Foundations

by Greg Egan

2: From Special to General

The first article in this series described some of the ways in which the geometry of spacetime affects travellers moving (relative to their destinations, or each other) at a substantial fraction of the speed of light. By generalising from the Euclidean metric, which captures such familiar aspects of geometry as Pythagoras's Theorem, to the Minkowskian metric suggested by the fact that the speed of light in a vacuum is the same for everyone, we analysed the "rotated" view of spacetime that two observers in relative motion have with respect to each other, and derived formulas for time dilation, Doppler shift and aberration.

This article and the next will build the framework needed to provide a similar account of the strange effects that have been predicted to take place in the vicinity of a black hole. To do this, we need to generalise yet again: from flat geometry, to curved.

Gravity as Spacetime Curvature

The basic premise of general relativity is simple: the correct way to account for the acceleration of objects due to gravity is to consider spacetime to be curved in the presence of matter and energy. How does curvature explain acceleration? If two explorers set off from different points on the Earth's equator, and both head north, their paths will grow steadily closer together, despite the fact that they started out in the same direction. In spacetime, if two nearby stars start out being motionless with respect to each other, their world lines will draw closer together, despite the fact that those world lines were initially pointing in the same direction. We could say that the force of gravity is pulling the stars together … but we don't say there's a "force" acting on the explorers, do we? Of course, the two-dimensional surface of the Earth is a visibly curved object embedded in a larger (and more or less flat) space, but we have no reason to believe that spacetime is embedded in anything larger. Rather, general relativity assumes that whatever gives rise to spacetime geometry in the first place is tied up with the presence of matter and energy in such a way that the resulting geometry is sometimes curved.

Manifolds

Before exploring curved geometry, it will be useful to take a look at a kind of geometry that's neither flat nor curved: geometry without any metric at all. Essentially, this is like asking what you can say about lines drawn on a sheet of rubber that remains true

however much you stretch or squeeze the sheet: distances and angles lose all meaning, but you can still talk about such things as whether or not two lines intersect. Why is this relevant to general relativity, which *does* assign a metric to every part of spacetime? Firstly, everything that's true without reference to a metric can be safely carried over to regions of spacetime where the metric varies from place to place. Secondly, it reflects the situation you find yourself in when you begin to solve a problem in general relativity: initially, you have no idea what the metric is, since that's the very thing the equations are supposed to tell you.

Let's start with a familiar situation — two-dimensional space, with flat geometry — and see what concepts can survive the loss of the metric. Choose a city's central post office as the origin for coordinates, measure distances north and east of it, and ignore the curvature of the Earth. Every building in the city can be identified by a pair of numbers, (x,y), specifying distances east and north of the post office. Vectors associated with objects in the city can also be given coordinates in much the same way. For example, a train's velocity can be assigned two coordinates — call them v^x and v^y — stating how fast the train is travelling east, and how fast it's travelling north.

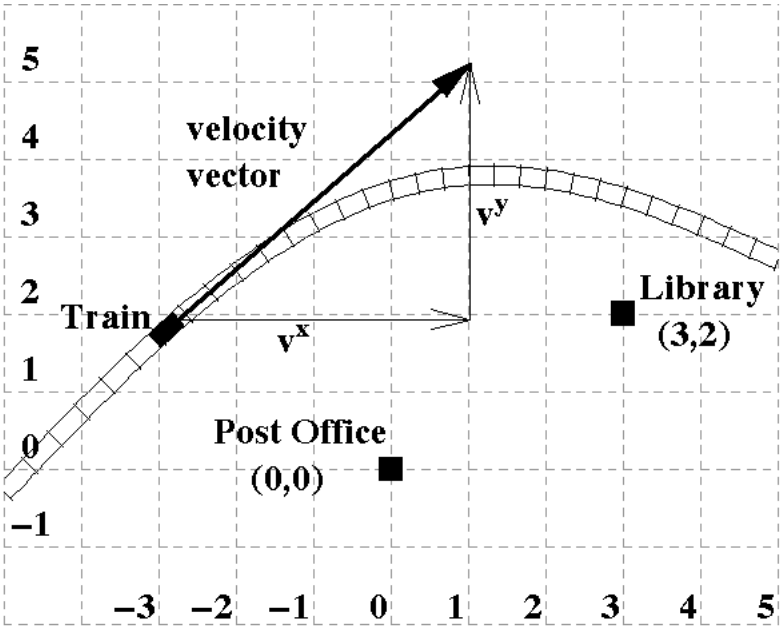

Fig 1: City with North/East Coordinates

Now, imagine that this city lies, not on solid rock, but on a vast sheet of rubber. The railway track and all the buildings are made of equally flexible material, so the whole city can be stretched and squeezed without anything being disrupted. What's more, the imaginary grid lines that initially measured distances north and east of the post office have been painted onto the ground, so they too can flex with it. Then a giant hand descends from the sky and gives one edge of the city a mighty tug.

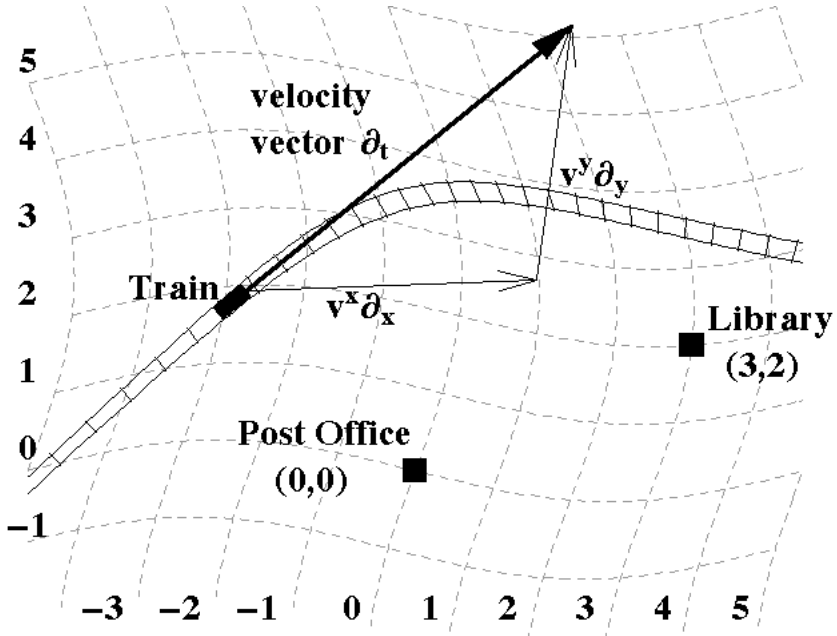

Fig 2: Distorted Version of City

Figure 2 shows the result. But the cosmic intervention doesn't stop here; further stretching and compression is applied, at random, 24 hours a day. The city dwellers simply have to adapt to the fact that streets no longer meet at the same angle from hour to hour, and buildings are no longer separated by fixed distances. These concepts are soon discarded as irrelevant.

The first thing to note is that the idea of assigning coordinates to every point doesn't have to be thrown out. It no longer makes sense to talk about measuring fixed distances in a fixed direction, but a coordinate grid painted on the ground can do just as good a job at identifying buildings, even if the numbers are now entirely arbitrary. The library's coordinates remain (3,2), whatever shape the city is in, simply by virtue of the fact that the building lies at the intersection of two lines called “x=3” and “y=2”.

What's more, if civilisation collapsed, the paint faded, and some later generation decided to construct their own new coordinates without ever having heard of distance, any scheme they adopted would be just as good as the old one, so long as it avoided certain pitfalls. For example, if two grid lines for different values of the x-coordinate intersected, that would leave the x-coordinate of the point of intersection undefined. We cope with this happening to longitude at the north and south poles, but even linguistically it's a bit of a nuisance. Sudden jumps in the value of a coordinate — like the jump from longitude 180° west to 180° east — would also add unwelcome complications. Of course, in the city these problems are easily avoided, but for the surface of the Earth as a whole (and many other examples) they turn out to be inevitable for any *single* set of coordinates. In such cases, the best that can be done is to use as many overlapping sets of **local coordinates** as necessary to cover the whole surface, each of which, individually, is suitably grid-like.

Any mathematical space upon which it's possible to “paint” locally well-behaved coordinates like this is known as a **manifold**. An idealised sheet of rubber is a two-dimensional manifold. So is the surface of the Earth; having a metric doesn't disqualify you, it's just not part of the definition. In general relativity, spacetime is assumed to be a four-dimensional manifold: at least locally, spacetime can always be given coordinates like those of a four-dimensional grid.

Returning to the city, one question we still haven't dealt with is the fate of the idea of a “vector”. Does this make any sense at all, without distances and angles? Surprisingly, it does. For example, we can still compare one velocity to another, at least at the same point. A train passing over twenty sleepers per second is, in a very real sense, moving *twice as fast* as a ghostly second train sharing exactly the same stretch of track but doing just ten sleepers per second; you don't need to worry about distances between the sleepers to know who's being left behind. And the train is certainly moving in *a different direction* than a car driving over a level crossing at the same location; the angle between the road and the railway line might be undefined, but no matter how the ground flexes to bring them together, the two different paths can't be made completely indistinguishable.

It's easy to assign coordinates, v^x and v^y, to the train's velocity: just observe the rate of change of the train's x- and y-coordinates. This tells you how fast it's passing over the city's coordinate grid lines, rather than how fast it's passing over sleepers along the track. In Euclidean space, this method agrees exactly with the usual way of splitting up a velocity into components; with the grid in Figure 1 it would yield the expected values in kilometres per hour. Without a metric, the values are just “coordinate units per hour”. The particular values of v^x and v^y depend on the coordinate system being used, but that's equally true with Euclidean coordinates: if you rotate your axes from north and east to some other orientation, you measure components of the velocity in different directions, so the values are different.

Still, the motion of the train along the track is something quite independent of any grid painted over the city, and it ought to be possible to characterise it on its own terms. In fact, there's a definition of a vector on a manifold that does this perfectly. Think of the familiar “number line” of high school mathematics, and imagine drawing part of it on a manifold, so it passes through some point P. There are lots of different ways you could do this: crossing into P from different neighbouring points, and crowding or stretching the numbers on the line by various degrees as you approach. (Even though there's no such thing as distance, once you pick a certain curve through P you can compare two ways of drawing the number line along that curve.) The combination of the *direction* in which the line passes through P, and the *rate* at which the numbers are changing at P, together define a unique **tangent vector** at P.

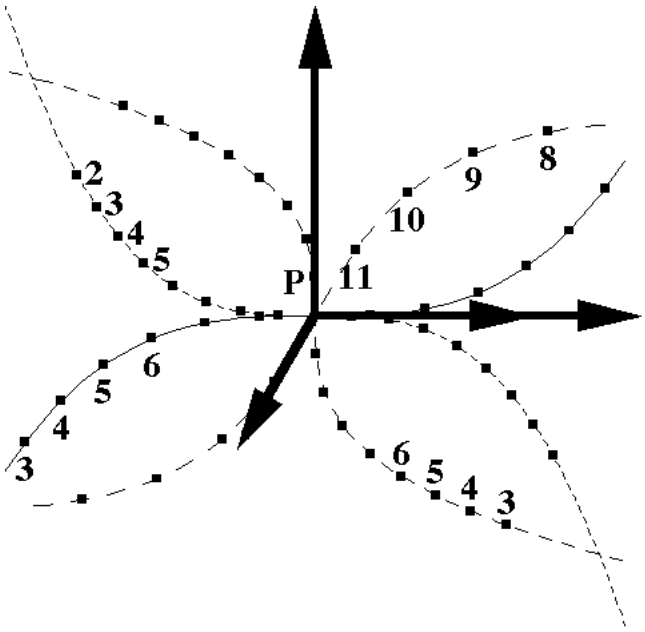

Fig 3: Some Tangent Vectors at a Point

Tangent vectors are often drawn as arrows of different lengths, tangent to the curves that define them. Figure 3 shows several such curves and arrows (successive dots mark successive integer values along the curves). The lengths of these arrows and the precise angles between them are arbitrary; without a metric, the most you can talk about meaningfully is the *ratio* of the size of one vector to another that's pointing in the same (or precisely the opposite) direction.

Velocity vectors are also drawn as arrows in Figures 1 and 2, and though this is a convenient thing to do, it's worth stressing that these arrows aren't “part” of the city, in the way that a road or a railway line is. Two different diagrams have actually been overlaid here, for convenience: on top of the drawing of the city is a drawing of the abstract “space” of velocity vectors for the train. In general, the vectors at each point in a manifold comprise what's known as the **tangent space** for that point.

How, exactly, do we make use of our new definition of vectors in terms of numbered curves? (Or to introduce the correct terminology, **parametrised curves**, with the numbers along the curve known as its **parameter**.) The railway track in Figure 2 is just like one of the curves in Figure 3, and the passage of the train along the track assigns a parameter to every point: t, the time when the train passes. (Think of a machine strapped to the front of the engine, time-stamping every sleeper.) And what we can do with such a curve is compute the rate of change of any other quantity that we associate with points in the city, with respect to the parameter t.

We've already done just that: in defining the velocity's coordinates, v^x and v^y, we took the rate of change (with time) of x and y, two numbers that label every point in the city. But there's no reason why we have to confine ourselves to coordinates. To make our city even more fanciful, assume that the ground magically drags the air along

with it when it's deformed, so that at any given moment, every point in the city will have a certain temperature, a certain air pressure, a certain carbon monoxide level. The train's velocity then offers a way of computing a rate of change of any one of these, as it ploughs across thermal (or pressure, or pollution) gradients.

A quantity with a value at every point in a manifold is called a **scalar field**. We write the rate of change of a scalar field, f, along a curve with parameter t, as $\partial_t f$. Don't be put off by the symbol; this is just shorthand for: "Look at two points on the curve, with t values t_1 and t_2, and f values f_1 and f_2. Then $\partial_t f$ is the value that the ratio $(f_2–f_1)/(t_2–t_1)$ approaches as the points get closer together".

Using this terminology, we can write:

$$v^x = \partial_t x \qquad \textbf{(1a)}$$

$$v^y = \partial_t y \qquad \textbf{(1b)}$$

for the coordinates of any vector, **v**: these are the rates of change of the ordinary manifold coordinates as we move along the curve that defines the vector. In cases other than velocity the curve parameter need not have anything to do with time, so in general the most that can be said about vector coordinates is that they're measured in "coordinate units per parameter unit".

In effect, the vector **v** *is* the operation described by the symbol ∂_t, since that captures everything about the train's motion (or whatever) at the point in question. If a vector **v** is defined by a curve with parameter t, we'll use **v** and ∂_t interchangeably, and write **v**(f) for the rate of change, $\partial_t f$, of a particular scalar field.

We can add and subtract vectors, or multiply a vector by a numeric factor, to yield a new vector. The simplest way to do this is to declare that the rate of change of any scalar field is **linear** in terms of the vector:

$$(a\mathbf{v} + b\mathbf{w})(f) = a\mathbf{v}(f) + b\mathbf{w}(f) \qquad \textbf{(2)}$$

for any numbers a and b, and any vectors **v** and **w**. We can then write any vector **v** as a linear expression in terms of **coordinate vectors**:

$$\mathbf{v} = v^x\partial_x + v^y\partial_y \qquad \textbf{(3a)}$$

and the rate of change for a particular scalar field, f, as:

$$\mathbf{v}(f) = v^x\partial_x f + v^y\partial_y f \qquad \textbf{(3b)}$$

What does this mean? The two coordinate vectors, ∂_x and ∂_y, are defined by grid lines,

rather than the time-stamped railway track; the symbols ∂_x and ∂_y just mean "take the rate of change of anything, with respect to x (along a curve of constant y), or with respect to y (along a curve of constant x)". Equation (3a) then tells us that a certain linear combination of those two rates of change will give the rate of change for the train as it moves along the railway line. To make this concrete, let's take the rate of change of air pressure, and rewrite Equation (3b) in English:

> (The rate of change of air pressure with time, for the moving train) equals (the rate of change of x with time, for the train) times (the rate of change of air pressure with x), *plus* (the rate of change of y with time, for the train) times (the rate of change of air pressure with y).

The advantage of describing **v** with Equation (3a), rather than just stating its coordinates (v^x,v^y), is that it's an expression for *the vector itself*, true regardless of the coordinate system. Sure, **v** is described in terms of particular coordinate vectors, but if we want to change coordinates, and if we know how to describe ∂_x and ∂_y in terms of the new coordinate vectors, we can just substitute those descriptions into Equation (3a).

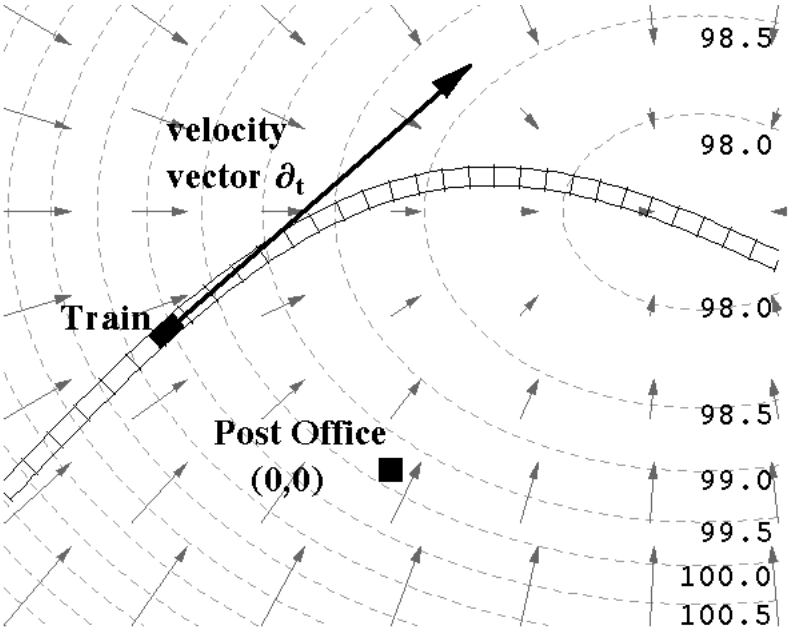

Fig 4: Contour Lines and Gradient Vectors

There's one more geometrical object on manifolds that we need to talk about, which arises directly out of the idea of scalar fields. Figure 4 shows a series of contour lines for air pressure over the city: lines along which the pressure is constant, or "isobars". The arrows crossing the isobars show the direction in which the pressure difference is driving the air (ignoring all real meteorological complications like the Coriolis force).

The arrows in Figure 4 are drawn perpendicular to the contour lines, so for a

moment let's stick to Euclidean space, where “perpendicular” actually means something. There are plenty of other examples where this idea is useful. Which way will a boulder roll at a certain point on a complicated hilly terrain? Draw contour lines for altitude, then draw an arrow perpendicular to the contour line passing through the boulder. There are two possible arrows you could draw; pick the one in the direction of decreasing altitude, and that's the way the boulder will fall. You can even give the arrow a length proportional to the steepness of the ground; that is, the rate of change of altitude with distance. Rate of change? That sounds like a vector! It is; it's called the **gradient vector** for the scalar field of altitude. Actually, the gradient vector is defined as the direction in which the scalar field *increases* — it's uphill, not downhill — so the motion of the boulder is in the opposite direction to the gradient vector. Similarly, the motion of air in Figure 4 is in the opposite direction to the gradient vector for the air pressure.

In a manifold without a metric, there's no way of classifying two directions as being “perpendicular”, so you can't construct gradient vectors from a scalar field. However, you can still ask how rapidly the contour lines are crossed when you're moving with a certain velocity. This is really the same thing as asking for the rate of change of the scalar field; you just have to take account of the interval between the contour lines and the direction in which you're crossing them.

For example, the train in Figure 4 is crossing the isobars in the direction of decreasing pressure, so if p is the pressure, $\partial_t p$ will be negative, and equal to –0.5 times the number of isobars the train is crossing per second (since the isobars are drawn at intervals of 0.5 kiloPascals). If the train was moving in the opposite direction, $\partial_t p$ would be positive, and if the track was at a tangent to the contour lines, $\partial_t p$ would be zero. But you can't single out any one direction in which $\partial_t p$ is greatest, because you can make $\partial_t p$ as large as you like just by travelling faster. In Euclidean space, you can pick the direction which, for a given speed, yields a faster rate of change of air pressure than any other direction. Without a metric, though, “a given speed” means nothing.

The geometrical object the contour lines themselves represent is known as a **1-form**. (There's a reason for this: the process we used to get contour lines from a scalar field can be generalised to make structures called **2-forms**, **3-forms**, and so on.) The 1-form generated from the scalar field f is called df, or “the differential of f”.

What can we do with a 1-form? As we've seen, we can combine it with a vector at any point to get a number. This number is known as the **inner product** of the 1-form and the vector. The inner product is written by enclosing the 1-form and the vector in angle brackets, e.g.:

$$\langle df, \mathbf{v} \rangle = \mathbf{v}(f) \qquad (4)$$

Equation (4) is just convenient shorthand for the things we've been discussing. The 1-

form df means “the contour lines of f”, and <df,**v**> means “the rate at which someone travelling with velocity **v** crosses the contour lines of f, times the interval at which those contours are drawn … times minus one if they're being crossed in descending order”. On the right hand side, **v**(f) is “the rate of change of f for someone travelling with velocity **v**.”

While keeping the picture of contour lines in mind, we can also give a slightly more mathematical definition of a 1-form: it's anything that we can combine with a vector to yield a number at any point in the manifold, such that the number we get is **linear** with respect to the vector. In other words, doubling the vector doubles the result, and a vector that's the sum of two others gives the sum of the individual results. Since the rate of change that a vector gives from a scalar field is linear in terms of the vector — see Equation (2) — the inner product defined by Equation (4) will also be linear:

$$\begin{aligned} \langle df, a\mathbf{v} + b\mathbf{w} \rangle &= (a\mathbf{v} + b\mathbf{w})(f) \\ &= a\mathbf{v}(f) + b\mathbf{w}(f) \\ &= a\langle df, \mathbf{v} \rangle + b\langle df, \mathbf{w} \rangle \end{aligned} \qquad \textbf{(5)}$$

There's no reason why you can't have a 1-form — call it **m** — that satisfies these equations in place of df, but doesn't happen to be the differential of *any* scalar field. We could draw a stack of lines at any point that yielded the right inner product for **m**, but they wouldn't intermesh from point-to-point like the true contour lines in Figure 4. In this case, it makes more sense to think of the lines being drawn in *the tangent space* for each point, since they don't really belong in the manifold. The inner product then counts how many lines of the 1-form a given vector pierces, from base to tip (with the usual corrections for interval and direction).

Just as we can add and subtract vectors, and multiply them by numeric factors, we can do the same with 1-forms, by declaring that the inner product will be linear, not just in the vector, but in the 1-form as well:

$$\langle a\mathbf{m} + b\mathbf{n}, \mathbf{v} \rangle = a\langle \mathbf{m}, \mathbf{v} \rangle + b\langle \mathbf{n}, \mathbf{v} \rangle \qquad \textbf{(6)}$$

And just as we can give a vector **v** coordinates v^x and v^y, we can give a 1-form **m** coordinates m_x and m_y. (Note that vector coordinates are written with superscripts, and 1-form coordinates are written with subscripts.) These are defined by:

$$m_x = \langle \mathbf{m}, \partial_x \rangle \qquad \textbf{(7a)}$$

$$m_y = \langle \mathbf{m}, \partial_y \rangle \qquad \textbf{(7b)}$$

which, for the special case of **m** = df, become:

$$(df)_x = \partial_x f \quad \textbf{(7c)}$$
$$(df)_y = \partial_y f \quad \textbf{(7d)}$$

Here, the units for the 1-form's coordinates are “scalar field units per coordinate unit” (e.g. kiloPascals per coordinate unit). In general, when **m** is not a differential, all we can say is that m_x and m_y are in “1-form units per coordinate unit”.

We can write the 1-form itself in terms of two **coordinate 1-forms**:

$$\mathbf{m} = m_x dx + m_y dy \quad \textbf{(8)}$$

where dx and dy are just the differentials of the manifold coordinates x and y, which are two perfectly good scalar fields after all. Note that:

$$\langle dx, \partial_x \rangle = 1 \quad \textbf{(9a)}$$
$$\langle dx, \partial_y \rangle = 0 \quad \textbf{(9b)}$$
$$\langle dy, \partial_x \rangle = 0 \quad \textbf{(9c)}$$
$$\langle dy, \partial_y \rangle = 1 \quad \textbf{(9d)}$$

which in turn lets us write:

$$\begin{aligned} \langle \mathbf{m}, \mathbf{v} \rangle &= \langle m_x dx + m_y dy,\ v^x \partial_x + v^y \partial_y \rangle \\ &= m_x v^x + m_y v^y \end{aligned} \quad \textbf{(10)}$$

If you multiply the units involved, the manifold's coordinate units cancel out, yielding “1-form units per parameter unit”. Unlike the individual coordinates v^x, v^y, m_x and m_y, this number is completely independent of the coordinate system. Why? Because it's almost just counting the number of intersections of one line with some other lines — and you can't get much more objective than that. The only proviso is that the 1-form and the vector's curve parameter do involve a particular choice of units; for example, in the case of our train crossing the air pressure contours, the inner product $\langle dp, \mathbf{v} \rangle$ would be in kiloPascals per second.

Curved Geometry

Now that we have a set of geometrical objects at our disposal that can function independently of any coordinate system — let alone any metric — the idea of doing geometry with a variable, curved metric becomes much less daunting. When Einstein was developing general relativity he postulated that the laws of physics should be

generally covariant, which means it should be possible to formulate those laws in a way that works equally well for *any* choice of coordinates. This is not to claim that there is no objective difference between, say, the way we normally measure coordinates around us in nearly flat spacetime (x, y, and z measured at right angles to each other, t measured in the direction of our world lines), and some arbitrary set of wiggly lines, resembling Figure 2. We can tell the difference very easily! But you can state the laws of physics, and solve the appropriate equations, without knowing *in advance* how to construct a coordinate grid with lines as straight as possible, meeting at right angles. The metric that defines these things is very real, but unless you have firsthand experience of the region of space in question, it's something you don't know until you've calculated it.

To study curvature, though, we're initially going to use examples where we don't have to solve equations in general relativity to find the metric. Rather, we'll look at curved surfaces in flat space, where the metric is "inherited" from the three-dimensional space in which the surface is **embedded**. Spacetime curvature doesn't arise like that — the metric is *not* inherited from some larger, flat "hyperspace" — but we can still define measures of curvature that are equally applicable in both cases.

Figure 5 shows the surface of the Earth, covered with a familiar coordinate grid: x and y are just longitude and latitude. As mentioned earlier, there are problems with this grid at the poles and the 180° meridian, but we'll stick to a region that avoids those mathematical trouble spots. We will make one slight change from the standard way of measuring latitude and longitude: we'll measure these angles in **radians**, rather than degrees. If you haven't come across this before, it's an extremely simple idea; 360° is equal to 2π radians — the circumference of a circle with a radius of one — and any smaller angle's equivalent in radians is just equal to the length of a proportionately smaller arc.

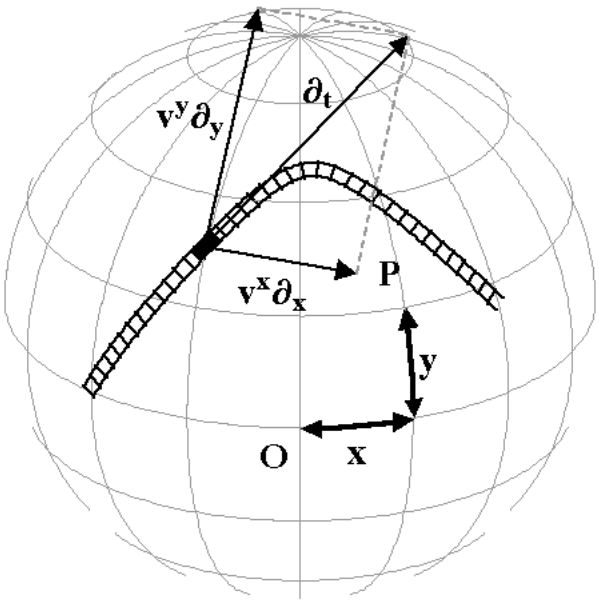

Fig 5: Points and Vectors on a Sphere

As before, imagine a train in motion across this surface. But now that the notion of "distance" is allowed, we can ask: what's the relationship between the train's speed, in kilometres per hour, and v^x and v^y, the rate at which its x- and y-coordinates are changing? If the train was heading purely in the y-direction, due north along a meridian, its speed would be equal to Ev^y, where E is the radius of the Earth. Why? Because v^y is just $\partial_t y$, the rate of change of the train's angle from the equator. Multiplying by the radius of the meridian the train is travelling on — which is just the radius of the Earth — converts an angle (measured in radians) into an arc length; equally, it converts a rate of change with time of that angle into the rate of change of the arc length.

If the train was heading purely in the x-direction, along a line of latitude, almost the same argument would apply. However, the radius of each circle of latitude obviously depends on the latitude, and it's not hard to see that the value is E cos y. So the train's speed would be $(E \cos y)\, v^x$.

Given that the train is actually heading partly in each of these two directions, we can use Pythagoras's Theorem to find its speed. This is where we make use of the plain old Euclidean metric of flat, three-dimensional space. The two directions, ∂_x and ∂_y, are at right angles to each other, and from a satellite's point of view they can be treated like vectors in a Euclidean plane. The speed of the train is:

$$|\mathbf{v}| \quad = \quad \sqrt{[(E \cos y)^2 (v^x)^2 + E^2 (v^y)^2]} \qquad \textbf{(11)}$$

What metric for the curved surface of the Earth will agree with this result? A metric applied to the same vector twice is supposed to yield the square of the length of that vector, $|\mathbf{v}|^2 = g(\mathbf{v},\mathbf{v})$, so:

$$g(\mathbf{v},\mathbf{w}) = (E \cos y)^2 v^x w^x + E^2 v^y w^y \qquad \mathbf{(12)}$$

is clearly compatible with Equation (11), since this gives:

$$g(\mathbf{v},\mathbf{v}) = (E \cos y)^2 (v^x)^2 + E^2 (v^y)^2$$

We can use Equation (12) to find the train's speed from the rates at which its longitude and latitude are changing, anywhere on the surface of the Earth (except for the poles and the 180° meridian). However, we *can't* substitute, say, the x- and y-coordinates of the point P in Figure 5 in place of v^x and v^y, and expect Equation (12) to give us the distance from O to P. Why not? Because the metric's relationship to the coordinates changes as you change latitude; that “cos y” in the formula means you can't just add up the length of all the steps you took as you walked from O to P, and expect them all to bear an identical relationship to the amounts by which they increased your latitude and longitude. In Euclidean space, if you head off along a straight line, equal strides always involve the same increments in your x- and y-coordinates. In general, this isn't true. What you *can* always do is use calculus to add up the distance along any path, taking account of the varying relationship of distance to the coordinates. That's not difficult, but we won't go into the details here.

Equation (12) can be rewritten in a way that makes it clear that the metric is a geometrical object, independent of the coordinate system. Equations (3a) and (9) allow us to rewrite the coordinates of the vectors in terms of inner products with the coordinate 1-forms: $v^x = \langle dx,\mathbf{v}\rangle$, $v^y = \langle dy,\mathbf{v}\rangle$, and similarly for the coordinates of **w**.

$$g(\mathbf{v},\mathbf{w}) = (E \cos y)^2 \langle dx,\mathbf{v}\rangle\langle dx,\mathbf{w}\rangle + E^2 \langle dy,\mathbf{v}\rangle\langle dy,\mathbf{w}\rangle$$

This in turn can be expressed with a more compact notation:

$$g = (E \cos y)^2 dx\otimes dx + E^2 dy\otimes dy \qquad \mathbf{(13)}$$

The symbol ⊗ is called the **tensor product**. Just as dx and dy can be combined with a single vector to yield a number, dx⊗dx and dy⊗dy can be combined with *pairs* of vectors to yield numbers. You can multiply as many 1-forms as you like in this fashion, and they need not be the coordinate 1-forms. For example, if **m**, **n** and **p** are 1-forms and **u**, **v** and **w** are vectors:

$$\mathbf{m}\otimes\mathbf{n}\otimes\mathbf{p}(\mathbf{u},\mathbf{v},\mathbf{w}) = \langle\mathbf{m},\mathbf{u}\rangle\langle\mathbf{n},\mathbf{v}\rangle\langle\mathbf{p},\mathbf{w}\rangle$$

You can also combine vectors in the same fashion, or even a mixture of vectors and 1-forms:

$$\mathbf{u}\otimes\mathbf{v}(\mathbf{m},\mathbf{n}) = \langle\mathbf{m},\mathbf{u}\rangle\langle\mathbf{n},\mathbf{v}\rangle$$
$$\mathbf{m}\otimes\mathbf{n}\otimes\mathbf{w}(\mathbf{u},\mathbf{v},\mathbf{p}) = \langle\mathbf{m},\mathbf{u}\rangle\langle\mathbf{n},\mathbf{v}\rangle\langle\mathbf{p},\mathbf{w}\rangle$$

The tensor product of *r* vectors and *s* 1-forms is called a **tensor** of rank (r,s). Since each vector can create a number if combined with a 1-form, and each 1-form can create a number if combined with a vector, you can "feed" *r* 1-forms and *s* vectors to such a tensor, and it will give you a number. Because this number is produced by multiplying a string of inner products together, and each inner product is linear in both the 1-form and the vector, tensors are completely linear: feed them anything twice as big, and the number they produce from it will be doubled.

Equation (13) shows that the metric for the surface of a sphere is a tensor of rank (0,2): it can be fed two vectors, and from them it produces a number. In the previous article, we showed that both the Euclidean and Minkowskian metrics are linear, so this aspect should come as no surprise. We also showed that those metrics were *symmetric*, i.e. $g(\mathbf{v},\mathbf{w})=g(\mathbf{w},\mathbf{v})$, which is also clearly true of Equation (13), since it's the sum of two tensors which combine identical 1-forms with the first and second vectors fed to them. The most general form a two-dimensional metric can take is:

$$g = g_{xx}dx\otimes dx + g_{xy}dx\otimes dy + g_{yx}dy\otimes dx + g_{yy}dy\otimes dy \quad \mathbf{(14)}$$

where each of the numbers g_{xx}, g_{xy}, g_{yx}, g_{yy} can vary from point to point in the manifold, but g_{xy} must equal g_{yx} everywhere, to ensure that the metric is symmetric. These numbers are called the **coordinates** of the metric, and like those of vectors and 1-forms they depend on the particular coordinate system being used. But Equation (14) as a whole is independent of the coordinate system. To take a simple example, if we switched from x and y to new coordinates u and v, where u=x/2 and v=y/2, we'd make the substitutions du=dx/2, dv=dy/2 (i.e. dx=2du, dy=2dv):

$$\begin{aligned} g &= 4g_{xx}du\otimes du + 4g_{xy}du\otimes dv + 4g_{yx}dv\otimes du + 4g_{yy}dv\otimes dv \\ &= g_{uu}du\otimes du + g_{uv}du\otimes dv + g_{vu}dv\otimes du + g_{vv}dv\otimes dv \end{aligned}$$

So the metric's new coordinate g_{uu} is equal to $4g_{xx}$, but that's balanced by the fact that du⊗du yields numbers one-quarter the size of those dx⊗dx yields from the same vectors. The same thing holds for g_{uv}, g_{vu}, and g_{vv}. The metric's coordinates have changed, but the thing itself is unaltered, just as a train travelling down a railway track is unaltered by the coordinates used to measure its velocity.

Parallel Transport

It's common knowledge that the “straightest possible” line — the technical name for this is a **geodesic** — between two points on the surface of a sphere is an arc of a great circle, a circle whose radius is equal to the radius of the sphere. If you travel along a great circle, there's a sense in which your velocity vector is always pointing in “the same” direction, as opposed to swerving from side to side. But what exactly does this mean? It's easy to say when two vectors at different points are parallel in Euclidean space — in a rectangular coordinate system the two vectors will have identical coordinates — but in curved space, the issue is a great deal more subtle.

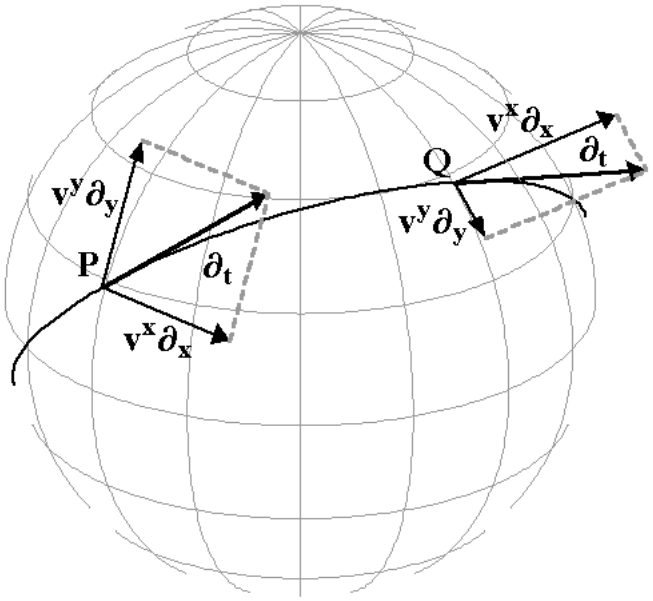

Fig 6: Tangents to a Geodesic

Figure 6 shows the velocity vector ∂_t at two different points for a ship travelling at a constant speed along a great circle. From a satellite's view, the velocity is clearly different from P to Q, but both vectors lie in the same plane: the plane of the great circle. As the ship moves around the Earth, it *can't* head off along a truly straight line; that would take it up into space! Instead, its velocity vector has to rotate downwards, in order to stay horizontal as the Earth curves away. This rotation is entirely perpendicular to the surface of the Earth; if it was partly sideways, the ship would swerve off course.

That's simple enough, but how can we characterise this from the ship's point of view? The coordinates of its velocity clearly don't stay constant: at P the ship is heading north-east, and at Q it's heading south-east. In general, v^x and v^y are going to change in a complicated manner, even as the ship maintains a constant speed along a geodesic.

The idea that we can take a vector at P and move it along some path to Q —

without “really” changing it, even though its coordinates might change — is known as **parallel transport**. It's this idea that defines a geodesic: parallel transport tells you how to carry a kind of “reference copy” of your initial velocity with you as you go. If your actual velocity agrees with the reference copy all the way, you're moving along a geodesic; if it doesn't, you're not.

Physically, you could do this by setting a gyroscope spinning with its axis in the direction of your initial velocity (and subtracting out the gradual tilt of the axis towards the vertical, as the local definition of “horizontal” changed), but what we need is a mathematical recipe to *predict* the difference between the gyroscope's direction and a fixed compass bearing. We'll assume that parallel transport is a linear process for the vectors being “transported”: adding two vectors or multiplying a vector by some number at the starting point, and then transporting the result, ends up giving you the same final vector as doing the transporting first and the adding or multiplying later. We'll also assume that over very short distances the difference between gyroscope and compass bearing is linear for the vector describing your motion; this is really just saying that the recipe for parallel transport, though it will vary from point to point (like the metric), doesn't undergo any wild jumps that would stop you from treating it as fixed over a small enough region. Then if we know what happens to ∂_x when we move a short distance ε in the x-direction, what happens to it if we move in the y-direction, *and* what happens to ∂_y after the same two moves, all these assumptions of linearity let us work out what happens to *any* vector, moved (a short distance) in *any* direction.

It's standard notation to write the effects of these four possible moves as:

$$\nabla_x \partial_x = \Gamma^x{}_{xx}\, \partial_x + \Gamma^y{}_{xx}\, \partial_y \qquad \textbf{(15a)}$$

$$\nabla_y \partial_x = \Gamma^x{}_{xy}\, \partial_x + \Gamma^y{}_{xy}\, \partial_y \qquad \textbf{(15b)}$$

$$\nabla_x \partial_y = \Gamma^x{}_{yx}\, \partial_x + \Gamma^y{}_{yx}\, \partial_y \qquad \textbf{(15c)}$$

$$\nabla_y \partial_y = \Gamma^x{}_{yy}\, \partial_x + \Gamma^y{}_{yy}\, \partial_y \qquad \textbf{(15d)}$$

Don't be put off by the unfamiliar symbols; as ever, this is just shorthand for things we've just discussed. After increasing our x-coordinate by a small amount ε, “the difference between the ∂_x the coordinate grid gives us and the parallel-transported reference copy of the ∂_x we started off with, divided by ε” is what we'll call $\nabla_x \partial_x$, with the subscript x in ∇_x indicating the direction of the move. We divide out the particular change in x, ε, because the effect is linear, and what we really care about is *the rate per coordinate unit*. This is a vector, with coordinates $\Gamma^x{}_{xx}$ and $\Gamma^y{}_{xx}$. Once you include the two coordinate vectors and the two directions you can consider moving in, there are eight of these numbers at every point, and together they characterise a particular way of doing parallel transport. They're known as the **connection coefficients** for the geometry, or sometimes the **Christoffel symbols**.

This gives us the language in which to talk about parallel transport, but we still don't know what recipe to use in any given case — what the values of the connection coefficients are, in terms of the metric. It turns out that there are two equivalent ways to pin this down. You can assume either that: (1) parallel transport must yield geodesics that give the shortest possible distance between nearby points (or in spacetime, the *longest* distance), or (2) parallel transport of vectors changes neither their length nor the angle between them, and what's more $\nabla_y\partial_x=\nabla_x\partial_y$: the x coordinate vector changes when you move in the y-direction in exactly the same way as the y coordinate vector changes when you move in the x-direction.

It's beyond the scope of this article to prove that these requirements are identical, but it's worth knowing both. Condition (1) is very simple to talk about, and makes great intuitive sense. In Euclidean space, straight lines are the shortest paths between points, so it's only reasonable that the “straightest” path between points in curved space should be the shortest. In Minkowskian spacetime, straight lines are the longest paths, so the same should apply to geodesics in curved spacetime. In fact, the fundamental reason for all of this lies in quantum mechanics, but that's a topic for a later article. Condition (2) is a little more complicated, but the first part sounds pretty reasonable: if the reference copy of our ship's velocity changed length as we moved, what sort of standard would it be? And if we carried a reference copy of “starboard” (to the right) that didn't continue to lie at 90° to the reference copy of our velocity, we wouldn't know which to steer by, since the two would contradict each other.

It's possible to use Condition (2) to write a completely general formula for the connection coefficients in terms of the metric, but we'll just state the results for the surface of a sphere:

$$\nabla_x\partial_x = (\sin y \cos y)\,\partial_y \qquad \textbf{(16a)}$$
$$\nabla_y\partial_x = (-\tan y)\,\partial_x \qquad \textbf{(16b)}$$
$$\nabla_x\partial_y = (-\tan y)\,\partial_x \qquad \textbf{(16c)}$$
$$\nabla_y\partial_y = 0 \qquad \textbf{(16d)}$$

Equation (16a) says that ∂_x (east) tilts in the y-direction (north) when you travel in the x-direction (east), to an extent that's zero either on the equator or on the poles (since either sin y or cos y is zero in those cases), and negative in the southern hemisphere. To see this, find the point on Figure 6 where the great circle is tangent to a circle of latitude, just west of point Q. The two curves are initially parallel, but then the circle of latitude — the definition of “east” — veers north relative to the great circle. Equation (16b) says that ∂_x decreases in the x-direction (i.e. shrinks) as you travel in the y-direction: as you travel north, a velocity measured in degrees or radians per hour east comes to mean less and less in terms of kilometres per hour, as the circles of latitude shrink. Equation (16c) says

that as you watch ∂_y while moving in the x-direction (i.e. keep an eye on what happens to "north", relative to your gyroscope, as you travel east), it tilts in the negative x-direction (west) by an amount that's greater at greater latitudes (tan y starts off equal to zero at y=0, and grows with latitude). That makes sense: the north pole lies *precisely* on your left when you start out east, but if you travelled along a great circle like the one in Figure 6, it would swing to the rear of you, which is west, as you turned to the south.

The Curvature Tensor

What happens when you parallel-transport a vector around a closed loop? Figure 7 shows this happening on the surface of the Earth: a vector **v** is transported along each of the three sides of a triangle PQR built from geodesic arcs. The angle between the vector and each geodesic it moves along remains unchanged, and to make the example even simpler, **v** is chosen initially to be a tangent vector to one side of the triangle, PQ, at P. After being transported from P to Q this is still true — the angle between the vector and the curve remains zero. Transport from Q to R is almost as easy: the vector just ends up making the same angle with QR at R and at Q … and then the same can be said about its angle with RP at R and at P.

At the end of the process, back at P, **v** no longer points in the same direction. This couldn't happen in Euclidean space, where the vector would remain parallel to its initial position throughout. In curved space, though, there is no absolute criterion for saying one vector is parallel to another vector at another point. All you can do is transport a copy of the first vector over to the second one, *by a particular route*, and see if they agree — and whether they do or don't will generally depend on the path you take.

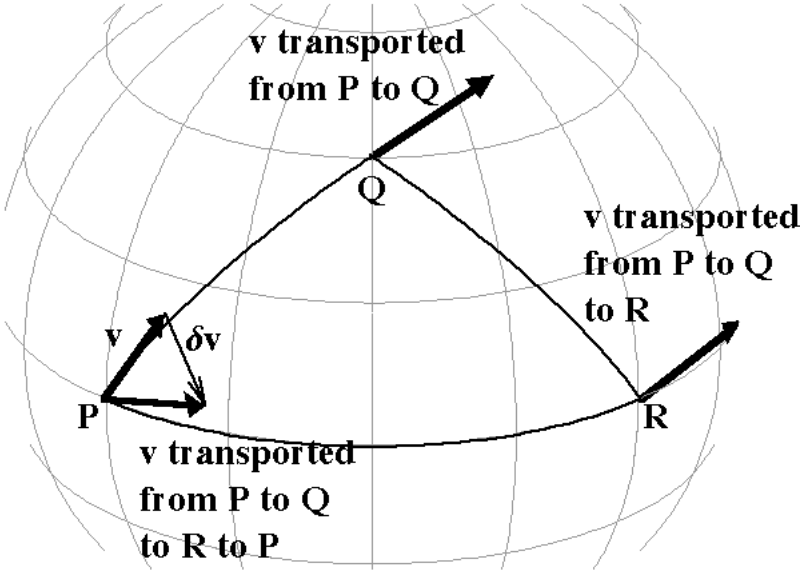

Fig 7: Parallel Transport Around a Loop

The failure of the vector to return to its initial direction can be linked to a well-known property of triangles in curved space: the failure of their angles to add up to 180°. The angle the vector **v** makes with successive sides of the triangle jumps at each vertex, by 180° minus the angle at the vertex; in total, this means it ends up rotated by 540° minus the sum of the angles. If that sum was 180°, the net rotation would be 360° and the transported vector would match the original. On a sphere, the angles of a triangle always add up to something *more* than 180°, so the transported vector in Figure 7 is rotated by less than 360° and fails to line up with the original.

It's clear that this discrepancy, $\delta\mathbf{v}$, between the original vector and the transported version must shrink as the triangle (or any other loop) gets smaller, because on a small enough region of the Earth's surface the geometry is indistinguishable from Euclidean geometry. However, we can still ask exactly how rapidly $\delta\mathbf{v}$ shrinks, and it turns out that for small enough loops it's proportional to the *area* of the loop, and unaffected by its shape. So we can create a kind of standardised small loop, a square which consists of travelling ε units in the x-direction, ε units in the y-direction, $-\varepsilon$ units in the x-direction, and $-\varepsilon$ units in the y-direction, where ε is a small number whose precise value doesn't matter. All that's needed to compute the discrepancy in any vector transported around this loop are the connection coefficients, and the rates at which they're changing in the x- and y-directions.

The result for the surface of a sphere turns out to be quite simple:

$$\delta\mathbf{v} \quad = \quad \varepsilon^2 \left(-v^y \, \partial_x + v^x (\cos y)^2 \, \partial_y\right) \qquad \textbf{(17)}$$

Note that this is *perpendicular* to the original vector:

$$\begin{aligned} g(\mathbf{v},\delta\mathbf{v}) &= \varepsilon^2 \left((E \cos y)^2 \, v^x(-v^y) + E^2 \, v^y v^x \, (\cos y)^2 \right) \\ &= 0 \end{aligned}$$

and its size relative to the original vector is:

$$\begin{aligned} |\delta\mathbf{v}| &= \varepsilon^2 \sqrt{[(E \cos y)^2 \, (-v^y)^2 + E^2 \, (\cos y)^4 \, (v^x)^2]} \\ |\mathbf{v}| &= \sqrt{[(E \cos y)^2 \, (v^x)^2 + E^2 \, (v^y)^2]} \\ |\delta\mathbf{v}| \, / \, |\mathbf{v}| &= \varepsilon^2 \cos y \end{aligned}$$

which is proportional to the area of the loop, once you take into account the fact that ε units of longitude comprise a shorter distance at higher latitudes, by a factor of cos y. For a given loop, δ**v** is perpendicular to **v** and proportional to it in length, so this confirms that the parallel-transported version of **v** has simply been *rotated* (if δ**v** wasn't perpendicular, that would imply some expansion or contraction as well).

Equation (17) can be rewritten as:

$$\mathbf{R} = \partial_x \otimes dy - (\cos y)^2 \, \partial_y \otimes dx \qquad \textbf{(18a)}$$

$$\delta\mathbf{v} = -\varepsilon^2 \, \mathbf{R}(\mathbf{v}) \qquad \textbf{(18b)}$$

where **R** is a tensor of rank (1,1). From our earlier description of tensors, a tensor of rank (1,1) would be fed a 1-form and a vector, to produce a number. But you can also leave the vector part “unfed” — combined with nothing — and use the tensor to produce one vector from another. In other words, Equations (18a) and (18b) are the same as Equation (17) because $(\partial_x \otimes dy)(\mathbf{v}) = v^y \partial_x$ and $(\partial_y \otimes dx)(\mathbf{v}) = v^x \partial_y$, with the 1-forms dy and dx combining with the vector **v**, and the vectors ∂_x and ∂_y left untouched as vectors.

The tensor **R** is *almost* the **Riemann curvature tensor** — a famous geometrical object containing all the information about the curvature of a given space or spacetime. There's one slight omission, though; we haven't considered the fact that in general, you need to specify *what plane* the loop lies in. There's only one possibility in two dimensions, but in three or four you have to specify a choice. The full Riemann tensor takes this into account: as well as feeding it the original vector **v**, you have to feed it two other vectors, say **w** and **z**, which together single out the plane of the loop. What's more, you can move the factor of ε into those vectors, so they specify the *size* of the loop as well.

We can modify Equations (18a) and (18b) to work like this:

$$\mathbf{R} = (\partial_x \otimes dy - (\cos y)^2 \, \partial_y \otimes dx) \otimes (dx \otimes dy - dy \otimes dx) \qquad \textbf{(19a)}$$

$$\delta\mathbf{v} = -\mathbf{R}(\mathbf{v}, \varepsilon \, \partial_x, \varepsilon \, \partial_y) \qquad \textbf{(19b)}$$

Here, the (dx⊗dy–dy⊗dx) part of **R** simply converts the last two vectors fed to **R** into the number ε^2. But it would cope just as well with whatever vectors you fed it; for example, if you reversed the roles of ∂_x and ∂_y and traversed the standard loop in the opposite direction, it would give $-\varepsilon^2$ instead.

In the next article, we'll see how the curvature of spacetime can be linked to the distribution of matter and energy, through Einstein's equation — and then we'll look at one solution of that equation in detail: a black hole.