12x$, then the tangent in y' is horizontal at $x = 2$. This is because substituting into the tangent equation, we get $y' + 12x =$

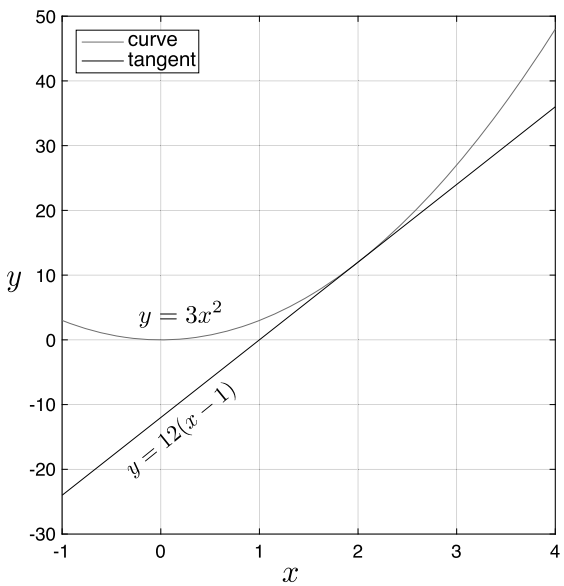

Figure 1.1 The curve $y = 3x^2$ and its tangent at $x = 2$, $y = 12(x - 1)$.

$12x - 1$ or just $y' = -12$. We perform the same linear transformation on the curve, so that $y' + 12x = 3x^2$, and we obtain $y' = 3x^2 - 12x$. Plotting $y'(x)$ for both curve and tangent leads to Figure 1.2.

We find that in the transformed variable y' the tangent has zero gradient, and the curve has a minimum at the point of intersection and rises quadratic in $x - 2$ on either side. In fact, we have $y' = 3(x - 2)^2$. In the coordinates x', y', related to x, y by the linear transformation $y' = y - 12x$, $x' = x - 2$, so that $y' = 3(x')^2$ in the transformed, primed coordinates, the deviation of the curve from a flat straight line in the neighbourhood of $x' = 0$ is quadratic in x'. This deviation vanishes faster than the deviation of x' from $x' = 0$. We could have chosen any other point on any curve, found the gradient of the curve at that point, and deduced a linear transformation that maps the curve onto one with zero gradient at that point and quadratic divergence from flatness moving away from that point, so long as the curve can be differentiated. This is guaranteed by the Taylor expansion. For any differentiable curve $y(x)$, we can write

$$y(x) = y(x_0) + \left.\frac{dy}{dx}\right|_{x_0} (x - x_0) + \frac{1}{2}\left.\frac{d^2y}{dx^2}\right|_{x_0} (x - x_0)^2 + \cdots . \tag{1.3}$$

The tangent curve at the point $x = x_0$ is

$$y_T(x) = y(x_0) + \left.\frac{dy}{dx}\right|_{x_0} (x - x_0). \tag{1.4}$$

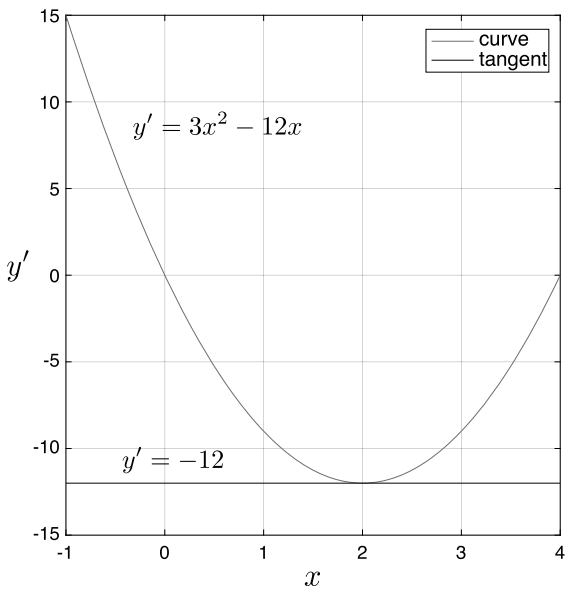

Figure 1.2 The same curve as in Figure 1.1 and its tangent at $x = 2$ following the linear transformation $y' = y - 12x$.

The transformation

$$y' = y - \left.\frac{dy}{dx}\right|_{x_0} x \tag{1.5}$$

results in a curve $y'(x)$ having a local minimum or maximum at $x = x_0$. The difference between the curve $y'(x)$ and the tangent straight line of zero gradient are quadratic in the distance $x - x_0$ moving away from the point x_0. A linear transformation can be made on any differentiable function $y(x)$ to a coordinate y' that has a local minimum or maximum at any point x and is quadratic in departures from that point. This can be shown by substituting the transformation of Equation (1.5) into the expression for the Taylor expansion of the function $y(x)$ about the arbitrary point x_0:

$$y'(x) + \left.\frac{dy}{dx}\right|_{x_0} x = y(x_0) + \left.\frac{dy}{dx}\right|_{x_0} (x - x_0) + \frac{1}{2}\left.\frac{d^2y}{dx^2}\right|_{x_0} (x - x_0)^2 + \cdots$$

$$y'(x) = y(x_0) - \left.\frac{dy}{dx}\right|_{x_0} x_0 + \frac{1}{2}\left.\frac{d^2y}{dx^2}\right|_{x_0} (x - x_0)^2 + \cdots . \tag{1.6}$$

The first two terms on the right are constants, and the third term is quadratic in $x - x_0$, so this curve has a local extremum at $x = x_0$ and deviates from the horizontal as a quadratic on either side of that point.

What have we learned from this? We have learned that for any differentiable curve, we can find a coordinate transformation into a coordinate system where that curve is flat at one point and quadratic in its deviations from flat as you move away from that point.

1.4 Curved Surfaces and Tangent Planes

We can also make tangents to two-dimensional curved planes. This is illustrated in Figure 1.3.

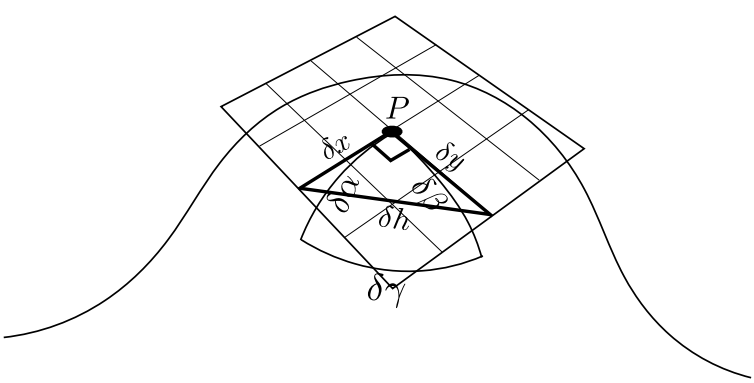

Figure 1.3 A curved surface with a tangent plane at point P. In the tangent plane, a small right-angled triangle has shorter sides δx and δy. In the curved space touching the plane, a curved triangle with edges corresponding to geodesics, or the straightest lines available in a curved space, has shorter sides $\delta\alpha$ and $\delta\beta$.

The flat plane touches the curved surface at point P. This time, instead of dealing with quantities outside the two-dimensional surfaces, let us confine ourselves to quantities that are within the surfaces. So, we are not going to determine the height z of the curved hill above some imagined flat base. Instead, we are going to start to think of ourselves as embedded in the two-dimensional curved surface, like a sort of two-dimensional ant, or alternatively embedded in the flat tangent plane. This is in preparation for considering the three-dimensional space and four-dimensional space-time, where we will not be able to imagine those spaces being embedded in some higher-dimensional space in the way we can imagine a two-dimensional surface embedded in a three-dimensional space.

One thing we can certainly do in the flat plane is to draw a right-angled triangle with its corner at the point P where it touches the curved sheet. We measure the two shorter sides of this triangle, δx and δy, and then the hypotenuse δh. We find that Pythagoras' theorem is obeyed, so that

$$\delta h^2 = \delta x^2 + \delta y^2. \tag{1.7}$$

We now ask the ants in the curved surface to try and reproduce this experiment in their curved space. Of course, they cannot make lines as straight as the ants in the tangent space. However, we can imagine that they can make their lines as straight as possible. For example, suppose they were to stretch an elastic band between two points in the curved surface. It would find the path for which its length was a minimum, and this would be as straight of a line as you can get confined to the curved surface. We will be more precise later in defining geodesics, but for now, it is enough to know that it is possible to determine the straightest possible path, even in a curved space. So, the ants do this for two lines that are at right angles at point P, and then they measure the distance between the two far ends of these lines, again along the geodesic joining the far ends, and they discover that the three lines are of lengths $\delta\alpha$, $\delta\beta$, and $\delta\gamma$. Do you think that these three lengths obey Pythagoras' theorem? It turns out that they do not. For example, suppose the curved surface was the surface of a sphere. Do the sides of right-angled spherical triangles obey Pythagoras' theorem? No they do not, and those of you who have studied spherical trigonometry in astronomy know that there are special rules governing the geometry of spherical triangles.

However, there is in fact a modified version of Pythagoras' theorem that triangles embedded in curved surfaces do obey. Here is a way of writing it:

$$\delta\gamma^2 = g_{11}\delta\alpha^2 + 2g_{12}\,\delta\alpha\,\delta\beta + g_{22}\delta\beta^2, \tag{1.8}$$

where g_{11}, g_{12}, and g_{22} are numerical factors known as metric coefficients, and they reflect the curvature of the surface at point P. Another way of writing this is

$$\delta\gamma^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} g_{ij}\delta\xi^i\delta\xi^j, \tag{1.9}$$

where $\xi^1 = \alpha$, $\xi^2 = \beta$, and $g_{12} = g_{21}$. You are going to have to get used to indices written 'upstairs' in the position where sometimes you will see powers. Whether you are looking at a power or an index should be clear from the context.

Now that we have seen the modification of Pythagoras' theorem for a curved surface, we can write the ordinary Pythagoras' theorem in a flat surface in a way that makes it clear how they are related:

$$\delta h^2 = 1 \times \delta x^2 + 0 \times (\delta x)\,(\delta y) + 1 \times \delta y^2. \tag{1.10}$$

We can see that in the case where our curved surface is in fact flat as well, g_{11} and g_{22} would be 1, and g_{12} would be zero. Evidently, the quantities g_{11}, g_{12}, and g_{22} contain information about the curvature of the surface at point P. Equation (1.10)

can also be written

$$\delta h^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \eta_{ij}\delta x^i \delta x^j, \tag{1.11}$$

where $x^1 = x$, $x^2 = y$, $\eta_{11} = \eta_{22} = 1$, and $\eta_{12} = \eta_{21} = 0$. The coefficients η_{ij} are particular cases of the coefficients g_{ij} that apply specifically to flat spaces described in Cartesian coordinates.

You can see how if the triangles were very small, the coefficients g_{ij} of the modified Pythagoras' theorem would be very close to the coefficients η_{ij} for a flat plane. In fact, exactly at point P, were our triangles to tend to zero size, there would be no difference between the geometry of the curved space and the geometry of the flat plane. This is calculus at work again – curves look straight at high magnification; curved surfaces look flat at high magnification too. However, as you move away from the point P where the two surfaces touch, and your triangles start to get bigger, you start to see discrepancies between the flat surface and the curved one. Those discrepancies show up as changes in the metric coefficients g_{ij}. We can write all this in terms of a Taylor expansion, this time a two-dimensional one, of the metric coefficients about the point (x^1, x^2) where the plane and the tangent surface intersect:

$$\begin{aligned} g_{ij}\left(x^1+\delta x^1, x^2+\delta x^2\right) &= g_{ij}\left(x^1, x^2\right) + \frac{\partial g_{ij}}{\partial x^1}\delta x^1 + \frac{\partial g_{ij}}{\partial x^2}\delta x^2 \\ &\quad + \frac{1}{2}\frac{\partial^2 g_{ij}}{\partial (x^1)^2}\left(\delta x^1\right)^2 + \frac{1}{2}\frac{\partial^2 g_{ij}}{\partial (x^2)^2}\left(\delta x^2\right)^2 \\ &\quad + \frac{\partial^2 g_{ij}}{\partial x^1 \partial x^2}\left(\delta x^1\right)\left(\delta x^2\right) + \cdots \\ &= \eta_{ij} + \frac{\partial g_{ij}}{\partial x^1}\delta x^1 + \frac{\partial g_{ij}}{\partial x^2}\delta x^2 \\ &\quad + \frac{1}{2}\frac{\partial^2 g_{ij}}{\partial (x^1)^2}\left(\delta x^1\right)^2 + \frac{1}{2}\frac{\partial^2 g_{ij}}{\partial (x^2)^2}\left(\delta x^2\right)^2 \\ &\quad + \frac{\partial^2 g_{ij}}{\partial x^1 \partial x^2}\left(\delta x^1\right)\left(\delta x^2\right) + \cdots . \end{aligned} \tag{1.12}$$

Importantly, all the derivatives of the metric coefficients g_{ij} appearing in Equations (1.12) are evaluated at the point of contact (x^1, x^2) between the curved space and the tangent plane. In the second equality, I have substituted $g_{ij}(x^1, x^2) = \eta_{ij}$.

Recall that in one dimension, I can make a coordinate transformation such that any given point on a curve has zero gradient in the transformed coordinates. In the same way, in two dimensions, I can always find a coordinate system where the two

first derivatives of g_{ij} with respect to the two new coordinates are zero,

$$\frac{\partial g_{ij}}{\partial y^1}(y^1, y^2) = \frac{\partial g_{ij}}{\partial y^2}(y^1, y^2) = 0. \tag{1.13}$$

The coordinates for the contact point are (y^1, y^2) in the new coordinate system, and we can therefore write

$$g_{ij}(y^1 + \delta y^1, y^2 + \delta y^2) = \eta_{ij} + \frac{1}{2}\frac{\partial^2 g_{ij}}{\partial (y^1)^2}(\delta y^1)^2 + \frac{1}{2}\frac{\partial^2 g_{ij}}{\partial (y^2)^2}(\delta y^2)^2 + \frac{\partial^2 g_{ij}}{\partial y^1 \partial y^2}(\delta y^1)(\delta y^2) + \cdots . \tag{1.14}$$

This equation can also be written using two summation signs:

$$g_{ij}(y^1 + \delta y^1, y^2 + \delta y^2) = \eta_{ij} + \frac{1}{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{\partial^2 g_{ij}}{\partial y^j \partial y^k}\delta y^j \delta y^l + \cdots . \tag{1.15}$$

Though there are four terms in the double sum, the two where $j \neq k$ are equal. Therefore, in two dimensions, there are three independent second derivatives of g_{ij} in the first non-zero corrections to the flat space metric η_{ij}. This equation is only true in the special coordinates where the first derivatives of the metric with respect to displacements from the contact point are zero. We refer to coordinate systems of this type as locally coincident coordinates, or, since this is a bit of a mouthful, somewhat humorously as 'pigeon' coordinates. The nickname pigeon will be justified in the next couple of sections, where we next consider how these ideas apply to four-dimensional space-time.

1.5 Four-Dimensional Space-Time

All of you have encountered four-dimensional space-time in special relativity, and before we introduce the ideas of Einstein, we need to figure out how to carry over the ideas of curved spaces in two dimensions to possibly curved space-times in four! Unfortunately, it is impossible to visualise four dimensions, but let us just think about what the four-dimensional space-time equivalents of the curved surfaces we have been discussing might be.

In special relativity, we neglected the action of forces on observers, and therefore bodies in special relativity tend to move at constant velocity. We thought of those non-accelerating observers as having their own reference frames, coordinate systems in which the position and time coordinates of events are recorded. There are other observers moving with respect to any given observer who have their own reference frames, and the coordinates of the same event as measured by different observers are related by the Lorentz transforms. For example, if two observers have