relativity's modification of ideas of space and time. Even before relativity, the painter Claude Monet was already fascinated by issues of simultaneity, speed, time and the alteration of space. When physics offered a world of non-Euclidean space-time and a fusing of temporality and spatiality, those notions, or at least metaphorical analogues of them, fell on fertile ground.

Third, after the British astronomer Arthur Eddington's eclipse expedition of 1919 proclaimed that Einstein's theory had correctly predicted the bending of starlight, Einstein became a cult figure standing all at once (at least for his adoring fans) as individual genius, pre-war pacifist, post-war conciliator and moral exemplar. Misunderstood, vilified and then lionized beyond measure, Einstein became a symbol of hope for anyone doing anything against the grain. For his enemies he was, of course, the anti-hero: cosmopolitan, anti-nationalist, Jew, abstract theorist, democrat, cut off from the so-called intuitions of earth, blood and nation. Even before World War II, Einstein, and through him his most famous equation, stood for the mixture of philosophy, physics and modernity that alternately seduced and horrified the world around him.

With the long hot and then cold war stretching from 1939 to 1989, the equation came to stand for something else – nuclear weapons – encapsulating in its sparse symbols both power and knowledge. Here the 'sextant equation' gained a fourth meaning, because these weapons seemed to combine the most esoteric understanding with the most terrible destructiveness. The equation came to signify an almost mystical force, embodying instantaneous and apocalyptic death.

It is in the confluence of these various cultural currents that we find the lines of affect that cluster around this equation. At once philosophy and genial fantasy, practical physics and terrifying weapon, $E = mc^2$ has become metonymic of technical knowledge writ large. Our ambitions for science, our dreams of understanding and our nightmares of destruction find themselves packed into a few scribbles of the pen.

The Rediscovery of Gravity

The Einstein Equation of General Relativity

Roger Penrose

Introduction

Einstein's theory of general relativity provided an extraordinary revolution in our understanding of the physical world. Yet it did not come about through the findings of experimenters' laboratories. It was purely a product of one particular theoretician's insight and imagination. It was thus a revolution that stood in stark contrast with the conventional picture of how a scientific revolution should take place. That picture would hold that a previously accepted scientific viewpoint would be overthrown only when there is a sufficiently impressive accumulation of observational data in contradiction with it. The twentieth century indeed saw some extraordinary revolutions in fundamental physics, each of which led to a thorough overhauling of basic principles and a shattering of previous views as to the nature of physical reality. In the main, they were in accordance with such a conventional picture. But we shall be seeing that general relativity was very different.

In a broad sense, there were two quite distinct fundamental revolutions in twentieth-century physics. The first was relativity, concerned with the nature of space and time, and the second was quantum theory, concerned with the nature of matter. But the theory of relativity itself involved what might be called *two* revolutions, going under the respective names of 'special relativity' and 'general relativity'.

Special relativity is concerned with the strange modifications that must be made to Newtonian physics when bodies travel with speeds that approach that of light, whereby space and time coordinates mysteriously transform among one another, leading to the combined notion of *space-time*. This theory essentially grew out of observational conflict with the idea of an all-pervasive 'ether', which would have defined an absolute state of rest. The most famous conflict with this notion of an ether came from the Michelson–Morley experiment (1887), which attempted to measure the speed of the Earth through the ether; it produced a null result. That experiment, among others, made it increasingly difficult to hold to a Newtonian view of space and time. The revolution that was special relativity came somewhat relentlessly through the work of several scientists: George Fitzgerald, Joseph Larmor, Hendrik Lorentz, Henri Poincaré, Albert Einstein and Hermann Minkowski. I believe that it should, accordingly, be viewed as an example of a revolution of the 'conventional' type, where it was experiments, in the main, that drove the theorists to move away from the Newtonian scheme of things (even though Einstein's own route to the special theory was not particularly experiment-based).

Quantum theory, also, was very experiment-driven. In fact this was true to a far greater degree than in the case of special relativity. Physicists were forced to introduce this new theory to cope with the behaviour of very small-scale matter when they were faced with a vast body of observational data that was in gross conflict with ordinary Newtonian ideas.

The *general* theory of relativity, on the other hand, with its description of gravity as an effect of the 'curvature of space-time' rather than as Newton's gravitational *force*, seemed to have been pulled out of the blue by Einstein, with no apparent need at all for such a revolutionary new viewpoint. At the turn of the twentieth century Newton's beautiful picture of universal gravitation, acting according to an inverse square law of force between particles, was in wonderful accord with observation, to an accuracy of something like one part in ten million. There were still a few minor anomalies, but these all eventually turned out to result from errors of observation or calculation, or from the fact that some disturbing influence had not been taken into account. Well, not quite all – for there was still something not completely accounted for in the tiny details of the motion of the planet Mercury. This was not unduly troubling to astronomers at the time, however, and it was believed that a more careful analysis of the situation would also resolve this apparently minor problem within Newton's scheme of things. Observationally, so it seemed, there was no real expectation that Newton's theory would not suffice.

But Einstein had found himself to be guided to a very different perception of gravitation from that of Newton. It was not observational data that influenced Einstein. Perhaps this is not quite fair. There *was* one piece of observational data that he relied upon, but it was not of the twentieth century nor of the nineteenth, nor even of the eighteenth or the seventeenth. What troubled Einstein had been well established by Galileo in the late sixteenth century (and had been noticed by others even earlier), and was a familiar part of accepted gravitational physics. For more than four centuries the true significance of Galileo's observation had lain dormant. But Einstein saw it with new eyes, and only he perceived its hidden meaning. It led him to the extraordinary view that gravitation is a feature of *curved space-time geometry*, and he produced an equation – now known as Einstein's equation – of unprecedented elegance and geometric simplicity. Yet, to calculate its implications would present enormous technical difficulty, though the results would be almost invariably indistinguishable from those of Newton. Occasionally they would not be, however, and remarkable new effects would come out of Einstein's theory. In one case the precision of Einstein's theory could be seen to advance beyond that of Newton's by another factor of about ten million!

What is this paradigm of a beautiful equation, the *Einstein equation* that governs general relativity? It is commonly written

$$R_{ab} - \tfrac{1}{2} R\, g_{ab} = -8\pi \mathrm{G}\, T_{ab},$$

but what does this mean? Why should this conglomeration of symbols be regarded as beautiful? Clearly, without the meaning that lies behind these symbols, there is neither beauty nor physical significance. We shall come to some real understanding of what this equation means shortly, but for the moment we must settle for a brief interpretation. The quantities on the left-hand side of this equation refer to certain measures of this mysterious 'space-time curvature'; those on the right, to the energy density of matter. Einstein's $E = mc^2$ tells us that energy is essentially equivalent to mass, so the right-hand terms refer equally to the *mass* density. Recall, also, that mass is the source of gravity. Einstein's field equation[1] thus tells us how space-time curvature (left-hand side) is directly related to the distribution of mass in the universe (right-hand side).

Before we begin, a few words about reading mathematical equations may be helpful, as there are indeed some equations in what follows. If you find these things intimidating, then I recommend a procedure that I

normally adopt myself when I come across such an offending line. This is, more or less, to ignore that line and skip over to the next line of actual text. Well, perhaps one should spare the equation a glance, and then press onwards. After a while, armed with new confidence, one may return to that neglected equation and try to pick out its salient features. The text itself should be helpful in telling us what is important about it and what can be safely ignored. If not, then do not be afraid to leave an equation behind altogether.

The Principle of Equivalence

Let's see if we can appreciate what Einstein was striving to achieve in putting forward his general theory of relativity. Why did he feel that there was a physical need to go beyond Newton's highly successful theory? Why did Einstein introduce the notion of space-time curvature? What, indeed, *is* space-time curvature?

The central principle that Einstein believed must be incorporated into gravitational theory in a fundamental way was what he referred to as the *principle of equivalence*. The essential ingredient of this principle was, in effect, known to Galileo at the end of the sixteenth century (and before him by Simon Stevin in 1586, and by others going back to Ioannes Philiponos in the fifth or sixth century). Imagine that a large and a small rock, each of whatever material composition may be chosen, are dropped simultaneously from (say) the top of the Leaning Tower of Pisa. If we may ignore the effects of air resistance, then the rocks will fall at the same rate and reach the ground together. Let us picture a video camera placed on the large rock, aimed at the small one. Since the two rocks fall exactly together, the image that the video camera records is of a small rock just hovering, seeming to be stationary and therefore apparently unaffected by gravity. To the rocks (until they hit the ground), the Earth's gravity seems to have completely vanished!

This observation contains the essence of the principle of equivalence. By falling freely in under gravity, one can eliminate its local effects, so that apparently the gravitational force has disappeared. Conversely, it is possible to produce effects indistinguishable from those of gravity by referring things to an accelerating reference frame. This apparent gravity due to acceleration is a familiar feature of modern high-speed transport. As a car accelerates forwards, the occupants are pressed to the backs of their seats as though a new gravitational force had suddenly appeared, pulling the occupants to the rear. Similarly, if the driver suddenly applies the brakes, then the occupants seem to be pulled forwards, as though there is a gravitational force pulling them to the front of the car. If the car swings to the right, then there would appear to be a gravitational force pulling the occupants to the left, and so on. These effects are particularly manifest on an aeroplane, as it is often difficult to tell which direction is actually 'down' – i.e. towards the Earth's centre – owing to the confusion of effects from the plane's acceleration and the Earth's actual gravitation. The principle of equivalence tells us that this confusion is a fundamental property of gravity. The physical laws that appear to be operating if measurements are taken with respect to an accelerating reference frame are just the same as those that operate if the reference frame is considered to be *un*accelerating but where an appropriate gravitational field of force is introduced, in addition to those forces already present.

It should be remarked that this 'equivalence' property is something that holds only for the *gravitational* field, and not for any other type of force. It certainly does *not* hold if we take an electric field in place of a gravitational one. Consider, for example, a corresponding situation to that outlined above, where rocks are imagined to be dropped from the Leaning Tower, but now with electric forces replacing the gravitational ones. The acceleration rate at which a body 'falls' in a background electric field is by no means independent of its compositional nature. This acceleration depends upon what is referred to as the body's charge-to-mass ratio. To take an extreme case, we could imagine that the two bodies have equal mass but their charge values are opposite (so that one is positively charged and the other negatively). Then the bodies would accelerate in the background electric field in opposite directions! A video camera placed on one body would certainly not register the other as being unaccelerated.

The issue with regard to the charged bodies in a background electric field, as opposed to the massive bodies in a background gravitational field, is that the force on the charged body is proportional to its *charge*, whereas its resistance to motion – i.e. its *inertia* – is proportional to its *mass*. What is special about the gravitational case is that the force on the body and its resistance to motion are *both* proportional to its mass. From the perspective of Newtonian theory, this fact seems entirely fortuitous. The equality between *gravitational mass* (controlling the strength of the gravitational force on a body) and *inertial mass* (controlling resistance to change of motion) is by no means an essential requirement for a dynamical theory of the Newtonian type, but this equality in the case of gravity makes things a little simpler, since one does not have two kinds of mass to worry about.

Although these matters were known for a long time – basically since Galileo's early considerations and certainly appreciated by Newton – it was Einstein who first realized the profound *physical* importance of the principle of equivalence. What importance was this? Let us first recall Einstein's development of *special* relativity. He had then taken the 'principle of special relativity' to be a fundamental principle. According to this principle, the laws of physics are the same with respect to any uniformly moving (unaccelerated) observer. Although Larmor, Lorentz and Poincaré before him had had the basic transformation laws of special relativity, none of them had adopted Einstein's viewpoint that this *relativity principle* should be fundamental and therefore respected by all the forces of nature. Einstein's fundamentally 'relativistic' attitude on this matter had led him to ponder upon whether there is really anything particular about the restriction to *uniform* motion in the statement of the relativity principle. What about the way in which physical laws are perceived by an *accelerating* observer?

At first sight, it would appear that accelerating observers simply perceive laws that are *different* from those perceived by uniformly moving observers. In Newtonian language, one needs to introduce 'fictional forces' (i.e. 'unreal' forces) to cope with the effects of acceleration. Here is where the principle of equivalence comes in. According to Einstein, such fictional forces are no less real (and no more real) than the gravitational force that we all seem to feel pulling us downwards to the centre of the Earth. For the force of the Earth's pull can appear to be eliminated if we fall freely with it. Recall our imagined video camera attached to one of Galileo's falling rocks. In the accelerating frame of the video camera, the Earth's field seems to have disappeared. It seems to have been rendered 'fictional' by the simple procedure of referring things to a reference frame at rest with respect to the video camera.

With Einstein's viewpoint, an accelerating observer perceives the same laws as those of the unaccelerated one provided that an appropriate new *gravitational field of force*, arising from the acceleration, is introduced in addition to all the other forces involved. In the case of the falling video camera, this additional field would be a gravitational field directed upwards which just cancels the Earth's downward field. In the video camera's reference frame, therefore, the gravitational field has been reduced to zero.

In a speech Einstein gave in Japan in 1922, he recalled the moment at which he happened on this idea, which occurred to him late in 1907:

> I was sitting in my chair in the patent office when all of a sudden a thought occurred to me: 'if a person falls freely he will not feel his own weight'. I was startled. This simple thought made a deep impression on me. It impelled me toward a theory of gravitation.

Elsewhere, Einstein referred to this realization as 'the happiest thought of my life'. For it contained the seeds of his wonderful general theory of relativity.

Yet the reader may be forgiven for worrying that Einstein seems to have eliminated gravity altogether with this point of view. Surely there *is* an effect that we call gravity! The planets surely *do* move in ways that are beautifully accounted for by Newtonian theory. And there surely *does* seem to be something that holds us to our chairs! The Einsteinian view would appear to be telling us that there is no such thing as gravity, since we can always eliminate the gravitational force by simply choosing a frame of reference that is in free fall. Where has gravity gone in this Einsteinian view? In fact it has not gone away, but is concealed in some subtleties that I have glossed over. In the next section, we shall see where the gravitational field is indeed hiding.

Tidal Forces

The considerations of the previous section are essentially local. I have ignored how Newton's gravitational field of force might be varying from place to place. The direction 'down' is not quite the same here in Oxford as it is in London, owing to our differing locations on the globe. If I try to eliminate the gravitational field where I sit here at my desk, by considering my descriptions relative to a rigid reference frame that falls freely to the ground here in Oxford, then this frame will not quite do the job for someone in London. Thus, the 'elimination' of the gravitational field by adopting a freely falling frame is not really a straightforward matter.

To make the situation a little more specific, let us imagine an astronaut called Albert – but we shall refer to him simply as '*A*' for short – who falls freely in the vicinity of the Earth. We could imagine that *A* simply drops towards the ground, but this might be considered to a little inhumane. We are concerned just with accelerations and not velocities directly, so it is just as good to suppose that Albert is safely in free orbit about the Earth. Let us

suppose that A is surrounded by a small sphere of particles, initially at rest with respect to A. Each particle will have an acceleration towards the Earth's centre C, and this will be in accordance with Newton's inverse square law. The two particles P_1 and P_2 that lie on the straight line CA will have accelerations in the direction of C, but the acceleration of the lower point P_1 will be a little greater than that at A, and the acceleration at the higher point P_2 a little less than that at A. Thus, *relative* to Albert, P_1 will be accelerating slowly down towards the Earth's centre C but P_2 will be accelerating up away from C. Both P_1 and P_2 will appear, to A, to be accelerating away from A. On the other hand, any particle P_3 on the horizontal circle of particles centred at A will accelerate slightly inwards, as it is pulled towards the Earth's centre C, since C is a definite finite distance from A, with a slightly different 'down' direction. Relative to A, the acceleration of such a point P_3 will appear to be inwards towards A. The entire sphere of particles will become distorted into a prolate (cigar like) ellipsoidal shape, moving inwards towards A in horizontal directions relative to A, and moving outwards along the line from A to the centre C. (See Figure 1.)

This distortion effect is referred to as the *tidal effect* of gravity. The reason for the description 'tidal' is that it is precisely this same effect that is responsible for the tides of the Earth's oceans, as governed by the location of the Moon. To see this, let us now imagine that A represents the centre of the Earth and that the sphere of particles represents the surface of the Earth's oceans. Let C now represent the location of the Moon. Again there will be slightly differing accelerations towards the Moon's centre C at all the points on the ocean's surface. The resulting effect, relative to the Earth's centre A, will be to cause a prolate ellipsoidal distortion of the ocean surface, which bulges in a direction towards the Moon (C) and also in the opposite direction. This is precisely the main effect that gives rise to the tides. (Subsidiary influences are the Sun's similar, but smaller, tidal effect and the frictional and inertial influences on the actual motion of the water in the oceans.)

It is a particular (defining) feature of Newton's *inverse square law* that the *volume* of the sphere of particles remains initially constant in its momentary distortion into an ellipsoid. (What this amounts to saying is that the outward acceleration at P_1 and P_2 is twice the inward acceleration at the horizontal points like P_3.) This fact depends upon there being no mass density within the sphere itself. If there were a significant amount of massive material *within* the sphere, then there would be an additional inward

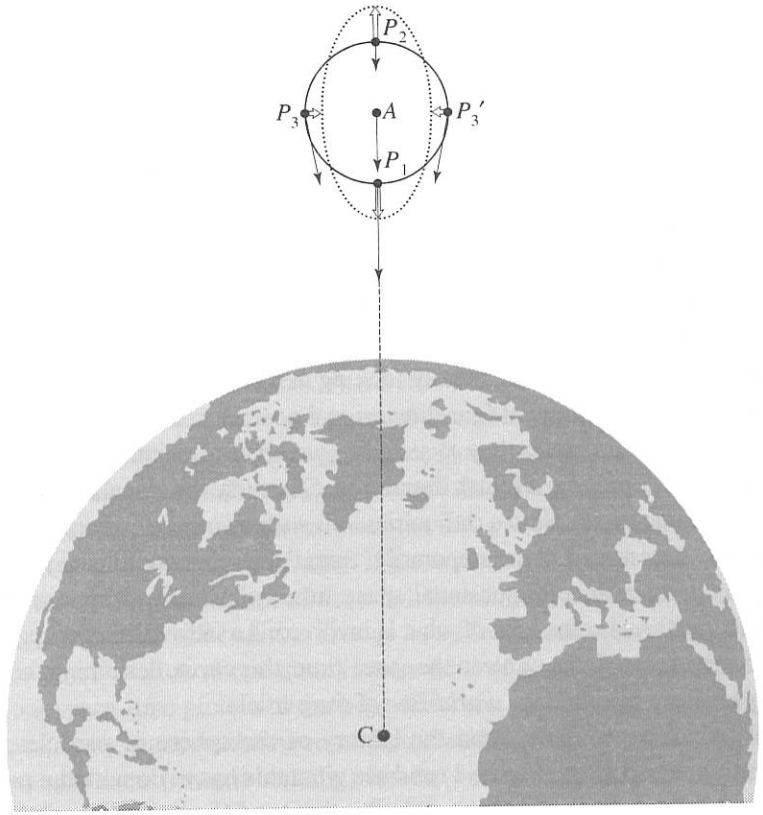


Figure 1 The tidal effect. Open arrows show relative acceleration.

acceleration that would serve to *reduce* the volume of the sphere in its initial motion. The amount of this (initial) volume reduction is, quite generally, proportional to the total mass surrounded by the sphere. In fact, Newton's magnificent gravitational theory is effectively encompassed within the simple facts that I have just described.

A particular example of this volume reduction would occur if we consider our sphere of particles to surround the Earth completely, in the vicinity of the Earth's surface, where we are now concerned with the Earth's gravitational field *itself*, rather than the small corrections due to the Moon which are (mainly) responsible for the tides. The distortion of our sphere is now of the pure volume-reducing type. This is an inward acceleration all around the Earth, and it supplies us with the familiar gravitational field that indeed holds us to our chairs.

Space-Time Curvature

Although the idea of space-time has not yet featured in these considerations, and we shall be coming to this more fully in the following section, it is useful to get some feeling for why the above way of looking at Newtonian gravity is actually telling us that Einstein's perspective on gravitational theory, where the principle of equivalence is regarded as fundamental, leads naturally to the notion that gravitation is manifested in a form of *space-time curvature*. Let us try to imagine that the history of the universe is laid out before us as a *four*-dimensional continuum. We are not, for the moment, trying to depart from Newtonian physics; we are merely looking at the Newtonian universe in an unusual way – as a piece of four-dimensional geometry! In addition to having three spatial coordinates, say x, y and z, we shall also introduce the time coordinate t describing a fourth dimension. Of course the visualization of the full four dimensions creates difficulties, but such a complete visualization is not really necessary. Let us temporarily 'forget' the space coordinate y, so that we now have a three-dimensional space-time coordinatized by x, z and t. Figure 2 gives us some idea of what is involved. An individual point particle is now represented as a *curve* in the space-time; this curve, describing the particle's history, is called the *world-line* of the particle.

We shall try to understand the history of the sphere of particles surrounding Albert (from Figure 1) and see what this has to do with the notion of space-time curvature. At the right-hand side of Figure 2, I have tried to

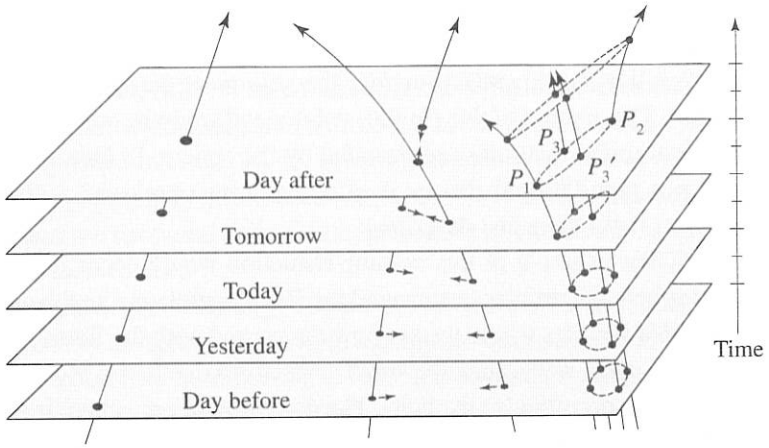
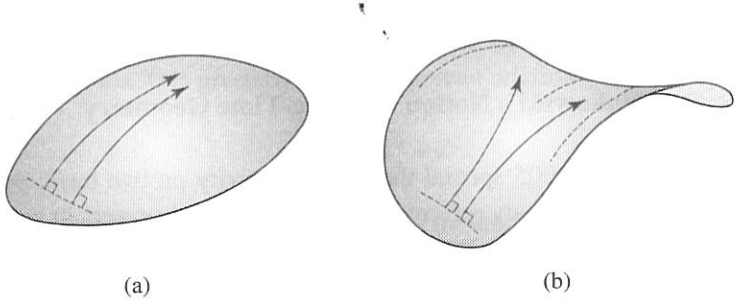


Figure 2 Space-time (Newtonian case). Geodesic deviation (tidal effect) is illustrated on the right.

depict the history of the evolution of this sphere, with one of the spatial dimensions (namely the horizontal dimension coordinatized by y) suppressed. The sphere (in this reduced dimensionality) now appears as a circle, and as time evolves it gets distorted into an ellipse. Note that there is a bending outwards of the world-lines of the vertically displaced particles P_1 and P_2 (major axis of the ellipse), this bending being outwards away from Albert's central world-line. On the other hand there is an inward bending for the world-lines of the horizontally displaced particles P_3 and P_3' (minor axis of the ellipse).

We are to compare this 'bending effect' with the behaviour of geodesics on a curved surface. A geodesic is a curve of minimal length on such a curved surface. We may think of a piece of string stretched taut over the surface. It will describe a geodesic on that surface. If the surface has what is called *positive curvature* (like the curvature of an ordinary spherical surface), then slightly displaced geodesics that start out parallel to each other will begin to curve in towards each other. If the surface has what is called *negative curvature* (like the surface of a saddle), then slightly displaced geodesics starting out parallel will begin to diverge away from each other. (See Figure 3.) This manifestation of curvature is referred to as *geodesic deviation*.

In our space-time picture of the tidal distortion, as illustrated at the right in Figure 2, we see a combination of these two kinds of curvature. There is positive curvature (inward-bending) for the horizontally displaced world-lines of P_3 and P_3', whereas for the vertically displaced world-lines of P_1 and P_2 we have negative curvature (outward-bending). This interpretation of the distortion of world-lines that occurs in the *tidal effect* as a geodesic deviation of some kind becomes justified when we are able to think of the world-lines of particles, freely moving under gravity, as geodesics in space-time. For this



(a) (b)

Figure 3 (a) Positive curvature causes convergence of geodesics – like the surface of an orange
(b) Negative curvature causes divergence of geodesics – like a saddle.

we shall need to have an appropriate notion of 'distance' in space-time. We shall be coming to that in the next two sections. We shall be seeing that the tidal effect is indeed an instance of geodesic deviation, and it is thus a direct measure of space-time curvature.

We observe that the notion of curvature in higher dimensions is a more complicated thing than it is in the two-dimensional case. In two dimensions, we find that the curvature at any point is given simply by a *single number*,[2] which would be a *positive* number in the sphere-like case of positive curvature and a negative number for the saddle-like case of negative curvature. In more than two dimensions the curvature is described by *several* numbers, called *components* of the curvature, these basically measuring the two-dimensional type of curvature in various different directions. In the example just considered we have seen, in effect, a positive curvature component referring to the horizontal direction from A to P_3 and P_3', and a negative curvature component referring to the vertical direction from A to P_1 and P_2. In fact, in the four dimensions of space-time, there are *twenty* independent components to the curvature, and these can be collected together to describe a mathematical entity referred to as the 'Riemann curvature tensor'. I shall defer discussion of the notion of a tensor until a later section, but it is worth pointing out here that the Einstein equation is itself a tensor equation, and the little indices (such as the a and b on R_{ab}) simply provide a labelling for such components in different directions.

So far, we have not really been doing general relativity, but merely Newtonian gravitational theory from the Einsteinian perspective.[3] In order to move forward to full general relativity, we shall have to understand a little more about *special* relativity: why is it really a four-dimensional space-time theory, and what is the appropriate notion of 'distance' in this space-time geometry? Let us come to all this next.

Minkowski's Notion of Space-Time Geometry

Einstein based his 1905 special theory of relativity on two basic principles. The first was already referred to earlier; for all observers in uniform motion the laws of nature are the same. The second was that the speed of light has a fundamental fixed value, not dependent upon the speed of the source. A few years earlier, the great French mathematician Henri Poincaré had a similar scheme (and others, such as the Dutch physicist Hendrik Lorentz, had moved some way towards this picture). But Einstein had the clearer vision that the underlying principles of relativity must apply to *all* forces of nature.

Historians still argue about whether or not Poincaré fully appreciated special relativity before Einstein entered the scene. My own point of view would be that whereas this may be true, special relativity was not *fully* appreciated (*either* by Poincaré *or* by Einstein) until Hermann Minkowski presented, in 1908, the four-dimensional space-time picture. He gave a now famous lecture at the University of Göttingen in which he proclaimed, 'Henceforth space by itself, and time by itself are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.'

Einstein seems not to have appreciated the significance of Minkowski's contribution initially, and for about two years he did not take it seriously. But subsequently he came to realize the full power of Minkowski's point of view. It formed the essential background for Einstein's extraordinary later development of *general* relativity, in which Minkowski's four-dimensional space-time geometry becomes *curved*.

The physical interpretation of this curvature is basically that which has already been given, but there is still an essential missing ingredient, namely the interpretation of the world-lines of particles moving freely under gravity as *geodesics* in space-time geometry. Examples of such geodesics would be the world-lines of our astronaut A and of the surrounding sphere of particles. To understand this interpretation, it will be important first to appreciate the *flat* four-dimensional mathematical structure that Minkowski actually introduced in order to describe special relativity.

For this it is helpful to start by considering familiar three-dimensional Euclidean geometry. It is convenient to introduce Cartesian coordinates x, y, z to label points in Euclidean 3-space. Then the *distance* l from the origin (with coordinates $x = y = z = 0$) to the point (X, Y, Z) (i.e. coordinates $x = X$, $y = Y$, $z = Z$) is given by the Pythagorean relation

$$l^2 = X^2 + Y^2 + Z^2.$$

(The reader will recall the Pythagorean theorem, which states that the squared length of the hypotenuse of a right-angled triangle is equal to the sum of the squared lengths of the other two sides. This would be the two-dimensional formula $l^2 = X^2 + Y^2$, since the distance between two points in the plane is the hypotenuse l of a triangle whose other two sides' lengths are X and Y. The extension to three dimensions is a two-step consequence of

this.) We can also use the above formula to express the Euclidean distance between *any* two points, where now X represents the *difference* between the x-coordinate values of the two points, and similarly for Y and Z.

It is easy to generalize the formula to four dimensions and obtain the squared distance from the origin to the point $w = W$, $x = X$, $y = Y$, $z = Z$ in Euclidean *4-space* as

$$l^2 = W^2 + X^2 + Y^2 + Z^2.$$

However, Minkowski's space-time geometry differs subtly but importantly from this. Although space and time coordinates indeed get mixed up with one another in relativity theory according to a kind of rotation (a 'Lorentz transformation'), the way that an ordinary Euclidean rotation mixes up the (w, x, y, z)-coordinates does not give us quite the correct prescription. There is a qualitative distinction between the space and time coordinates in Minkowski's description, which shows up as a *sign* difference in the above distance formula.

In place of the fourth spatial coordinate w we introduce a time coordinate t. How do we modify the above formula so as to obtain the correct Minkowskian measure of 'distance' τ? In fact, in order to arrive at the most directly physical such measure, it is appropriate to reverse the signs of *all* the spatial contributions, leaving the one temporal coordinate $t = T$ to contribute with a positive sign:

$$\tau^2 = T^2 - X^2 - Y^2 - Z^2.$$

Here I am using units of distance and time so that the speed of light comes out as *unity*. Thus, if we were to use the year as the time unit, then we should have to use the light year as the unit of spatial measure; if we use the second as the time unit, then we must use the light second as the unit of spatial measure (about 186,000 miles).

What kind of 'distance' is then defined by the quantity τ? It is better to think of τ as a measure of *time*. It is what is called the *proper time*. If the space-time point P with coordinates $t = T$, $x = X$, $y = Y$, $z = Z$ is such that the quantity on the right-hand side of the above expression is *positive*, then P is *timelike separated* from the origin O, which means, physically, that it is theoretically possible for the world-line of a particle to pass from O to P (if T is positive) or from P to O (if T is negative). If this particle moves uniformly in a straight line from O to P, then the quantity τ (taken with the positive sign) is the *time* (proper time) actually experienced by the particle between O and P as measured by an ideal clock situated on the particle. (The fact that this time is not simply the Newtonian t, but involves the spatial coordinate differences also, is an expression of the 'relativity of time' that occurs with special relativity.) As with Euclidean geometry above, these considerations apply also when the origin O is replaced by some arbitrary point P', but where now the quantities T, X, Y, Z refer to the *differences* between the respective t, x, y, z coordinates of the two space-time points P and P', and where t is the time experienced by the particle moving inertially from P to P'.

Minkowskian geometry has the curious property that the 'distance' between two points P and P' can sometimes be zero even though P and P' do not coincide. This happens when a light ray can contain both P and P' (which we think of as a 'particle of light', or photon, travelling with the speed of light). Thus, noting the above interpretation of 'Minkowski distance' as proper time, we find that a photon would not experience any passage of time at all (if photons could actually experience anything!). For fixed P, the locus of such points P' constitutes the (future) *light cone* of P. The light cones are important because they determine the *causality properties* of Minkowski space, but I shall not be much concerned with them here. The one essential point that will be needed is that the world-line of a particle with mass must lie *within* the light cone at each of its points. This simply expresses the fact that the particle does not exceed the speed of light anywhere. Such a world-line is referred to as a *timelike* curve. Any massive particle's world-line must be a timelike curve.

Now any timelike curve (i.e. allowable particle world-line) has a Minkowski 'length' whether or not the curve is straight. A curved world-line describes an *accelerating* particle. This 'length' is, physically, simply the (proper) time that is experienced by the particle. To obtain this length mathematically, we just do the same thing that we would do in ordinary Euclidean geometry, except that we must take into account the sign differences, noted above, that are involved in passing from Euclidean to Minkowskian geometry. To do this explicitly, we need the *infinitesimal* expression for the length that measures the 'distance' between two infinitesimally separated points. We then 'add up' (technically: *integrate*) all these infinitesimal separations along the curve to get the total length. In Euclidean three-dimensional geometry, this infinitesimal separation 'dl' is related to standard Cartesian coordinates x, y, z by the formula

$$dl^2 = dx^2 + dy^2 + dz^2.$$

In the Minkowski case, we must modify this to

$$d\tau^2 = dt^2 - dx^2 - dy^2 - dz^2,$$

but the interpretation is completely analogous. (Those unfamiliar with the relevant calculus notations can imagine dt to stand for $t'-t$ and dx to stand for $x'-x$, etc. where P' lies infinitesimally close to P within the light cone of P.) The total lapse of time (proper time) between two points on a world-line, as measured by an ideal clock, is the total 'length' of the world-line between these points.

An important feature of length in Euclidean geometry is that among all the curves joining two points, the length is *minimum* when the curve is straight. ('The shortest distance between two points is a straight line.') There is a closely analogous property in Minkowskian geometry, except that things are the other way around. If we select a pair of timelike-separated points, then among all timelike curves joining them, the proper time is a *maximum* when the curve is straight. Physically, this provides us with what is sometimes referred to as the 'clock paradox' (or 'twin paradox'), whereby a traveller to a distant star and back ages less (because of a shorter 'Minkowski distance') than his twin sister whom he leaves behind on Earth. The Earthbound twin has a straight world-line, and therefore she experiences a greater duration of time than does her space-travelling brother, whose world-line is curved because of the acceleration. It is very misleading, however, to think of this as a paradox. Admittedly it takes some getting used to, but it is not actually paradoxical, and many experiments have now confirmed this effect to great accuracy. Minkowski geometry makes the time difference between the two twins seem almost 'ordinary'.

Why was Einstein led to modify Minkowski's beautiful space-time geometry and introduce curved spacetime? We have seen that in special relativity, particles moving freely in the absence of forces – i.e. *inertially* moving particles – have straight world-lines in Minkowski space. Einstein's desire to incorporate the principle of equivalence into physical theory led him to the view that a *new* concept of 'inertial motion' was required. Since the gravitational force can be locally eliminated by use of a freely falling reference frame, we are not to consider the gravitational force as 'real' according to Einstein's viewpoint. So Einstein found that he needed to introduce a different notion of inertial motion, namely *free fall under gravity*, with no other forces acting. Because of the tidal effects that we encountered above, we cannot think of the 'inertial' particles (in Einstein's sense) as having straight (i.e. geodesic) world-lines in Minkowski's geometry. For this reason we need to generalize this geometry so that it becomes *curved*. Einstein found that, indeed, the world-lines of his inertial particles could now be *geodesics* in this curved geometry – locally maximizing the 'length' rather than minimizing it, in accordance with the above – and the tidal distortion is indeed an instance of geodesic deviation, providing a direct measure of space-time curvature. Let us try to understand this curvature a little more fully.

Curved Space-Time Geometry

In the nineteenth century two great German mathematicians, Carl Friedrich Gauss and Bernhard Riemann, introduced the general notion of 'curved geometry'. To get a feeling for this kind of geometry, think of the surface of a tennis ball divided in half. It can be flexed in various ways, but what is called its *intrinsic* geometry remains unchanged under such deformations. Intrinsic geometry is concerned with distances measured *along* the surface. It is not concerned with the space (here our ordinary Euclidean three-space) in which the surface may be pictured as embedded. Distances measured directly across from one point to another taken *outside* the surface are not the concern of intrinsic geometry. The length of a curve drawn *on* the half tennis ball is unchanged by the flexing, however, and such lengths are the basic concern of intrinsic geometry.

Gauss introduced this idea of intrinsic geometry in 1827 in the two-dimensional case, like our tennis-ball surface just considered. He showed that there is a notion of curvature in this geometry that is entirely intrinsic, so that it is completely unaffected by changes in the way that such a surface might be embedded. This curvature can be calculated from the length measures along the surface, where we think of the lengths of curves on the surface as obtained by integrating an *infinitesimal* measure of length dl along the curve, just as above. In practice, one introduces some convenient system of coordinates on the surface, say u, v, and we find an expression for dl in the form

$$dl^2 = A\,du^2 + 2B\,du\,dv + C\,du^2$$

where A, B and C are functions of u and v (this expression being locally the same as the infinitesimal 'Pythagorean' expression for distance $dl^2 = dx^2 + dy^2$ that we had earlier but now written in terms of the general coordinates u, v).

In 1854 Riemann showed how to generalize Gauss's intrinsic geometry of surfaces to higher dimensions. The reader might be puzzled about the motivations here. Why might mathematicians be interested in higher-dimensional intrinsic geometry? Ordinary space has just three dimensions and it is hard to see how to make sense of 'flexing' a three-dimensional 'surface' – let alone a higher-dimensional surface – within it. The first point to make is that this picture was helpful only for getting us started in understanding the notion of 'intrinsic geometry'. We should really be thinking of the intrinsic geometry of our surface as being something that stands on its own, without the need for an embedding space at all. Indeed, one of Riemann's original motivations was that the physical three-space within which we actually find ourselves might have a curved intrinsic geometry, without it having to 'reside' within some higher-dimensional space.

But Riemann also considered n-dimensional intrinsic geometries, and one might question the motivations for that. Two considerations are relevant here. It turns out that the mathematical formalism that has been developed for handling curved three-spaces is basically the same as that for handling curved n-spaces in general, so there is nothing to be gained by restricting attention to the case $n = 3$. The other consideration is that curved (intrinsic) n-geometry is important in many contexts where the n does not refer to the number of dimensions of ordinary space, but to the number of degrees of freedom of some system. There are abstract mathematical spaces known as 'configuration spaces', a single point of which represents the entire arrangement of parts of some physical structure. The dimension n of such a space can be very large indeed, when the system has many parts, and Riemann's higher-dimensional geometry can be of great relevance to these spaces.

The notions of 'metric' and 'curvature' in the n-dimensional case are natural generalizations of those introduced by Gauss for ordinary two-dimensional surfaces, but because of the large number of components involved, we need a suitable notation in order to handle them all. In place of the three 'metric components' A, B and C that occur in the above expression for dl^2 in the two-dimensional case, we need *six* such quantities for three dimensions. These are the components of the *metric tensor*, generally denoted g_{ab}. This quantity serves to define the appropriate notion of a 'distance' between neighbouring points, frequently denoted by ds.[4]

In Riemann's geometry, we obtain the *length of a curve* in the space by *integrating* ds along the curve in just the same way as in the flat-space case

already discussed. A geodesic in a Riemannian manifold is a curve that (locally) minimizes length (so it describes 'the shortest distance between points', in an appropriate sense). The *curvature* of the Riemannian space is the quantity that describes the amount of geodesic deviation in all the various possible directions in the space (as indicated above). Not surprisingly, there are many components to the curvature, there being lots of possible directions in which this geodesic deviation may be measured. In fact, all this information can be collected together in the quantity called the *Riemann tensor*. The Riemann tensor (or its collection of components) is commonly written R_{abcd}, where those little indices refer to all the different possible ways in which the geodesic deviation might be measured.[5]

Einstein's general relativity is formulated in terms of a concept of curved four-dimensional space-time which bears the same relation to Minkowski's flat space-time as Riemann's concept of curved geometry bears to flat Euclidean geometry. The metric g_{ab} can be used to define curve lengths, but as in Minkowski's flat space-time geometry, this 'length' is best thought of as defining the *time*, as measured by a particle along its world-line. Those world-lines that locally maximize this time measure are the geodesics in space-time and are considered to be the world-lines of inertially moving particles (where 'inertial' is taken in Einstein's sense of 'freely moving under gravity', as already described).

Now we recall that the geodesic deviation in space-time that is caused by gravitation (in Newtonian theory) has the property that in vacuum there is initially no volume change, whereas when there is matter present in the vicinity of the deviating geodesics, the *volume reduction* is proportional to the total mass that is surrounded by the geodesics. This volume reduction is an *average* of the geodesic deviation in all directions surrounding the central geodesic (this central one being the astronaut A's world-line). Thus, we need an appropriate entity that measures such curvature averages. Indeed, there is such an entity, referred to as the *Ricci tensor*, constructed from R_{abcd}. Its collection of components is normally written R_{ab}. There is also an *overall* average single quantity R, referred to as the *scalar curvature*.[6] We recall that R_{ab} and R, together with g_{ab}, are precisely the things that appear on the left-hand side of Einstein's equation.

The quantities g_{ab}, R_{abcd}, and R_{ab} are (sets of components of) entities called tensors, and tensors are fundamentally important in the study of Riemannian geometry. The reason for this has to do with the fact that in this subject, one is not really interested in the specific choice of coordinates that happen to be used for a description of the manifold. (This is an implication

of a strict adherence to the principle of equivalence.) One set of coordinates may be used, or another set may be used equally well. It is just a matter of personal convenience. The *tensor calculus* was a marvellous technical achievement, developed in the late nineteenth century by several mathematicians as a means of extracting *invariant information* about the manifold, its metric and its curvature, where 'invariant' essentially means 'independent of any particular choice of coordinates'.

In Einstein's deliberations about how to incorporate the principle of equivalence fully into a physical theory of gravitation, he eventually realized that he needed a formulation that is 'invariant' in the sense referred to above. He called this requirement the *principle of general covariance*. The space-time coordinates that refer to two differently accelerating frames of reference can be related to each other in some (often complicated) way, and neither is 'preferred' over the other. Einstein had to enlist the help of his colleague Marcel Grossmann to teach him what he needed to know of the 'Ricci calculus' (as the tensor calculus was then called). The only essential difference between the curved space-time geometry that he required and the Riemannian geometry that the Ricci calculus was designed for (in the four-dimensional case) was the change in 'signature' that was needed in passing from the locally Euclidean structure of Riemannian spaces to the locally Minkowskian structure needed for a relativistic space-time.

Full General Relativity

Let us return to Albert, our astronaut A surrounded by a sphere of particles. All these particles, as well as A, are moving inertially, in Einstein's sense (i.e. freely under gravity), and he postulated that inertially moving particles should have world-lines that are geodesics in space-time.[7] We recall that the initial volume reduction of this sphere is proportional to the mass enclosed, in Newtonian theory, and that it is the Ricci tensor that measures this volume change. Accordingly, we may expect that the appropriate relativistic generalization of Newton's theory would be one in which there is an equation relating the Ricci tensor of space-time to a tensor quantity that appropriately measures the mass density of matter. The latter quantity is what is referred to as the *energy–momentum tensor*, and its family of components is normally written T_{ab}. One of these components measures the mass-energy density; the others measure momentum densities, stresses and pressures in the material.

There is a factor of proportionality, in Newton's theory, between the inward acceleration and the mass density, that is Newton's gravitational constant G. This led Einstein to anticipate something like the equation

$$R_{ab} = -4\pi G\, T_{ab}.$$

The 4π comes from the fact that we are dealing with densities rather than individual particles, the minus sign coming from the fact that the acceleration is inwards, where my own conventions for the sign of the Ricci tensor are such that outward acceleration counts positively – but there are innumerable different conventions about signs etc. in this subject.

This equation is indeed what Einstein first suggested, but he subsequently came to realize that it is not really consistent with a certain equation,[8] necessarily satisfied by T_{ab}, which expresses a fundamental *energy conservation* law for the matter sources. This forced him, after several years of vacillation and uncertainty, to replace the quantity R_{ab} on the left by the slightly different quantity $R_{ab} - \frac{1}{2}Rg_{ab}$ which, for purely mathematical reasons, rather miraculously *also* satisfies the same equation as T_{ab}! By this replacement, Einstein restored the necessary consistency of the resulting equation, which is his now justly famous and very remarkable *Einstein equation*[9]:

$$R_{ab} - \tfrac{1}{2}R\, g_{ab} = -8\pi G\, T_{ab}.$$

This 'volume reduction' in the geodesic deviation that this equation gives rise to is just slightly different from what we expect from Newtonian theory, because of the additional term '$-\frac{1}{2}Rg_{ab}$' that now occurs in the left-hand side of the above equation. Instead of the 'source of gravity' (i.e. source of volume reduction) being simply $4\pi G$ multiplied by the *mass* density (in the sense of the mass-energy term in T_{ab}), it now turns out to be $4\pi G$ multiplied by the mass density *plus the sum of the pressures* in the material, in three mutually perpendicular directions (coming from other components of T_{ab}). For ordinary materials, like that composing ordinary stars and planets, the pressures are very small as compared with the mass densities (because the constituent particles of such bodies move slowly in comparison with the speed of light), so agreement with Newtonian theory is very precise. There are, however, certain circumstances (such as with the instability of supermassive stars, as they collapse to become black holes) in which this difference actually has important effects.

Classical Tests of General Relativity

It might seem from the preceding discussion that Einstein's general relativity is simply a technical modification of Newtonian theory, the latter having been rephrased so that it is in accordance with relativity and the principle of equivalence. Indeed, this could be said to be the case, although the way in which I have presented the comparison with Newtonian theory is not the way in which this was originally done. By concentrating on the tidal force of Newtonian gravity as something that cannot be eliminated in free fall, we have been able to see more directly its relationship to space-time curvature and, therefore, to the framework of Einstein's general relativity.

In fact, it is remarkably hard to find clear-cut observational differences between the two theories. Originally, there were the so-called 'three tests' of general relativity. The most impressive of these three was the explanation of the perihelion advance of the planet Mercury in its orbit around the Sun. It had been known from work in the nineteenth century that there was a curious discrepancy with Newtonian theory in Mercury's motion. When the perturbing effects of all the other known planets are taken into account, there is still a slight extra component to Mercury's motion, amounting to a swing in the axis of its orbital ellipse of 43 seconds of arc per century. This amount is so tiny that it would take about 3 million years for the ellipse of Mercury's orbit to swing completely around owing to this effect alone. Astronomers had tried various explanations, including the prediction of another planet within Mercury's orbit, which they had christened Vulcan. None of these ideas worked, but Einstein's theory exactly accounted for the discrepancy, and provided a rather impressive test of the theory.[10] The other two tests concerned the slowing of ideal clocks in a gravitational field and the bending of light by the Sun's field. The clock-slowing effect was convincingly confirmed by an experiment by Pound and Rebka in 1960, although it was recognized that this was a rather weak test of general relativity, being a direct consequence of energy conservation and the equation $E = hf$ for the energy of a photon.

The light-bending effect has a more interesting history. Before he had found the full general relativity, Einstein had used considerations from the principle of equivalence to predict, in 1911, that the Sun would bend light by an amount that is only one half of what the full theory actually predicts. This effect should be observable during a favourable solar eclipse and it was proposed to make an expedition to the Crimea in 1914 to test Einstein's 1911 version of his theory. From Einstein's point of view, it was fortuitous

that World War I prevented the expedition from taking place. By the time Arthur Eddington led a corresponding expedition to the Island of Principe to view light bending during the eclipse of 1919, Einstein fortunately had found, in 1915, the correct theory, and the observations were hailed as a triumph for that theory. In the light of modern analysis, these observations may be regarded as less convincing than they were thought to be at the time, when they were taken as a resounding success for Einstein's theory. Nevertheless, modern observations of this effect, and of a related time-delay effect noted by Shapiro, supply convincing support for Einstein's prediction.

Einstein's light bending is now so well established that it is used as a very impressive tool for observational astronomy and cosmology. Distant galaxies provide complicated lensing influences on even more distant light sources. This can give important information, not reliably obtainable in any other way, concerning the distribution of mass in the universe. Einstein's prediction has been turned around to provide a superb probe of matter in the distant universe.

Gravitational Waves

One of the most striking predictions of Einstein's theory is the existence of *gravitational waves*.[11] Maxwell's theory of electromagnetism had led to the prediction that waves of oscillating electric and magnetic field should be able to propagate through space at the speed of light, and Maxwell had postulated, in 1865, that light itself is an effect of this nature. Maxwell's prediction is now thoroughly confirmed in many experimental situations. Einstein's theory of gravity has many similarities with Maxwell's theory of electromagnetism, one being the existence of corresponding gravitational waves, these being distortions of space-time that propagate with the speed of light. Such waves would be emitted by gravitating bodies in orbit about one another, but the effect is generally very small. In our solar system, the largest emission of energy in the form of gravitational waves comes from the motion of Jupiter about the Sun. The amount of this energy loss is only about that in the light of a 40-watt light bulb!

In fact (perhaps partly owing to the influence of his colleague the esteemed Polish physicist Leopold Infeld) Einstein seems to have wavered in his belief that a freely gravitating system might actually lose energy in the form of gravitational waves. In the early 1960s, when I was first becoming

actively interested in Einstein's theory, there was a debate raging concerning this issue. At about this time, some important advances were beginning to be made in general relativity. For many years earlier, even stretching back to the time of the theory's conception, little interest was shown by serious physicists, and the subject was thought of as rather a playground for pure mathematicians. But in the early 1960s something of a renaissance of interest in general relativity occurred. In particular, the work of several theoreticians provided what to me was a convincing demonstration of the existence and generation of gravitational waves as a *real* physical phenomenon, the energy loss due to these waves being in accord with a formula that Einstein had put forward much earlier, in 1918.

In more modern times Einstein's theory has acquired an extraordinary boost from the observations (and theoretical analysis) of Joseph Taylor and Russell Hulse. In 1974 they first observed pulsar signals from the double neutron-star system PSR 1913+16. The variations in these signals give detailed information about the masses of the stars and their orbits, and one can cross-check this information with what general relativity predicts. There is an extraordinary overall agreement between theory and observation. Over the twenty-five-year period during which this system has been observed, there is a precision in the timing of the signals to roughly one part in 10^{14}; that is, one part in one hundred million million. To a first approximation, this gives a check on the Newtonian orbits of the stars. To a second, there is detailed confirmation of the general-relativistic corrections to the orbits (of the nature of that which occurs with the perihelion advance of Mercury). Finally, the loss of energy from the system in the form of gravitational waves, which is predicted by Einstein's theory, is seen to be in precise agreement with the theory. In 1993 Hulse and Taylor were awarded the Nobel Prize for Physics for the discovery and analysis of this remarkable system. From its uncertain beginnings, when general relativity had seemed an outlandish and rather flimsily supported theory, it now stands in extraordinary agreement with observation. In this one instance at least, a physical theory appears to be in detailed accord with nature to a precision greater than that which has been ascertained for any other individual physical system.

The existence of gravitational waves seems very well established in the PSR 1913+16 system. But such waves have not yet been convincingly observed *directly* here on Earth. There are several detectors in various stages of construction which should be able to observe such waves in the future. Moreover, the totality of these detectors, at different locations about the globe, should, within a few years, provide us with a remarkable Earth-scale gravitational-wave telescope able to obtain information about cataclysmic events (such as collisions between black holes and the like) occurring in very distant galaxies. This should give a completely new kind of window on the universe in which gravitational waves replace the usual electromagnetic ones. As with the light-bending effect, Einstein's prediction of gravitational waves may thus be turned around to provide a wonderful new observational tool to tell us something important about the distant universe.

Some Difficulties with General Relativity

We have now seen something of the extraordinary successes of general relativity. What about its limitations? The traditional view of the subject has been that its equations are notoriously difficult to solve. Indeed, despite the relatively very simple appearance of the Einstein equation, it hides a very considerable complication that is revealed when the expression $R_{ab} - \frac{1}{2}Rg_{ab}$ is written out explicitly in terms of the components g_{ab} and their first and second partial derivatives with respect to the coordinates. For many years only few solutions of the equations were known explicitly, but more recently numerous mathematical devices have been employed to find hosts of different solutions. Many of these are of mainly mathematical interest and do not directly relate to situations of particular physical relevance. Nevertheless, quite a lot is now known from the nature of exact solutions concerning, in particular, rotating bodies, black holes, gravitational waves and cosmology.

This notwithstanding, it is still hard to find particular exact solutions that describe situations that one may be interested in. Most notorious among these issues is the 'two-body problem': find an exact solution of Einstein's equation describing, say, two stars in orbit about one another. The difficulty here is that owing to the emission of gravitational waves, the two would spiral in towards one another, whence the situation possesses no symmetry. (The presence of symmetry is a great help in solving equations generally.) In fact, the difficulty in finding exact solutions to equations in physics is not now regarded as a particular limitation on a physical theory. With the advent of modern high-speed electronic computers, physicists can often get a much better picture of the evolution of the equations from a numerical simulation than they might obtain from an explicit exact solution. Considerable efforts have been devoted to developing computer techniques in general relativity, and some very good progress has been made.

Some of the main problems involved in solving the Einstein equation are of a rather different kind from sheer complication, however, and arise from an ingredient that is specific to general relativity: the principle of general covariance. Thus, when a solution is found, by computation or by analytical methods, it may not be clear what the solution *means*. Many features of the solution might merely reflect some aspect of the particular choice of coordinates, rather than expressing something of interest concerning the physics of the problem. Techniques have been evolved for answering such questions, but much more needs to be done in this area.

Finally, there is the profound issue of *singularities* in solutions of Einstein's equation. These are places where the solution 'diverges', thereby giving infinite answers rather than something physically sensible. For many years, there was great confusion in the subject because such singularities may turn out to be 'fictional'; that is to say, they may be merely the result of some inappropriate choice of coordinates rather than of some genuinely singular feature of the space-time itself. The most famous example of this kind of confusion occurred with the renowned *Schwarzschild solution* – the most important of all solutions of the Einstein equation. It describes the static gravitational field surrounding a spherically symmetrical star, and was found by Karl Schwarzschild in 1916 as he lay dying from a rare disease contracted on the eastern front in World War I, the same year that Einstein published his first full account of general relativity. At a certain radius, now known as the *Schwarzschild radius*, a singularity appeared in the metric components, and this region of the space-time used to be referred to as the 'Schwarzschild singularity'. People did not tend to worry much about this singularity, however, because the region would normally lie far beneath the surface of the star where, owing to the presence of a matter density (the T_{ab} of Einstein's equation), Schwarzschild's solution would cease to hold. But in the 1960s, the discovery of quasars led astronomers to wonder whether some highly compressed astrophysical bodies, as small as the scale of their Schwarzschild radii, might actually exist.

In fact, as early as 1933, Monseigneur Georges Lemaître had shown that with an appropriate coordinate change, the singularity at the Schwarzschild radius can be seen to be fictitious. Accordingly, this non-singular region is now *not* called a singularity, but is referred to as the Schwarzschild *horizon* – of a black hole. Indeed, any body that is compressed down to smaller than its Schwarzschild radius must collapse inwards towards the centre and a black hole is the result. No information can escape from within the Schwarzschild radius, which is why it is now referred to as a 'horizon'.

Space-Time Singularities

At this point it may be appropriate to relate how I myself became professionally involved with general relativity. In the late 1950s I was a young research fellow at St John's College, Cambridge. My official area of interest was in pure mathematics, but a friend and colleague of mine, Dennis Sciama, had taken it upon himself to acquaint me with many of the exciting things that were going on in physics and astronomy. I had had a significant but amateur interest in general relativity, since that subject was something that could be comprehended, and its beauty appreciated, by someone such as myself, with merely a love of geometry and an appreciation of the relevant physical ideas. Although Dennis had fired my interest in physics, I had not thought of general relativity as a subject into which I would research in a serious way, mainly because I had thought of it as somewhat peripheral to the main concerns of the fundamental quantum physics of the small-scale universe.

Nevertheless, probably some time in 1958, Dennis persuaded me to accompany him to attend a seminar given in London by David Finkelstein. This was on the extension of the Schwarzschild solution through its Schwarzschild radius. I remember being particularly struck by this lecture, but I had been troubled by the fact that although the 'singularity' at the Schwarzschild radius had been eliminated by a change of coordinates, the singularity at the *centre* (zero radius) still remained, and could not be removed in this way. Might it be, I had thought to myself, that there is some underlying principle that prevents the complete elimination of singularities from a broad class of solutions of the Einstein equation, including that of Schwarzschild?

Upon returning to Cambridge, I tried to think about this problem, though I was completely inadequately equipped to tackle it. At the time, I was concerning myself with a formalism known as the 2-spinor calculus, which has application to the study of spinning quantum particles. My pure-mathematical work had led me to study the algebra of tensors in a rather general way, and I had become intrigued by 2-spinors, because they seemed to be, in some sense, the square roots of vectors and tensors. In a certain clear sense, 2-spinors constitute a system that is even more primitive and universal in the description of space-time structures than that provided by tensors. Accordingly, I tried to see whether the employment of spinors might provide novel insights into general relativity, and whether these might be of use for the singularity problem.

Although I did not find that spinors told me much about singularities, I did find that they meshed extraordinarily well with the Einstein equation itself, providing unexpected insights that are not easy to come by, by other means. The elegance of the resulting expressions was striking, and I was hooked! For the ensuing forty-two years, general relativity has been one of my deepest passions, particularly in relation to its affinity to certain unusual mathematical techniques.

In 1964 I became interested in the singularity problem again, largely because John A. Wheeler pointed out that recent observations of those objects now known as quasars indicated that the Schwarzschild radius is being approached by actual astrophysical objects. Could the singularity that arises as a result of the collapse of a body down *through* that radius – the one at the centre that I had worried about in Finkelstein's lecture – actually be avoided? The exact solution of such a collapse (now referred to as a black hole), as found by Oppenheimer and Snyder in 1939, indeed possessed a genuine singularity at the centre. But a crucial assumption of their model was exact spherical symmetry. It could well be imagined that with irregularities present, the infalling matter might *not* simply be focused to an infinite-density singularity at the centre, but might instead pass through a complicated central configuration to be flung outwards again, and no actual singularity might be the result.

My earlier worries that such singularities may be inevitable had led me to doubt such a possibility, and I began to wonder whether some ideas that I had been more recently playing with, involving qualitative topological considerations – rather than the usual direct attempts at exact solution of Einstein's equation – might be able to resolve this issue. In due course, this unorthodox line of thinking led me to the first 'singularity theorem' of physical relevance in general relativity, which showed that, under some very reasonable general assumptions, *any* gravitational collapse to within a region that qualitatively resembles the Schwarzschild radius (but with no special assumptions of symmetry) results in a genuine space-time singularity.

Later work by Stephen Hawking, and by the two of us together, generalized this result, showing that in addition to the black hole situation, such singularities are also inevitable in the Big Bang origin of the universe, irrespective of any symmetry assumptions. The standard cosmological models derive from the original cosmological solutions to Einstein's equation found in 1922 by the Russian Alexander Alexandrovich Friedmann. Here, exact spatial homogeneity and isotropy is assumed, and the solution expands away from the initial Big Bang singularity. What the singularity theorems

show is that we cannot eliminate the Big Bang singularity just by dropping the symmetry assumptions of homogeneity and isotropy.

All this is dependent upon the validity of Einstein's equation (and upon some physically reasonable assumptions concerning T_{ab}). Some people regarded these singularity theorems as revealing a profound shortcoming in Einstein's general relativity. My own attitude is somewhat different. We know, in any case, that Einstein's theory cannot be the last word concerning the nature of space-time and gravity. For at some stage an appropriate marriage between Einstein's theory and quantum mechanics needs to come about. What the singularity theorems reveal is an inner strength in Einstein's classical theory, in that it points clearly to its *own* limitations, telling us where we must look for an extension into a quantum world, and telling us also something of what to expect from an eventual quantum/gravitational union. We shall try to glimpse something of this in the next section.

The Beginning and Ends of Time

In the discussion above, we have caught sight of two situations in which space-time singularities arise in Einstein's theory: in gravitational collapse to a black hole and in the Big Bang origin of the universe. It seems clear that Einstein was very unhappy about both of these seeming blemishes to his theory. He appears to have been of the opinion that realistic departures from the high symmetry that is assumed in the standard exact solutions ought to lead to *non*-singular solutions. Unfortunately we shall never know what his reaction to the singularity theorems would have been, but apparently one of his reasons for trying, in his later years, to generalize general relativity to some kind of 'unified field theory' was his attempt to arrive at a singularity-free theory.

Initially he favoured a spatially closed-up universe that is *static* – so it would remain unchanged for all time. He found that he could achieve this only by introducing (in 1917) a *cosmological constant* Λ into his equation, which then becomes

$$R_{ab} - \tfrac{1}{2}R\, g_{ab} + \Lambda\, g_{ab} = -8\pi \mathrm{G}\, T_{ab}.$$

Later, he regarded this modification as his 'greatest mistake'. If he had not insisted on a static model, but just let his original equation carry things along so as to obtain the Friedmann picture of a universe expanding away

from a 'big bang', then he would probably have predicted the expansion of the universe, which was actually discovered observationally by Edwin Hubble in 1929.

There is much discussion today of whether the observational evidence now actually favours the existence of a (very small) cosmological constant. Some cosmologists (especially the proponents of what is referred to as the 'inflationary scenario') claim that such a constant is necessary in order to fit recent observations. Yet there are some seeming contradictions as things stand, and it will be better to wait until the dust settles before coming to any clear conclusions about this.

According to my own perspective on these issues, while we must be cautious about claims concerning the observational status of the large-scale universe, we must accept that the Big Bang and black-hole singularities are indeed part of nature. Rather than shrinking from them we must try to learn from them something of the 'quantum geometry' that should ultimately replace them. What can we learn? Although little is known in detail about the nature of singularities, some general comments can be made.

The first is that although unstoppable gravitational collapse must sometimes occur (such as with a supermassive star or collection of stars at a galactic centre), we do not know for sure that a black hole would be the result even though the singularity theorems tell us to expect space-time singularities. There is a still unproved assumption, referred to as 'cosmic censorship' (that I pointed out in 1969), which asserts that the resulting singularity cannot be 'naked', which means in effect 'visible from the outside'. If naked singularities do not occur, then a black hole must indeed be the result. (In any case, naked singularities would be, in a clear sense, 'worse' than black holes!) A black hole swallows material in its immediate vicinity and (cosmic censorship being assumed) destroys it all in the singularity at the centre. To the infalling material, this singularity represents the 'end of the universe', and it plays a role like a big bang reversed in time.

Despite this particular unpleasant feature, the *exterior* space-time to a black hole has a large number of very elegant properties. Moreover, large black holes appear to lie at the centres of virtually all galaxies, and the extraordinary physics that sometimes goes on in their immediate neighbourhoods seems to be responsible for the stupendous energy output of quasars, which can easily outshine entire galaxies. They also represent the regions of highest entropy known in the universe, and a famous formula due to Bekenstein and Hawking tells us exactly what that entropy should be in terms of the surface area of the hole's horizon.

In effect, cosmic censorship may be interpreted as telling us that there are just two kinds of space-time singularity in the universe, the *past* type (in the Big Bang) and the *future* type (in black holes). Matter is created at the past-type singularity and it is destroyed at the future-type ones. At first sight, these two types of singularity would appear to be simply time-reverses of each other. However, when we examine this in a little more detail, we find a gross distinction between these two types of singularity. This is related to the enormously large entropy of black holes. In everyday terms, 'entropy' means 'disorder', and the famous second law of thermodynamics tells us that the entropy of the universe increases with time. It turns out that the physical origin of the second law can be attributed to this gross asymmetry between past- and future-singularity structure, where the past-type singularities are particularly special and simple, whereas the future-type ones are general and extraordinarily complicated. Using the Bekenstein–Hawking formula for black-hole entropy, one can conclude that the 'specialness' of the Big Bang was quite stupendous, namely to one part in at least $10^{10^{123}}$.

Quantum Gravity?

Where does this gross time-asymmetry in space-time-singularity structure come from? The issue continues to stir up much controversy, but my own view is that there is a clear implication that the 'quantum gravity' that is supposed to account for the detailed nature of space-time singularities must be time-asymmetrical. I am continually amazed by the fact that so few workers in the area of quantum gravity seem to have come to the seemingly obvious conclusion that whatever the nature of this still-missing 'quantum gravity' theory may be, it *must* be a fundamentally time-asymmetric scheme. It is true that Einstein's equation is symmetrical under reversal of time, and so also is the Schrödinger equation which governs the evolution of a quantum state. Accordingly, any 'conventional' application of the rules of quantum mechanics to Einstein's theory ought to lead to time-*symmetrical* conclusions. In my own opinion, this provides a clear indication that the sought-for 'quantum gravity' must be an *unconventional* quantum theory, according to which the rules of quantum mechanics must themselves be expected to *change*. This is in addition to changes that must in any case be expected to take place in the classical rules of Einstein's general relativity. Thus I agree with Einstein in his belief that quantum mechanics is incomplete.

However, this is not the position of the great majority of those who attempt to combine quantum theory with general relativity. Despite the wealth of unusual and fascinating ideas that have been put forward as candidates for a 'quantum gravity' theory – such as 'space-times' of ten, eleven or twenty-six dimensions and ideas involving supersymmetry, strings, etc. – none of these candidates takes on board the possibility that the very rules of quantum mechanics may have to change. In my own view (and in the view of a sizeable minority of researchers into the foundations of quantum mechanics), changes in the rules of quantum theory are to be expected in any case, because of what is known as the 'measurement problem'.

What is the measurement problem? For this, we need to understand a little of the actual rules of quantum theory. There is a mathematical quantity referred to as the *quantum state* (or wave function), frequently labelled Ψ, which is supposed to contain all the necessary information defining the quantum system under consideration. The time-evolution of the state Ψ is governed by Schrödinger's equation until a measurement is made on the system, whereupon the state *jumps* (randomly) to one of a set of allowed possibilities defined by the specific measurement being performed. This 'jumping' does not take place in accordance with Schrödinger's equation, however, and the measurement problem is to understand how this random jumping comes about, given that the state is supposed actually to evolve by the deterministic Schrödinger equation.

I believe that a strong case can be made that the pure Schrödinger equation does *not* apply rigorously at all scales, and needs modification when gravitational effects begin to become significant. Accordingly, such a modification would necessarily be part of the correct 'quantum gravity' theory. Moreover, the measurement problem would find its resolution within this 'correct quantum gravity' theory. One of the main reasons for believing this comes from strong arguments that point out a fundamental conflict between the principle of general covariance and the basic principles of standard Schrödinger wave function evolution. According to this reasoning, quantum jumping (which I take to be a physically *real* phenomenon rather than the 'illusion' that it is often assumed to be) comes into play as a feature of the resolution of this conflict.[12] Now whatever form this modification of Schrödinger's equation would take, it would have to be *time-asymmetric*, and a gross asymmetry between past and future singularities would be expected, in accordance with the arguments I have given here.

As things stand, no plausible such modification of Schrödinger's equation has yet come to light, so a unification of quantum theory with general

relativity along these lines remains as elusive as a unification along any of the more conventional lines that have so far been suggested. Finding the correct unification presents the twenty-first century with one of its greatest challenges. If this challenge is successfully met, then it will have profound implications running far beyond those that we can directly perceive at the moment. It will not be met, however, if the strange and wonderful principles underlying Einstein's beautiful equation are not thoroughly respected.

On Wheeler and the H-bomb, Peter Galison, *Image and Logic: A Material Culture of Microphysics* (University of Chicago Press, 1997).

Situating Einstein within the tradition of electrodynamics, see Olivier Darrigol, *Electrodynamics from Ampère to Einstein* (forthcoming).

A short but very useful volume with translations of the five papers of 1905 is: John Stachel, *Einstein's Miraculous Year: Five Papers that Changed the Face of Physics* (Princeton University Press, 1998).

Einstein's own popularization of his theory is obviously also to be valued: Albert Einstein, *Relativity: The Special and the General Theory* (New York: Crown Publishers, 1961).

Einstein's essays on everything from politics to the philosophy of physics: Einstein, *Ideas and Opinions* (New York: Bonanza Books, 1954).

And my favourite elementary textbook on special is still, after all these years: A. P. French, *Special Relativity*, the M.I.T. Introductory Physics Series (New York: W. W. Norton, 1968).

The Rediscovery of Gravity
The Einstein Equation of General Relativity

Notes

1 It is a common modern practice to refer to this equation in the singular rather than the plural, as had been usual originally, because it is better to think of it as a single equation on the entire tensors that are involved (see the section on curved space-time geometry below), rather than on the family of components of those tensors.

2 Technically this number is called the *intrinsic* curvature, or *Gaussian* curvature, of the surface. We shall be coming to the notion of 'intrinsic' a little more fully later on.

3 The full mathematical theory of the type of four-dimensional geometry that is involved for Newtonian theory was first worked out by the outstanding French mathematician Élie Cartan in 1923/24.

4 The previous Euclidean expression for dl^2 now generalizes to the well-known form $ds^2 = g_{ab}dx^a dx^b$ (our 'dl' being written 'ds' in most literature). For an n-dimensional space we need n independent coordinates, here denoted $x^1, x^2, \ldots, x^n$. This may be a bit confusing, because the notation 'x^2' does *not* now stand for 'x squared', nor does 'x^3' stand for 'x cubed', etc. The notation 'x^a' (or 'x^b', etc.) is a generic symbol for one of these coordinates. Similarly, 'g_{ab}' is a generic symbol for one of the quantities $g_{11}, g_{12}, \ldots, g_{nm}$, these being $n(n+1)/2$ independent functions because $g_{ab} = g_{ba}$. The *Einstein summation convention* is being adopted here, according to which repeated indices get summed over. Hence the expression '$g_{ab}dx^a dx^b$' stands for '$g_{11}dx^1dx^1 + g_{12}dx^1dx^2 + \ldots + g_{nm}dx^n dx^n$'. In our two-dimensional case, $g_{11} = A$, $g_{12} = g_{21} = B$ and $g_{11} = C$ are functions of the two coordinates u and v, where $x^1 = u$ and $x^2 = v$.

5 There are explicit but complicated expressions telling us how to calculate the R_{abcd} from the g_{ab} and their first and second partial derivatives, with respect to the coordinates x^a.

6 Using Einstein's summation convention we can define, R_{ab} and R by the relations $R_{ab} = R_{acbd}g^{cd}$ and $R = R_{ab}g^{ab}$, where g^{ab} is the *inverse* of g_{ab}, in the sense of matrix algebra.

7 In fact Einstein later showed that this postulate can be *deduced* from his field equation, together with some other reasonable assumptions.

8 The vanishing of the 'covariant divergence' of T_{ab}.

9 The mathematician David Hilbert also came upon this equation at a similar time to Einstein, but by a different route, in the autumn of 1915. This has resulted in an uncomfortable priority dispute. But Hilbert's contribution, though technically important, does not really undermine Einstein's fundamental priority in the matter. See, in particular, J. Stachel (1999), *New Light on the Einstein–Hilbert Priority Question* in Journal of Astrophysics and Astronomy, Volume 20, Numbers 3 and 4, December 1999, 91–101.

10 There was a curious 'scare' in 1966 when Robert Dicke claimed that careful observations of solar oblateness by himself and Goldenberg showed that the Sun possessed a quadrupole moment of such a magnitude that it would thoroughly spoil the agreement of Mercury's perihelion advance with general relativity. Fortunately, subsequent observations and theoretical considerations showed that Dicke's conclusion was wrong.

11 Interestingly, although he did not have the basic ideas of general relativity, Poincaré had already predicted the existence of gravitational waves in 1905, based on analogies with Maxwell's theory of electromagnetism.

12 I have proposed a technically difficult but apparently feasible experiment, one version of which would have to be performed in outer space, for testing whether or not this proposal is correct.

Further Reading

W. Rindler, *Relativity: Special, General and Cosmological* (Oxford University Press, 2001).

W. Rindler, *Essential Relativity* (New York: Springer-Verlag, 1997).

L. A. Steen (ed.) *The Geometry of the Universe*, in Mathematics Today: Twelve Informal Essays, (New York: Springer-Verlag, 1978).

K. Thorne and C. W. W. Norton, *Black Holes and Time Warps: Einstein's Outrageous Legacy*, (New York, 1994).

A. Einstein, *Relativity: The Special and the General Theory* (Reprinted by Three Rivers Press, California, 1995).

Erotica, Aesthetics and Schrödinger's Wave Equation

1 For biographical details, see W. Moore, *Schrödinger: Life and Thought* (Cambridge University Press, 1989).

2 See D. Cassidy, *Uncertainty: The Life and Science of Werner Heisenberg* (New York: Freeman, 1992).

3 Ibid, p. 137, as recalled by Max Born.

4 For biographical details, see A. Pais, *Niels Bohr's Times: In Physics, Philosophy and Polity* (Oxford University Press, 1991).

5 E. Schrödinger, 'Über das Verhältnis der Heisenberg–Born–Jordanschen Quantenmechanik zu der meinen', *Annalen der Physik*, 70, 734–56 (1926), p. 735. This is known as the third communication.

6 The quotes in this paragraph are from E. Schrödinger, 'Quantisierung als Eigenwertproblem', *Annalen der Physik*, 80, (1926) pp. 437–90, which is the second communication.