Lesson 4 : Geodesics and gravity

Notes from Prof. Susskind video lectures publicly available on YouTube

Introduction

Before we enter into the heart of the subject – geodesics and gravity – let's go through a quick review of some of the important points and equations we have established so far.

We saw various operations we can do on vectors and tensors : addition, tensor product, contraction, not to mention expression of their components in various coordinate systems.

Most importantly, we talked about the operation of *differentiation.* We have a surface or variety, with a tensor defined at every position. How to calculate the variation of the tensor when we move a bit from X to $X + dX$? If we did it "naively" we would produce something which would depend not only on the tensor itself, but also on the change of the curvilinear coordinate system from X to $X + dX$ [1]. And it would not be a tensor. It could not be used conveniently in frame-independent equations.

We defined therefore a special kind of differentiation called *covariant differentiation.* At any given point P on the surface, we change from the current coordinate system to a set of local Gaussian normal coordinates – that is coordinates which are closest possible to Cartesian coordinates. This can be done in several ways, because once we have one set of Gaussian normal coordinates we can rotate it and that produces another set. But considering how we use it it does not matter which set we use.

So we choose one set of Gaussian normal coordinates at P, and we calculate the *ordinary derivatives* of the compo-

[1] Notice that the idea of a change *not due to* the tensor itself is more delicate that it seems. It requires to be defined, and that is precisely the purpose of considering locally what is going on in Gaussian normal coordinates.

nents of the tensor with respect to each direction *in that system.* This produces a multi-indexed collection of components with one more index downstairs. And we treat this collection as the components of a tensor expressed in that Gaussian normal coordinate system.

We can then express the derivative tensor back into the initial curvilinear coordinates if need be. But remember : a tensor exists and is well defined irrespective of the basis we are using – just like vectors exist and are well defined in a vector space irrespective of any basis. Bases are used to work on vectors when we need to have them in component form. The tensor obtained by the above differentiation process is called the *covariant derivative* of the tensor we started with.

At any particular point P, if we happen to be in Gaussian normal coordinates, then the covariant derivative is just the ordinary derivative. But in other coordinates, it takes on a more complicated form. This is a consequence of the fact that we want it to be a tensor.

For the simplest example of a tensor, that is a covariant vector, the formula is

$$D_r V_m = \partial_r V_m - \Gamma^t_{rm} V_t \tag{1}$$

where Γ^t_{rm} is called a Christoffel symbol.

For a tensor with more covariant indices, the formula is a simple generalization of equation (1), carrying an extra term with a Christoffel symbol for each index. For instance

$$D_r T_{mn} = \partial_r T_{mn} - \Gamma^t_{rm} T_{tn} - \Gamma^t_{rn} T_{mt} \tag{2}$$

Equations (1) and (2) are valid in any coordinate system.

But let's stress again that, at any given point, if we are locally using a coordinate system which is as close as possible to Cartesian – instead of Gaussian normal we will sometimes call it a "best local coordinate system" – then the Christoffel symbols are zero, and the right hand sides reduce to their first terms, that is to ordinary derivatives.

Having recalled all this, let's turn to a specific tensor : the metric tensor. Cartesian coordinates are by definition a coordinate system in which the metric is everywhere constant, moreover is equal to the Kronecker delta tensor. And a space in which such a system can be found is called flat.

Similarly, locally, a Gaussian normal coordinate system is one in which the metric tensor is locally the Kronecker delta tensor up to second order (that is, still behaving like the Kronecker tensor in the fist order but not in the second). Therefore, in a set of Gaussian normal coordinates, at a given point P, the ordinary partial derivatives of the components of the metric tensor are zero :

$$\partial_r \, g_{mn} = 0 \tag{3}$$

This is true *only in a set of Gaussian normal coordinates.*

However, as a consequence, the covariant derivative of the metric tensor, which is itself a tensor, *in any coordinate system*, at any point P on the surface, is equal to zero :

$$D_r \, g_{mn} = 0 \tag{4}$$

Looking again at the Christoffel symbols appearing in equations (1) and (2), we saw in many ways why they are not tensors. Unlike tensors, they can be zero in one coordinate system and not zero in another. We calculated their value, in any given coordinate system, in terms of the ordinary partial derivatives of the components of the metric tensor :

$$\Gamma^t_{mn} = \frac{1}{2}\, g^{rt}\ [\ \partial_n g_{rm} + \partial_m g_{nr} - \partial_r g_{mn}\ ] \qquad (5)$$

We see again that if the ordinary partial derivatives of the metric tensor components are zero, as is the case in a best local coordinate system, then the Christoffel symbols are zero in that coordinate system. If they were tensors they would have to be zero in any coordinate system, but they are not.

Equation (5) should be memorized. It is not too hard. Remember that tensor indices must display a coherent structure, and follow the summation convention. On the left hand side, we have a Christoffel symbol. It has two lower indices, m and n for instance, and one upper index, t for instance. Its expression will involve a sum – expressed not with a big Σ but with the summation convention – over an index, say, r. The formula begins with a factor 1/2 followed by the inverse metric tensor g^{rt} with upper indices. These are the t of the Christoffel symbol and an r to operate a sum. Then, between brackets, the formula contains three partial derivatives of the form $\partial_i g_{jk}$, where ijk cycle over mnr, two terms with a plus sign and the term with g_{mn} with a minus sign[2]. That is the Christoffel symbol, with one upper index and two lower indices.

These Christoffel symbols go into equations (1) and (2). That tends to make covariant derivative rather complica-

[2] Equation (5) looks slightly different from equation (19) of chapter 3, because between the brackets we wrote, in the middle term, g_{nr} instead of g_{rn}. But remember that g is symmetric.

ted objects. If we plugged the expression of the Christoffel symbols in equations (1) or (2), and even wrote out the analytic form of the metric tensor that we might have at our disposal, it would be a nuisance. But that is what covariant derivative means.

Equation (1) was the covariant derivative of a vector with covariant components. Let's talk now about the covariant derivative of a vector with *contravariant* components. We denote it

$$D_r V^m$$

As always it starts out with an ordinary partial derivative, and there is another term. The calculations are exactly the same as what we did to calculate the covariant derivative of a covariant vector. They are left to the reader. But he or she is invited to use the following little trick. We can write

$$V^m = g^{mp} V_p$$

Then we take the covariant derivative of each side. Since, in a best set of coordinates, the covariant derivative is a standard derivative, it is easy to see that it will satisfy the rule of differentiation of a product (see chapter 2 of Volume 1 of the collection *The Theoretical Minimum*)

$$D_r V^m = (D_r g^{mp}) V_p + g^{mp} (D_r V_p) \tag{6}$$

Now what do we know about the first term on the right hand side ? It involves the covariant derivative of the inverse metric. We know that the covariant derivative of the metric is zero – that was equation (4). That means, as we saw, that in a best set of coordinates the metric is constant up to second order. Then its inverse must also be constant up to second order – go into matrix representation if necessary to convince yourself. Therefore the first term on the right hand side of equation (6) disappears, and the equation rewrites

$$D_r V^m = g^{mp}(D_r V_p) \tag{7}$$

Now we know how to calculate the covariant derivative of a vector with lower indices. That is given by equation (1). If you plug that in equation (7) and do a little algebraic manipulation, you will find out the formula for covariantly differentiating a vector with a contravariant index, that is with an upper index.

Here is the result

$$D_r V^m = \partial_r V^m + \Gamma^m_{rt} V^t \tag{8}$$

As before the formula begins with just a simple derivative. Then it has a term which would be zero in a set of best coordinates, because the covariant derivatives would simply be the ordinary ones. But they are not zero in general coordinates. In the term with the Christoffel symbol, there is a sum over t. Generally speaking, in equation (8), we see that all the indices are in place as expected.

The only peculiarity is the plus sign instead of the minus sign that appeared in the covariant derivative of a vector with a lower index. That minus sign was a convention. Here too, but it must be the other sign.

If we have memorized the covariant derivative of a covariant vector, to remember the formula for the derivative of a contravariant derivative, it is the only thing we have to remember. The place of the indices follow consistent rules. There is no way to be mistaken. The Christoffel symbols have one upper index and two lower indices. The upper index must the same as on the left hand side of equation (8). The second term on the right hand side is a sum, so there must be a dummy index t downstairs in the Christoffel symbol, and a corresponding t upstairs on V^t.

Just as we generalized the covariant derivative of a covariant vector to tensors with covariant indices, going from equation (1) to equation (2), we can now generalize the covariant derivative to a tensor with any collection of lower and upper indices. A lower index will entail an extra term with a Christoffel symbol with a minus sign, an upper index will entail an extra term with a Christoffel symbol with a plus sign.

Now we come to the idea of parallel transport. We have already touched upon it in the previous lesson. But let's now spell it out in detail.

Parallel transport

Suppose we have a curved surface – or a higher dimensional curved space – and some vector field defined on it. That is, at every point of our space, there is attached a vector. And, in what follows, the vectors of the vector field will always be in the tangent plane – or in the higher tangent flat space – to the space.

We are interested in knowing, when we move along a curve on the space, see figure 1, whether the field stays parallel to itself. In figure 1 we have represented the space and the curve, but neither the vectors of the vector field nor the curvilinear coordinates on the surface.

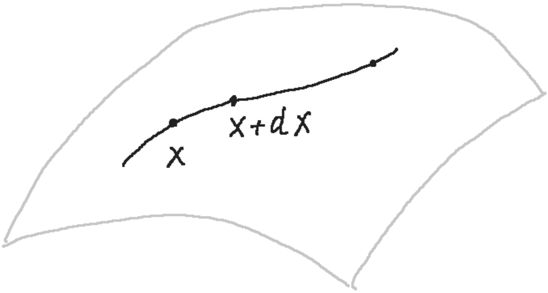

Figure 1 : Vector field and curve on a variety.

At each point of the curve, imagine there is a vector. Let's move along the curve. What we want to know is whether the field stays parallel to itself. "Parallel to itself" between X and $X + dX$ on the curve means that the covariant derivative of the vector is 0 in the direction of the curve at that point X.

The covariant derivative is the difference between the vectors at $X + dX$ and at X, as they are written in best local coordinates, divided by the components of dX. Let's write again the tensor which is the covariant derivative of a contravariant vector

$$D_m V^n = \frac{\partial V^n}{\partial X^m} + \Gamma^n_{mr} V^r \tag{9}$$

Now we want to consider the derivative along the trajectory or curve. How does the vector change from point to point ?

That simply corresponds to taking the covariant derivative $D_m V^n$ and multiplying it by dX^m. Hence, the small change in the vector is

$$D_m V^n dX^m \tag{10}$$

This formula eliminates the fact that the coordinates themselves may evolve as we go from point to point. That is the essence of covariant derivative.

Expression (10) is the small change in the vector V in going from one point to its neighbor, measured by the change of its components in a set of best coordinates, and then considered abstractly in any coordinate system. Let's give it a name

$$DV^n = D_m V^n dX^m \tag{11}$$

It is the *covariant change* in the vector going from one point to a neighboring point on the trajectory.

Let's express this covariant change with the building blocks we have. We multiply the right hand side of equation (9) by dX^m and get

$$DV^n = \frac{\partial V^n}{\partial X^m} dX^m + \Gamma^n_{mr} V^r dX^m \tag{12}$$

The first term on the right hand side has a simple interpretation. It is the ordinary differential change in V disregarding anything related to a possible change in coordinates. We denote it dV^n. Equation (12) becomes

$$\boxed{DV^n = dV^n + \Gamma^n_{mr} V^r dX^m \qquad (13)}$$

It reads as follows : the covariant change in V is equal to the ordinary change in V plus a term equal to a Christoffel symbol multiplied by V^r and by dX^m. This second term is of course a double sum according to summation convention.

Equation (13) is the formula which tells you how a vector changes from point to point.

Now suppose we are interested in finding a vector which is parallel to itself as we move along the curve. Parallel to itself means that it doesn't change as we move from X to $X + dX$. At each point X, we erect some best coordinates, and in those coordinates we test whether the vector is changing. If it doesn't change in the first order – that is, its first derivative is zero –, we say : good, the vector is constant along the little segment. We go to the next little segment, erect best coordinates at the new point, test again. We do that along the whole curve. It the tests say that the vector never changes in the first order, the vector is said to be parallel to itself along the curve[3].

In summary, if all along the curve the vector V satisfies

$$dV^n + \Gamma^n_{mr} V^r dX^m = 0 \tag{14}$$

then the vector maintains a relationship of being parallel to itself.

[3] There is a notion of calculus to remember : consider a function f continuous and differentiable over a segment $(a,\ b)$. Suppose $f(a) = 0$, and when we go from x to $x + \epsilon$, $f(x+\epsilon) - f(x) < o(\epsilon)$. The notation $o(\epsilon)$ means something an order of magnitude smaller than ϵ. That is, when ϵ goes to zero, $o(\epsilon)/\epsilon$ goes to zero too. Then the function f is equal to zero along the whole segment $(a,\ b)$. In other words, if over any small displacement the variation of f is an order of magnitude smaller than the displacement, f' is always zero, and f stays constant.

Taking a vector from one point and transporting it like this along a given curve, in such a way that it stays parallel to itself, is called *parallel transport*. Making up a benign neologism, we say that we "parallel-transport" the vector.

An important point about parallel transport on a curved space is that it is *trajectory-dependent*. On the surface in figure 2, if we start at point A, take a vector V there, which lives in the tangent plane, and parallel-transport it to point B, the vector we end up with at B will depend on the path we followed from A to B :

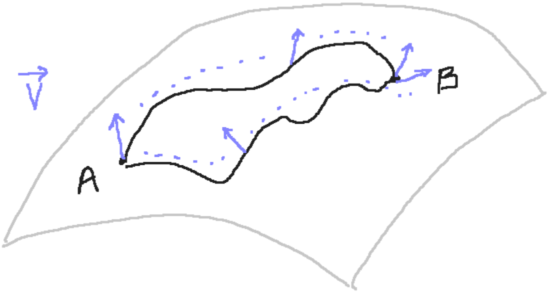

Figure 2 : Parallel-transporting V from A to B.

In figure 2, we represented the vector V (or $\vec{V}$) at A and suggested its evolution along two paths. We did not represent any coordinate system. Indeed, it is important to understand that parallel transport is dependent on the trajectory, but is independent of any coordinate system used to locate points on the surface. At each point, anyway, we use a set of best local coordinates to do the infinitesimal parallel transport of the vector there. When we arrive at B, the final vector we end up with depends not only on V or course, but also on the path we followed. The final vector depends on the bumps and troughs we encountered along the path, that is on the local curvatures along the path. Even if we came back to the same point A, depending on the loop we followed, we would end up with one or another vector. If there exists a flat *connected* region – that is with no hole – and we follow a loop entirely in that region we will end up with the same vector V.

We already saw this phenomenon on the cone – pointy or rounded, it doesn't matter – in the previous lesson. When we started with a vector on the side of the cone, and parallel-transported it *around the cone* we did not ended up with the same vector. An alternative path would be not to go around the top of the cone, in which case we would end up with the same vector – if we were careful to stay on the flat part. So we see that two paths don't lead to same result, figure 3.

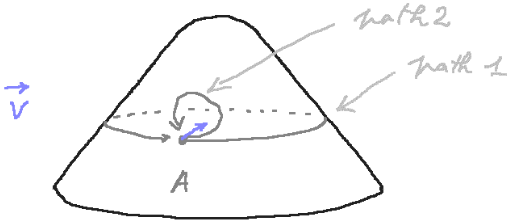

Figure 3 : Parallel-transport on a cone along two different paths.

Remember that the side of a cone is flat according to our definition, even though we see it embedded in 3D and in ordinary language it is not flat. The side of a cone is *intrinsically* flat, because any section of it with no hole can be laid out on a plane without exerting any distorsion on it. More mathematically, any connected section of the side is flat because there exists a coordinate system the metric of which is the Kronecker delta tensor over the whole section.

Parallel-transporting a vector, that is moving it on the surface while making sure its covariant derivative remains null, also preserves its length. It can be shown as a consequence of equation (14).

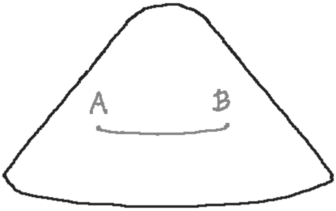

Figure 4 : Going from A to B on a cone.

The next topic will concern tangent vectors to a curve, and whether the tangent vector stays constant or not. When the tangent vector stays parallel to itself we will see that the curve is a geodesic. For instance on the cone of figure 3, if we go from A to A around the cone, we don't follow a geodesic because obviously the tangent vector changes direction. Even when we go from A to B as in figure 4, this is not a geodesic.

On the Earth boats and airplanes try to follow geodesics for reasons we shall see. When, while we are sitting on a plane, going from Paris to New York, the crew shows on a screen our trajectory, we are often surprised to discover that we do not follow "a straight line". The quotes are because the notion of straight line on the Earth unless we are in a small region is, as we know, something to handle with care.

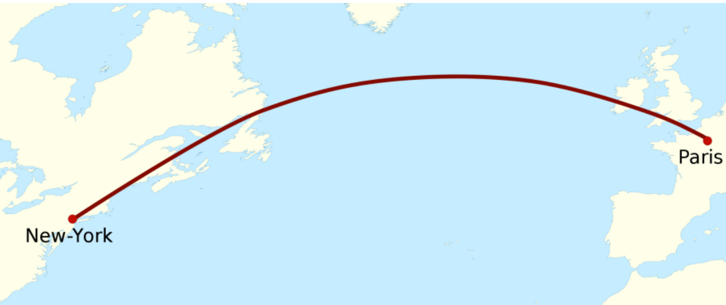

Figure 5 : On a plane from Paris to New York.

Tangent vectors and geodesics

We arrive at the notion of tangent vector to a curve and of geodesic.

On a surface where we consider two points A and B, a geodesic between A and B is a curve with certain properties. One can define a geodesic in several ways :

1. The curve with the shortest distance between A and B is a geodesic.
2. A curve whose length is stationary when you wiggle it is a geodesic.
3. A third better definition looks at what happens locally along the curve. A curve which at each point is as straight as possible is a geodesic.

Of course this last definition is more intuitive than mathematical. So let's make it more precise. If at each point along the curve, the covariant derivative of the tangent vector [4] is zero, that is if the tangent vector doesn't change, then the curve is as straight as possible.

Let's try to build more intuition about it before turning to the maths. Suppose you have a curved terrain, as in figure 6. For convenience, it is a two dimensional example, but there is nothing special about two-dimensional spaces in defining the notion of geodesic. Secondly suppose we have a car that we drive on this terrain. And assume that the size of a car, in particular the distance between the front wheels, is small by comparison with any curvature. In other words, the car is very small compared to the hills and the valleys.

We start from A in some direction, and driving straight in the above sense – never turning the steering wheel –, we end up at B. Our trajectory will wind between the hills – or we may even start from the top of a hill, that is, from a point with clear curvature. The curve that we will execute with our car in the space, keeping the steering wheel straight, will

[4] When we talk about the tangent vector without further specification, it is of length one.

nevertheless be as straight as possible. It will be a geodesic in the space.

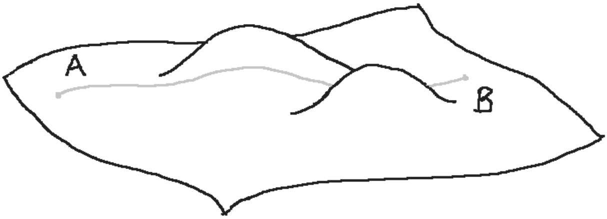

Figure 6 : Driving a car straight ahead, on a curved terrain.

Another way to characterize a geodesic is to say that the tangent vector along the curve is constant. We have an intuitive perception of what the tangent vector is. But let's define it more precisely.

Consider a curve, and a point at coordinate X on it. And take a neighboring point, figure 7. The points X and $X+dX$ are separated by dX which we can also denote, in tensor style, dX^m. And consider a vector starting a X, going through $X + dX$, and of length one.

Then take the limit when the second point $X + dX$ approaches the first point X. The resulting vector is called the tangent vector at X.

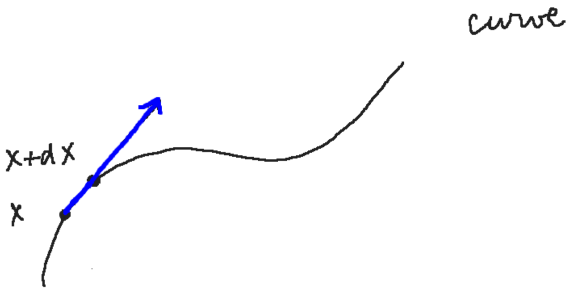

Figure 7 : Construction of the tangent vector at a point.

Consider the distance dS between the two points X and $X + dX$, as you remember, it is defined by

$$dS^2 = g_{mn}\, dX^m dX^n \tag{15}$$

The way we construct the tangent vector in the X coordinate system is very simple. The m-th component of the tangent vector is

$$t^m = \frac{dX^m}{dS} \tag{16}$$

It can be proved that equation (16) produces a vector of length one. The exercise is left to the reader. There is one such vector at each point along the curve. It is called the *tangent vector*. It points in the direction between two neighboring points and its length is one.

Let's now turn our attention to curves the tangent vector of which is constant ? If we plug-in the tangent vector in equation (14), these curves satisfy the following equation :

$$\boxed{dt^n + \Gamma^n_{mr} t^r dX^m = 0 \qquad (17)}$$

This equation holds because once you have set your steering wheel dead ahead, you are moving in as straight a line as you can. So the covariant change of the tangent vector is zero. See the example below to build your intuition.

Example of calculations with Christoffel symbols

In order to build our intuition about geodesics, particularly where a surface has curvature – like a rounded hill –, let's see a complete example with calculations.

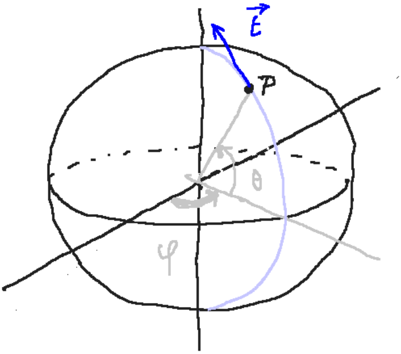

Figure 8 : 2-sphere with polar coordinates.

Consider a point P on the surface of a sphere. Mathematicians call such a surface a 2-sphere, because its points are located with two coordinates. Let's use the ordinary latitude θ and longitude ϕ, and the ordinary distance we are familiar with, for instance on the Earth.

The objective of the exercise is to show that a meridian is a geodesic. In other words, when we follow a meridian the tangent vector *doesn't change.*

Exercise 1 : We are on a 2-sphere with polar coordinates θ and ϕ, see figure 8.

1. Show that the metric tensor of the ordinary distance is

$$\begin{pmatrix} 1 & 0 \\ 0 & cos^2\theta \end{pmatrix}$$

2. Express the eight Christoffel symbols using this metric. Show that

$$\Gamma^1_{22} = sin\theta \; cos\theta$$
$$\Gamma^2_{12} = \Gamma^2_{21} = -tan\theta$$

and all the others are zero.

3. Show that the tangent vector to a meridian has everywhere components $t^1 = 1$ and $t^2 = 0$.
4. Show that the tensor which is the covariant derivative of this tangent vector is

$$\begin{pmatrix} 0 & 0 \\ 0 & -tan\theta \end{pmatrix}$$

5. Show that if we follow a meridian, the covariant change of the tangent vector is always zero.

Doing this exercise will show you that doing actual calculations with Christoffel symbols, even on a simple example, quickly fills pages. It will also show you that even on a surface with curvature there are paths where the tangent vector doesn't change. These are the geodesics.

In the exercise we looked at a meridian, because the polar coordinates make it simple to study, but of course by symmetry any great cicle is a geodesic.

We might feel, in figure 8, that the tangent vector changes when we move along a meridian, but it is because we look at the 2-sphere *embedded* in 3D Euclidean space.

If we turned our steering wheel, however, and in the tangent plane swerved from our straight path, then the tangent vector of our trajectory would change.

More on geodesics

We can write equation (17) of a geodesic in a slightly neater form. Let's divide both sides of the equation by dS, that is, by the little distance between two neighboring points with coordinates X and $X + dX$, see figure 7.

Equation (17) becomes

$$\frac{dt^n}{dS} = -\Gamma^n_{mr} \; t^r \; \frac{dX^m}{dS}$$

But $\frac{dX^m}{dS}$ is t^m, so we can rewrite equation (17) as

$$\frac{dt^n}{dS} = -\Gamma^n_{mr} \; t^r \; t^m \tag{18}$$

This equation only involves the tangent vector. It also of course involves the Christoffel symbols, but let's suppose we are given them. Then equation (18) is the equation of motion on a geodesic.

The gammas are made up out of the metric. So if we know the metric, we know what to put on the right hand side. And we can say one more thing : since the tangent vector t itself is a derivative, we can write the left hand side as a second derivative.

$$\boxed{\frac{d^2X^n}{dS^2} = -\Gamma^n_{mr}\ t^r\ t^m \qquad (19)}$$

Does this look like anything familiar ? May be not. But if we were to think of S as some measure of time as we moved along the curve, then the second derivative of position on the left would be acceleration.

So if S were like time, or if S were increasing uniformly with time, then we would have the following fact : an acceleration is equal to something that depends on the metric and on t. We will deal with $t^r\ t^m$ later.

Let's observe for the time being, that equation (19) has the look of a Newton equation : acceleration is equal to something that depends on the gravitational field, because, as we will see, the metric *is* the gravitational field.

We will see that equation (19) replaces Newton's equation for the motion of a particle in a gravitational field.

In other words, in some sense a particle in a gravitational field moves along the straightest possible trajectory. But it moves along the straightest possible trajectory *through space-time* not just through space.

Space-time

So now we have to come to space-time. So far we have been studying the mathematics of curved spaces as Riemann would have understood them.

The German mathematician Bernhard Riemann (1826-1866) is the one who invented most of the mathematics for curved spaces. But these were, so to speak, ordinary curved spaces in which distance was governed by the Pythagorean theorem. In Riemann spaces, the square of the distance is always positive.

A Minkowski space is a *space-time* which also has a natural measure of distance along curves or between points – also called *events* in the space-time –, be they neighboring points or distant points. As in Riemannian geometry, in Minkowski geometry the distance is generally expressed through its square. But in Minkowski space this square can be negative.

What is the name of the distance in a Minkowski space-time ? Answer : the *proper time.* It applies to any pair of events, far away, close, or infinitesimally close.

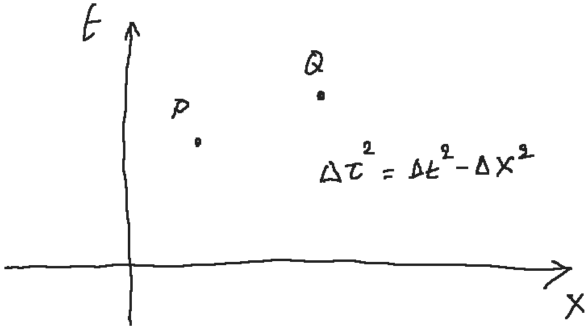

Figure 9 : Proper time between two events P and Q.

Given two points P and Q in space-time, the space-time distance between them – and now let's not call it the space-time distance, let's call it the proper time – is equal to Δt^2 not plus $\Delta X2$, but *minus* ΔX^2.

$$\Delta\tau^2 = \Delta t^2 - \Delta X^2$$

When P and Q are infinitesimally close, it becomes

$$d\tau^2 = dt^2 - dX^2 \tag{20}$$

And X may stand for three spatial coordinates $(x,\ y,\ z)$. With a Greek index it may even stands for the four coordinates, see X^μ below.

It is conventional to rewrite equation (20) as

$$dt^2 - dX^2 = -dS^2$$

where S is called the proper distance. Some authors prefer to work with τ, some prefer to work with S. In this book, we will mostly use the proper time τ.

Another convention it to write equation (20) as

$$dS^2 = g_{\mu\nu}\, dX^\mu\, dX^\nu \tag{21a}$$

or

$$d\tau^2 = -g_{\mu\nu}\, dX^\mu\, dX^\nu \tag{21b}$$

where the readers who have read volume 3 of *The Theoretical Minimum*, on special relativity, are familiar with the notation X^μ with a Greek index :

$$X^\mu = \begin{pmatrix} t \\ x \\ y \\ z \end{pmatrix}$$

According to standard convention this is also sometimes noted

$$X^\mu = \begin{pmatrix} X^0 \\ X^1 \\ X^2 \\ X^3 \end{pmatrix}$$

where the Greek index μ runs over 0 to 3.

When we use a Latin index, we mean only the three spatial coordinates, that is, if you read X^i, this means that i runs over 1, 2 and 3, or in other words X^i runs only over the spatial coordinates.

Let's comment on equation (21a). The indices μ and ν run over 0 to 3. The equation has exactly the same form as the usual equation for the distance in Riemannian geometry that we have already often used, see equation (1) of chapter 3 for instance.

The only new thing in the Minkowski geometry is the metric tensor $g_{\mu\nu}$ or the corresponding matrix. It is still diagonal, but it has a minus 1 corresponding to the time axis, and three $+1$ for the space axes. Since it plays a central role, it has a name. We use the Greek letter η (written and pronounced "eta"). It is called $\eta_{\mu\nu}$ (pronounced "eta mu nu").

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{22}$$

With this metric tensor, we can check that equation (21b) expressing the proper time is the same equation as

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

What do all the zeros mean in the metric tensor ? They mean that there are no cross terms like $dt\ dx$, or $dy\ dz$ in the definition of the proper time.

The coordinates $(t,\ x,\ y,\ z)$ are coordinates specific to Minkowski geometry. They are the analog of Cartesian coordinates in Euclidean geometry.

In special relativity, that is all there is to the coordinates and the metric tensor.

In general relativity, the metric tensor becomes a function of space and time. We then call it $g_{\mu\nu}(X)$ (where X clearly stand for an event with four coordinates). And equation (21b) becomes

$$d\tau^2 = -g_{\mu\nu}(X)\ dX^\mu\ dX^\nu \tag{23}$$

There is one more important thing, to mention at the outset, on the metric tensor in relativity. What is the difference between the matrix of equation (22) and the identity matrix of Euclidean metric ? Well, it has a minus 1 in first position. But more importantly there is an invariant concept about $g_{\mu\nu}(X)$: it has one negative eigenvalue and three positive eigenvalues. And this will always be the case in general relativity.

We are not going to spend much time dealing with the mathematical fact. The equations will always automatically take care of all of that. But what does it mean that there is one negative eigenvalue and three positive ? It means there is one dimension of time and three dimensions of space.

We could write a metric with two minus signs on the diagonal. It would correspond to a crazy space with two time directions and two space directions. That is not allowed in relativity.

At the beginning of his work on general relativity, Einstein realized that the metric of space-time should always have one negative and three positive eigenvalues – locally at any point in space-time. But, as said, we won't have to worry about it. It will be taken care of by the equations themselves.

Other than that, all that we have done in Riemannian geometry, all the equations involving metrics, covariant derivatives, curvature, geodesics, etc. will be exactly the same in the space-time geometry of general relativity.

What does flat mean in space-time ?

Flat no longer means that there is a coordinate system in which the metric is the Kronecker delta. Now it means that there is a coordinate system in which the metric has the form of $\eta_{\mu\nu}$, that is, the Kronecker delta except with one negative sign.

Remember that in Riemannian geometry, the condition for flatness was not that the metric we were dealing with was the unit matrix, but that there existed coordinates – other coordinates that we could find – in which it would be the unit matrix.

Similarly, in Minkowski geometry, the condition for flatness in space-time is not that the metric be $\eta_{\mu\nu}$, it is that *there exist coordinates in which* the metric can be brought to the very simple form of $\eta_{\mu\nu}$. Mathematically speaking this condition is equivalent to saying that the metric must have one negative eigenvalue and three positive eigenvalues.

That is the notion of a flat space-time. And if the space-time cannot be brought to the form of $\eta_{\mu\nu}$, then it is curved.

How do we check whether the space-time is curved ? We do all exactly the same things that we did in Riemannian geometry.

Here is a recap of the analogies that we have already made so far, as well as those we shall see :

Flat spaces

Euclidean geometry → Minkowski geometry

Kronecker δ tensor → η tensor

Newton physics → special relativity

Non flat spaces (always locally flat)

Curved metric → gravitational field

Riemannian geometry → Einstein general relativity

Before going into spaces whose curvature is due to real gravitational fields (i.e. to the presence of massive bodies), we shall spend some time with "flat" spaces with Minkowski geometry. Standard terminology, however, doesn't call them "flat", it simply calls them "spaces endowed with the Minkowski metric".

So for a while we are going to do special relativity. We will deal with a space with the $\eta_{\mu\nu}$ metric.

And we are going to wind up looking at it in polar coordinates – not ordinary polar coordinates but hyperbolic polar coordinates. The name is awe-inspiring, but the concept is simple and well adapted to space-time and particles moving in it, particularly particles accelerating in it.

Since we know from chapter one that there is a link between gravity and acceleration, and our ultimate objective is to describe relativistic motion of particles in gravitational fields, it is natural to start with studying particles accelerating in the framework of special relativity.

Special relativity

We are in the space-time of special relativity, which we can also call a Minkowski space. Its metric is defined by the tensor

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{24}$$

Our goal is to define the notion of a uniformly accelerated reference frame.

We have already met it in chapter one when we illustrated the principle of equivalence with a the gravitational field of the Earth (in a very small region where it can be viewed as uniform) and the apparent field we experience in an elevator being uniformly accelerated. But in chapter one, we staid at an elementary somewhat casual level.

In fact, in special relativity there is a difficulty with the notion of uniformly accelerated reference frame. Consider a bunch of points, separated by a fixed distance, as in figure 10. We think of them as forming a frame.

Figure 10 : Points in space with a fixed separation.
We try to move them "uniformly".

Suppose we start accelerating them along the X-axis, and let each have the same constant acceleration. We would think that they would maintain the same distance between them – that they would keep forming a frame. But distances, time and simultaneity do odd things when we start moving sizeable objects.

In fact if we gave all the points the same trajectory, in particular the same acceleration, we would discover that, in the rest frame of the first point for instance, they would neither stay fixed, nor evenly distributed. The distance between them would grow larger and larger. That means for example, if there were strings between them, that as they started moving, in an attempt to keep their distance uniform, the strings would stretch, and eventually break.

That is not what we would think of a uniformly accelerated reference frame, as we are accustomed to from non-relativistic physics. What is nice, in non-relativistic physics, about a uniformly accelerated reference frame, is that it keeps the same structure, the same shape. The distances between points stay the same. If you had strings connecting the points they wouldn't get strechted.

There is a second difficulty about the simplistic idea of uniformly accelerating a particle in special relativity : if we waited long enough, in the stationary frame, the particle would eventually exceed the speed of light.

So uniform acceleration, to the extent that it exists and makes good physical sense, is not as simple as just moving the points in figure 10 all with the same accelerated trajectory.

So we are going to construct what a relativist would call a uniformly accelerated reference frame. But to do so, we need to go back one step, to Euclidean space and talk about polar coordinates, fig. 11, because – surprisingly enough – the uniformly accelerated coordinate system is the analog of polar coordinates.

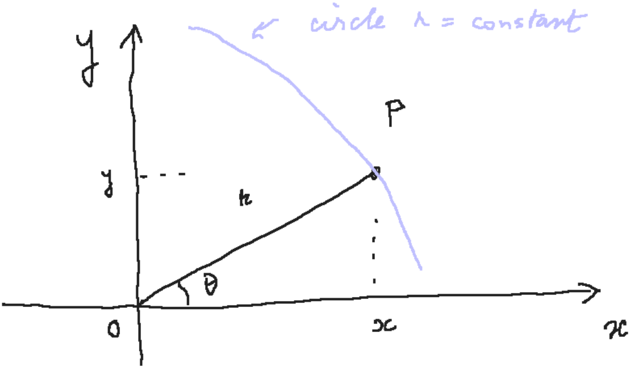

Figure 11 : Euclidean and polar coordinates in the plane.

Here are some equations, expressing the coordinate transformation from polar to Cartesian coordinates, which the reader should be familiar with

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned} \tag{25}$$

We also have

$$\cos^2\theta + \sin^2\theta = 1 \tag{26}$$

which is the same as saying that

$$x^2 + y^2 = r^2 \tag{27}$$

Finally there are two more equations to remember

$$
\begin{aligned}
\cos\theta &= \frac{e^{i\theta} + e^{-i\theta}}{2} \\
\sin\theta &= \frac{e^{i\theta} - e^{-i\theta}}{2i}
\end{aligned}
\tag{28}
$$

The reader can check that $\cos^2\theta$ plus $\sin^2\theta$ is equal to 1. It is a simple identity, true for all possible θ. Equations (25) to (28) are the basic equations governing ordinary polar coordinates.

What is the equation of a cicle around the origin ? It is just

$$r = \text{constant}$$

Imagine a point moving around the circle with uniform velocity, therefore uniform angular velocity. Then the magnitude of the acceleration of that point is constant around the circle. The vector acceleration constantly points toward the center of the circle.

What does it have to do with relativity ? In relativity we write basically the same equations to define a uniformly accelerated point.

We turn to the basic representation of space-time, which is the analog in special relativity of figure 11. In figure 12, we see the light cone : the two diagonal straight lines in grey. From the volume 3, in the collection *The Theoretical Minimum*, on special relativity, we know that they represent the trajectory of a light ray starting at time 0 from the origin and going either to the right or to the left.

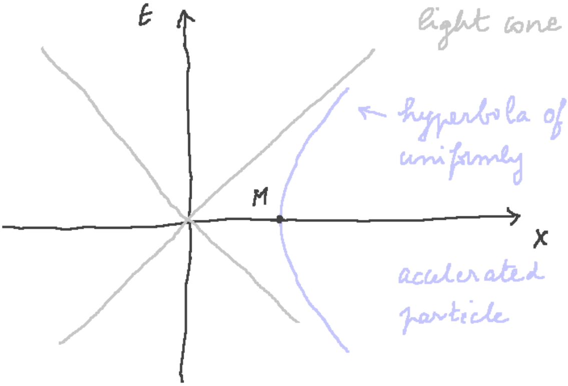

Figure 12 : Light cone (it would be a cone if we had two spatial coordinates).

Let's see now what is the analog of the uniformly accelerated point on the circle of figure 11.

In other words, what is a uniformly accelerated point in special relativity ? *It is defined* as a point moving on a hyperbola as in figure 12. The point is clearly accelerated. It is not moving with constant velocity. Constant velocity would correspond to a straight line with some slope higher the 45°.

In fig. 12, we see that from the past until time 0, the point, or particle, moves, spatially on the X-axis, to the left to a minimum point M. At M its velocity is zero. Therefore, in the $(X,\ t)$ diagram, at M the tangent to the trajectory is vertical.

Then, after point M, the particle is moving again to the right. In the $(X,\ t)$ diagram, as the point moves up and up, the tangent to the trajectory gets closer and closer to

45°, that is the particle gets closer and closer to the speed of light, but never exceeds it.

Next step is the following mathematical question : what is for a hyperbola the analagous of the equation for a circle in ordinary polar coordinates ? Answer : there exists a system of coordinates in which the hyperbola of fig. 12 has a very simple equation akin to equation (25) for the circle in polar coordinates. It uses hyperbolic polar coordinates.

In equation (25), expressing the transformation of coordinates between Cartesian and ordinary polar, we shall replace the sine and cosine functions by their hyperbolic version. Let's also replace the angle θ by a parameter ω. The correspondence is

$$\begin{aligned} \cos\theta &\rightarrow \cosh\omega \\ \sin\theta &\rightarrow \sinh\omega \end{aligned}$$

The mathematical definitions of the hyperbolic sine and cosine functions are very similar to those of ordinary sine and cosine. But there is no more $i = \sqrt{-1}$ coefficient.

$$\begin{aligned} \cosh\omega &= \frac{e^{\omega} + e^{-\omega}}{2} \\ \sinh\omega &= \frac{e^{\omega} - e^{-\omega}}{2} \end{aligned} \tag{29}$$

Analogously to equation (26), the reader can verify that

$$\cosh^2\omega - \sinh^2\omega = 1 \tag{30}$$

The coordinates of a point P in the $(X,\ t)$ diagram are now

$$\begin{aligned} X &= r\cosh\omega \\ t &= r\sinh\omega \end{aligned} \tag{31}$$

Equations (31) *define* r and ω from X and t. The parameter ω is not a geometric angle. But when we move along a hyperbola with the light-ray trajectories as asymptotes, see figure 12, it is what increases from $-\infty$ to $+\infty$ – just like θ was the parameter that changed as we moved along a circle centered at the origin. And on such a hyperbola, r doesn't change. The parameter ω plays on the hyperbola the role of the angle on the circle. It is sometimes called the *hyperbolic angle.*

As before, equations (31) express nothing more than a coordinate transformation between the Minkowski coordinates $(X,\ t)$ and the hyperbolic coordinates $(r,\ \omega)$.

Any point P in space-time can be located by its Minkowski coordinates $(X,\ t)$ or by its hyperbolic coordinates $(r,\ \omega)$.

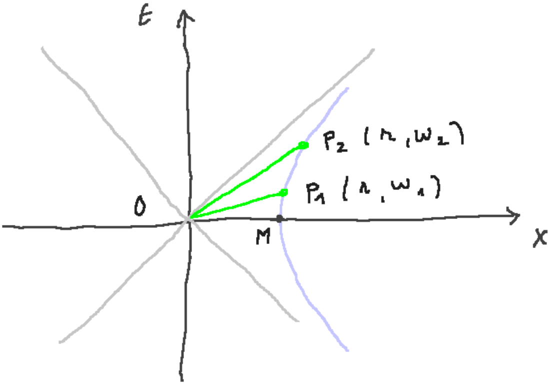

Figure 13 : Hyperbola in hyperbolic coordinates.

In figure 13, all the points on the hyperbola have the same *hyperbolic radius* r. Its value is the distance between O and

M. And it is a characteristic of the curve. On the other hand, the hyperbolic angle ω increases up to infinity as we move on the hyperbola closer and closer to its asymptote, that is, as the particle moves spatially farther and farther away on the X-axis.

So we have in Minkowski geometry the analog to the circle in Euclidean geometry : the hyperbola, as in figures 12 or 13, in hyperbolic polar coordinates given by equations (32), corresponds to a constant value r, and the parameter ω going from $-\infty$ to $+\infty$. This will be handy to study a uniformly accelerated particle because by definition it moves along such a trajectory.

Uniform acceleration

Now that we have a good understanding of what r and ω are in the hyperbolic coordinate system, which do you think is like time ?

On the X-axis, $\cosh \omega = 1$. So if we move to the right on this axis, we just increase r. Therefore *r is like a space coordinate.*

If from point M we travel upward on the hyperbola of figure 13, r stays fixed and we increase ω. Going upward is the analogy with traveling around the circle in figure 11. Therefore *ω is like a time coordinate.*

Just as there was a uniformity to the circle – at any point you could define the radius r and it was constant –, there is an analog uniformity to the hyperbola : the hyperbolic radius r is constant on a hyperbola. In figure 14, we can see hyperbolas for different values of r.

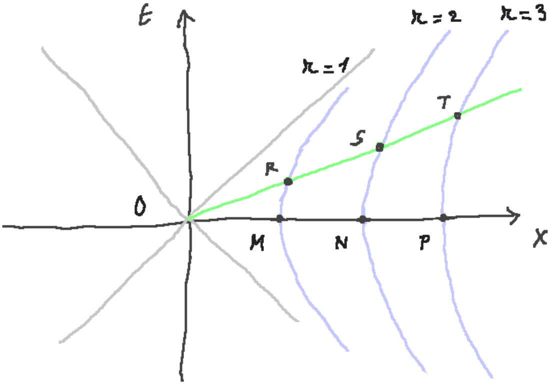

Figure 14 : Hyperbolas for different values of r.

The analog of equation (27) on a circle becomes on a hyperbola

$$X^2 - t^2 = r^2 \tag{32}$$

The hyperbolic polar coordinates are just new coordinates for every point in the standard Minkowski diagram $(t,\ X)$ of figure 14. Those coordinates in special relativity are the closest thing that exists to uniformly accelerated coordinates.

For a given ω the distances between the points corresponding to $r = 1$, $r = 2$, $r = 3$ etc. stay the same. In other words, in figure 14, not only $MN = NP$ but also $MN = NP = RS = ST$. That can be checked with the tools we learned in volume 3 of the collection *The Theoretical Minimum*, on special relativity.

That means that as the Lorentz frame of reference accelerates the distance between neighboring particles just stay the same. Notice also that points O, M, N and P are simultaneous, with $t = 0$. And points O, R, S and T are also simultaneous for another observer which is moving with a certain speed relative to O.

Exercise 2 : In figure 14, what is the speed, relative to the stationary frame, of the observer who sees R, S and T simultaneous ?

What is unusual here, and different than an ordinary accelerated frame of reference, is that the accelerations along the different trajectories corresponding to $r = 1$, $r = 2$, $r = 3$, etc. are different.

Let's look at a trajectory in the same collection as the hyperbolas of fig 14, but with a very small r, figure 15.

On the hyperbola with a very small r, the particle makes a sudden change of direction when it comes close to the origin, and then speeds off to the right again. That indicates that its trajectory has a much larger acceleration.

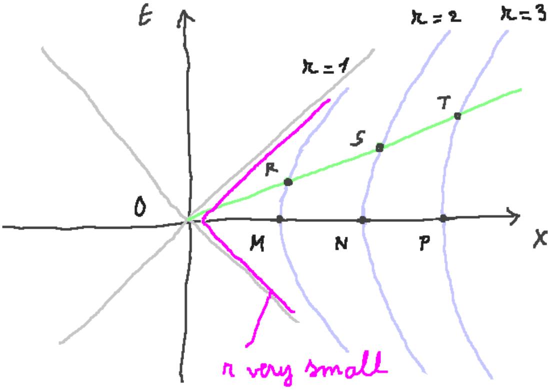

Figure 15 : Hyperbola with a very high acceleration.

Another important fact : on a given hyperbola, the proper acceleration, that is, the actual acceleration – the actual push – that an observer would feel, is uniform along the trajectory. This is the analog of the fact, on a given circle in fig 11, that at any point on the circle an observer feels the same acceleration in magnitude toward the center.

On the other hand, the accelerations on two different hyperbolas are different. Again that is the analog of the fact that on a circle with a larger radius the centripetal acceleration is smaller. Conversely the tinier the radius the bigger the acceleration.

The message to remember is this : if you want to define uniform acceleration, you ought to have acceleration which is constant in time. That is okay. But then you are necessarily going to have different accelerations at different points of space, that is, on trajectories with different r's.

The proper acceleration of a particle with coordinates (t, X) [5], is defined by $d^2X^n/d\tau^2$. It is the rate of change of the proper velocity, with respect to proper time. The proper velocity being itself $dX^n/d\tau$.

We have seen many times now the definition of proper time. Remember that it has a physical meaning. It is the time recorded and displayed by a watch attached on the wrist of an observer travelling in space-time.

So, if we consider the collection of particles of figure 10, fixed with respect to each other, the further out the particle is in space, that is, the further out to the right on the X-axis, the smaller its acceleration is. In other words, the further to the right, the straighter is its trajectory.

In fact we can write an equation for the proper acceleration. The proper acceleration, A, of a particle on one of the trajectories in figure 15 depends only on r. Let's choose one of them, and call it R. And let's work in units where the speed of light c is equal to 1. Then along the hyperbola with parameter R, the proper accelaration is given by

$$A = \frac{1}{R} \tag{33}$$

The bigger the R, the smaller the acceleration. It is the same thing that was going on on the circle in figure 11 : the smaller the radius of the circle was, the bigger the acceleration had to be for the particle to stay on the circle.

Equation (33) is does not look consistent unit-wise. What is the unit of acceleration ? It is length divided by time

[5] Here X stands for (X^1, X^2, X^3) or equivalently (x, y, z).

divided by time, i.e. l/t^2. Let's rewrite this dimension as

$$\frac{1}{l}\ \frac{l^2}{t^2}$$

In other words, acceleration has unit 1 over a length times a velocity squared. So in equation (33) – we are now familiar with this kind of reasoning on dimensions – there is a 1 actually with the dimension of the square of a velocity. It is the speed of light, with value 1 in the units chosen. If we worked in general units where the speed of light has value c, equation (33) would be

$$A = \frac{c^2}{R} \tag{34}$$

This means that for a fixed R, at human scale, the acceleration of a particle, in the collection of particles of figure 10, is very large. We have to go to a very large R before we get to particles with a moderate acceleration.

We usually work in this book with units in which $c = 1$, but the reader should keep in mind that the acceleration of a particle, whose coordinates are given by equation (31) with r fixed, is very big unless r is very big.

By the way, the acceleration on a given trajectory in figure 12, for example the acceleration on the hyperbola $r = 2$, is the ordinary acceleration at point N. And it is the constant acceleration that we would experience all along the trajectory if we rode a particle.

Notice that the uniformly accelerated reference frame we have been studying is defined relative to a particular origin.

The hyperbolic coordinates r and ω depend on the stationary frame selected. We could change for another stationary frame. This would produce different r and ω. But no real physics depends on the choice of the stationary frame.

Uniform gravitational field

We have introduced a somewhat arbitrary set of coordinates for our stationary frame[6]. In that set of coordinates we are now going to write the equation of motion along a geodesic. And we are going to see that the equation of motion along a geodesic looks like a particle falling in a uniform gravitational field.

Let's first talk about the metric of the Euclidean plane in ordinary polar coordinates as we did in figure 11

$$dS^2 = r^2 d\theta^2 + dr^2 \tag{35}$$

This just says that the distance between a point P, with coordinates $(r,\ \theta)$, and a close neighbor on the same circle, with coordinates $(r,\ \theta + d\theta)$, see fig. 11, for a fixed $d\theta$, increases linearly with r

This is the metric of the plane in polar coordinates. Notice that the metric is not the Kronecker delta. It has dr^2. The

[6] It just has to be Galilean, that is, it has to be such that in it Newton's equation has the simple form $F = ma$.

matrix of this two-dimensional metric looks like this

$$g_{mn} = \begin{pmatrix} r^2 & 0 \\ 0 & 1 \end{pmatrix} \tag{36}$$

Why is it not the Kronecker delta. Not because the space is curved, but because the coordinates are curvilinear. The space itself is flat. Indeed it is the plane, and we can go back to Cartesian coordinates $(x,\ y)$ in which the metric is the Kronecker delta.

Staying in the flat plane, the analog with the hyperbolic coordinates $(r,\ \omega)$ is

$$d\tau^2 = r^2 d\omega^2 - dr^2 \tag{37}$$

We are considering only two dimensions, time and one spatial coordinate[7]. For the moment we will ignore y and z, because we don't need them. We will be thinking about a particle falling in a gravitational field along a vertical axis denoted X or x. The coordinates y and z would be the other spatial coordinates and they don't matter for the problem we are going to discuss.

So the two coordinates that we will be interested in are ω and r, that is respectively, time and distance from the origin. And equation (37) is the metric.

We want to see what this metric has to do with gravitation.

[7] We express the metric in terms of the proper time τ rather than proper distance S. This makes no essential difference. It is the same geometry. When we want specifically to talk about a distance we will be careful to change the sign of the right hand side of equation (37).

Gravitation is supposed to have something to do with the metric. So consider again figure 15, which represents a bunch of particles fixed with respect to each other, in a frame that is uniformly accelerated relative to the stationary frame. The spatial x-axis, even though we said that it is a vertical axis along which a particle is falling, is represented as usual horizontally.

As we explained in the previous sections, a uniformly accelerated frame is a more subtle idea than it seems. In fact each particle keeps a constant acceleration on its trajectory. But the accelerations of different particles are different.

As we move to the right, with r increasing, the acceleration of the corresponding particle is smaller and smaller, fig. 15. If we go very far away, we can find a particle, corresponding to a hyperbolic radius R, and whose acceleration is g equal to 10 meters per second per second. Remember that the formula for its acceleration is

$$\frac{c^2}{R}$$

We set this equal to g. It gives

$$R = \frac{c^2}{g}$$

We have to go out this distance from O to find a particule with acceleration g. The speed of light c is 3 times 10^8 meters per second, so c^2 is approximately 10^{17}, and R about 10^{16} meters. Therefore we have to go out 10 000 billion kilometers to find a particle with approximately the acceleration of the Earth gravitational field on its surface. So let's go there.

Moreover, while there, if we don't move too much along the r direction, the acceleration $g = c^2/R$ won't change much. It is similar to moving on a vertical axis near the surface of the Earth : the gravitational field doesn't change much.

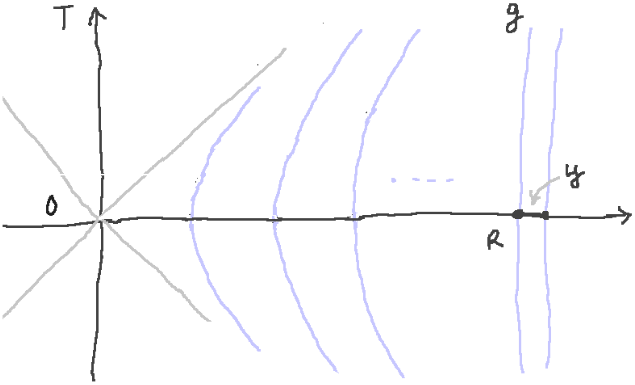

Figure 16 : Hyperbolas at R and at $R + y$.

Think of the X-axis as the axis of fall of a particle. Falling would be going to the left. And for the heck of it, we call y the distance from R on that axis. Also for convenience, we no longer call the time, t. We call it T.

y is a new first coordinate in the hyperbolic polar coordinates $(r,\ \omega)$.

$$y = r - R \tag{38}$$

All the particles with a moderate y have the same acceleration g.

Then we rewrite the metric of equation (37) using the new local coordinate y.

$$d\tau^2 = (R^2 + 2Ry + y^2)d\omega^2 - dy^2 \tag{39}$$

This creates an even more complicated looking metric, but not much. And we shall simplify it, because we are going to focus on a limited region around R[8]. We rewrite equation (39) as

$$d\tau^2 = (1 + \frac{2y}{R} + \frac{y^2}{R^2})\ R^2\ d\omega^2 - dy^2 \qquad (40)$$

One more step concerning coordinates : we give $R\omega$ a new name, we call it t. Before rewriting again equation (40), let's notice that y/R is very small. y is measured in meters or kilometers – for instance the stunt man Felix Baumgartner, in 2012, dived from 39 km – while R is 10^{16} meters. And y^2/R^2 is even much smaller. Wee are going to keep y/R and neglect y^2/R^2. Equation (40) becomes

$$d\tau^2 = (1 + \frac{2y}{R})\ dt^2 - dy^2 \qquad (41)$$

We end up with a metric which apart from the term $2y/R$, which is small, just looks like the good old Minkowski metric $dt^2 - dr^2$. It is space and time in a more or less ordinary way, but with a little correction.

The little correction $2y/R$ is what accounts for gravitation in the accelerated reference frame.

Keep in mind, however, that we are really talking about flat space. So far we have not introduced any curvature.

[8] Notice that when we talk about r, R, y or ω, we no longer talk about the Minkowski coordinates $(x,\ t)$, but about the hyperbolic polar coordinates $(r,\ \omega)$. There is, however, a simple correspondence between r and x : when $\omega = 0$, they are equal. That is why we can talk casually about R on the X-axis, while also meaning all the points on the hyperbola intercepting the horizontal axis at that point, because they all have hyperbolic radius R.

The space we are working in is really a flat space. With a change of coordinates we can go back to the Minkowski metric. Therefore any gravitation that we find is in a sense the same fake gravitation that we found in the accelerated elevator of chapter 1.

We are studying physics in an accelerated coordinate system. It is the elevator being pulled toward the right. And what do we expect to find ? We expect to find that in that elevator there is an *effective gravitational field.* We can also call it a *fictitious gravitational field.* It is this field that is associated with the term $2y/R$ in equation (41).

To get a better understanding of the connection, let's now study the motion of a particle in a metric given by equation (41). In units where c is equal to 1, we have $g = 1/R$. The metric can be rewritten

$$d\tau^2 = (1 + 2gy)\ dt^2 - dy^2 \tag{42}$$

Have you ever seen the expression gy in studying gravitation in a uniform field ? If we introduce the mass m of a the particle, mgy is simply the potential energy. The term gy is called the *gravitational potential.* And the term $(1 + 2gy)$ is one plus twice the gravitational potential.

It is extremely general. In any kind of gravitational field, as long as it is more or less constant with time, and not doing anything too radically relativistic, the coefficient in front of dt^2 in the metric is always one plus twice the gravitational potential.

Why do we call gy the gravitational potential other than that it just looks like it ? Answer : because if we work out the equation of motion of a particle in the metric given by equation (42), we will find that, as long as the particle is moving slowly, as long as we can make a good Newtonian approximation, as long as things are not too relativistic, the equation of motion that we will find is the same as that of a particle falling along the y-axis in a uniform gravitational field, as we have calculated it in classical mechanics, see volume 1 in the collection *The Theoretical Minimum.* What we mean by the y-axis of course is still the unique spatial axis, but near point R.

The whole point of the preceding section on uniform acceleration was to explain that a uniformly accelerated reference frame is something more subtle than just accelerating a thirty meter long steel beam along the X-axis. We had to *define* what we meant by a uniformly accelerated frame. It lead to this funny construction where points at different distances from O, measured for instance at time 0, each have a fixed acceleration, but the acceleration differs from point to point. On the other hand we can check with a Lorentz transformation, in the frame of one of the moving points at velocity v, that the distances measured simultaneously by P between the points don't change. And it also lead to hyperbolic coordinates and to figure 14.

Aside from that, it is really ordinary physics. There is an accelerated elevator at M, see figure 17, there is another accelerated elevator at N, there is another one at P, there is one at R, etc. In other words, it is just a bunch of elevators at different positions each being accelerated. The only specific idea is that, for them to form a uniformly accelerated frame, they each must have a different acceleration.

They have to be accelerated, according to equation (34), with acceleration c^2/R.

We are interested, in figure 17, in the accelerated elevator at hyperbolic radius R. Of course the elevator should be imagined on its side. The bottom of the elevator, to be precise, follows the trajectory with hyperbolic radius R, that is, the thick trajectory in figure 17.

It is important to understand that the trajectory *in space-time* of the floor of the elevator is not a little horizontal segment going to the right on the X-axis – although it is true that if we look only at the space coordinate X, the elevator does move very slightly to the right. Its trajectory in space-time is the almost vertical line, in figure 17, because it has only acceleration g, which is rather small.

And in the hyperbolic coordinates (r, ω), *the hyperbolic radius of the elevator remains* R. To the right of R there are other trajectories – trajectories of points inside the elevator –, and they also have an acceleration very close to g.

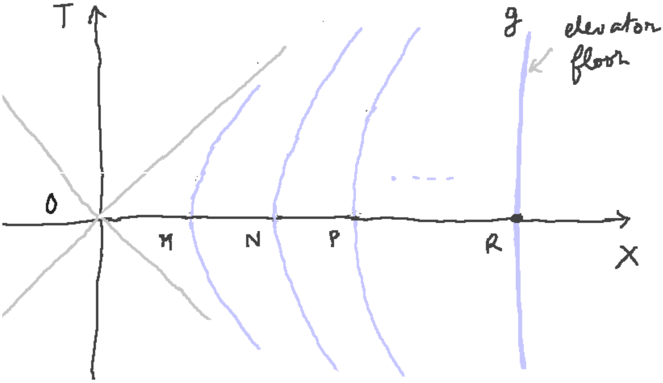

Figure 17 : Floor of the elevator.

The local coordinate y – it is hyperbolic radius shifted by R – is a coordinate to locate things inside the elevator, that is, it is the height of a point or a particle above the floor.

The metric of space-time in the coordinate system $(t,\ y)$, which we built to locate things inside the uniformly accelerated elevator, is given by equation (42), which we reproduce below

$$d\tau^2 = (1 + 2gy)\ dt^2 - dy^2$$

We had to do a little bit of work to arrive at this equation, but now for us it is a given. It is the metric tensor, in the coordinate system $(t,\ y)$, at points in the vicinity of a point moving with acceleration g. And in order to show the equivalence of a uniform acceleration and a gravitational field, we chose for the acceleration g the same value as the gravitational field of the Earth.

What is the rule to figure out how a particle moves ?

The rule about particle motion is that particles move on geodesics – not geodesics of space but geodesics of space-time. In other words, we take the metric of space-time, whatever it is, and we go through exactly these same operations.

The metric of space-time that we are given is equation (42). And the equation of motion of a particle is equation (19), which we reproduce below, with a change of sign and a new numbering,

$$-\frac{d^2X^n}{dS^2} = \Gamma^n_{mr}\ t^r\ t^m \tag{43}$$

It is the equation of motion which says : it is a geodesic. Go straight ahead. But not straight ahead in space now, *straight ahead in space-time.* Otherwise the equations are the same.

We shall write it slightly differently, using $d\tau$ instead of dS. Remember that $d\tau^2$ is just the opposite of dS^2. So the left hand side of (43) becomes

$$\frac{d^2X^n}{d\tau^2} \tag{44}$$

This is called the proper acceleration. As long as the elevator is moving slowly, in other words if it hasn't been accelerating long enough to get up near the speed of light, then the proper time and the ordinary time are essentially the same. And expression (44) is just the ordinary acceleration.

We choose X to be y. We want the y component of acceleration. Then expression (44) is simply

$$\frac{d^2y}{d\tau^2}$$

Now we turn to the right hand side of equation (43). The n-th component of X stands for y. What is t^r ? It is dX^r/dS, because we are on a geodesic. And dS is $id\tau$. So equation (43) becomes

$$\frac{d^2y}{d\tau^2} = -\Gamma^y_{mr}\, \frac{dX^r}{d\tau}\, \frac{dX^m}{d\tau} \tag{45}$$

Since m and r each run over four coordinates[9] , the right hand side has a whole bunch of terms – ten of them to

[9] To be consistent with standard notations in relativity, it would be better to use μ and ν. We leave to the reader to make the change in these dummy variables.

be precise because the gammas are symmetric in m and r. Fortunately most of them are extremely small as long as the elevator is moving slowly, and as long as the movement of the object we are interested in, namely the particle with coordinate y, is slow. Under these conditions, only one of the combinations $\Gamma^{y}_{mr}\ \frac{dX^r}{d\tau}\ \frac{dX^m}{d\tau}$ is significant.

What is the value of $dt/d\tau$ for slow motion ? It is essentially 1, because time and proper time in that case are almost the same.

On the right hand side of equation (45) the differential elements are the components of the 4-velocity of the particle. We just saw that $\frac{dX^0}{d\tau}$ is essentially 1. What are the derivatives of the space components with respect to τ ? They are proportional to their actual ordinary spatial velocity. We are assuming the spatial velocity is small compared to the speed of light so the only important contribution on the right hand side of (45) comes when r and m are time indices. Let's use t instead of 0 for the time index. Equation (45) reduces to

$$\frac{d^2y}{d\tau^2} = -\Gamma^{y}_{tt} \tag{46}$$

The right hand side must be the gravitational force. It must be the derivative of the gravitational potential energy.

Let's go back to the expression of the Christoffel symbols in terms of the metric

$$\Gamma^{p}_{rs} = \frac{1}{2}\ g^{pn}\ \left[\ \frac{\partial g_{nr}}{\partial X^s} + \frac{\partial g_{ns}}{\partial X^r} - \frac{\partial g_{rs}}{\partial X^n}\ \right] \tag{47}$$

We need the one with two time covariant index and one space contravariant index. The space index is y. Among the terms g^{yn}, the only term which is not negligeable is g^{yy} and it is 1. Since X^t is just what we denote t and X^y is what we denote y, we get

$$\Gamma^y_{tt} = \frac{1}{2} \left(\frac{\partial g_{yt}}{\partial t} + \frac{\partial g_{yt}}{\partial t} - \frac{\partial g_{tt}}{\partial y} \right)$$

The terms $\frac{\partial g_{yt}}{\partial t}$ and $\frac{\partial g_{yt}}{\partial t}$, which are equal, are both zero. So finally

$$\Gamma^y_{tt} = -\frac{1}{2}\frac{\partial g_{tt}}{\partial y}$$

And equation (46) can be rewritten

$$\frac{d^2y}{d\tau^2} = \frac{1}{2}\frac{\partial g_{tt}}{\partial y} \tag{48}$$

An equation like (48), where the second derivative of a spatial variable y with respect to time is proportional to the first derivative of some quantity with respect to y, reminds us of an equation of motion with potential energy. Somehow one half of g_{tt} must be the opposite of a potential energy. But we saw that : it is minus a potential energy with $m = 1$, also called gravitational potential.

In equation (42), which we rewrite below,

$$d\tau^2 = (1 + 2gy)dt^2 - dy^2$$

g_{tt} is the coefficient $-(1 + 2gy)$ in the metric defining dS^2, which is the same as $d\tau^2$ with a minus sign. So one half the derivative of g_{tt} with respect to y is $-g$. Equation (48) finally becomes

$$\frac{d^2y}{d\tau^2} = -g \tag{49}$$

That is the equation of motion of a particle in a uniform gravitational field. We went through a rather complicated derivation to reach it, but so doing we learned the following points :

1. Space-time has a metric. In arbitrary coordinates, the metric can have a fairly complicated structure. In uniformly accelerated coordinates, however, it is almost the Minkowski metric, but with the extra term $2gy$ in equation (42).
2. The equation of motion along a geodesic in space-time – at least as long as things are going slowly, that is as long as Newtonian approximation is valid [10] – is just Newton's equation in a uniform gravitational field.

Uniform gravitational field, constant acceleration... : it is what we expected. But to do it properly, using the metric, the Christoffel symbols, the geodesics and so forth, is a fairly complicated procedure.

Einstein guessed it : the hypothesis that a particle moves along a geodesic in space-time was his starting point, and he went in the opposite direction. He knew about uniformly accelerated coordinate systems, but he didn't know about Christoffel symbols. Somewhere along our own derivations is where he started. And for a uniform acceleration – with Newtonian approximation – the metric is simply given by equation (42).

So we have come around full circle from the fist lesson where we talked about accelerated elevators giving rise to gravitation. We have shown that in the Minkowski-Einstein space-

[10] This means setting the terms where c appears in the denominator to zero.

time a uniformly accelerated reference frame does give rise to an effective gravitational field.

But so far we haven't gotten to real gravitational fields. The gravitational field we got is not a real gravitational field because it really corresponds to a flat space.

If we were to take the metric of equation (42) and calculate the curvature tensor it would be exactly zero, indicating that there does exist coordinates where the metric has the simple form $dt^2 - dX^2$. So the gravitation that we are experiencing is really exactly the gravitation due to an accelerated frame of reference, not due to real gravitating matter.

Now we can guess what the effect of real gravitating matter would be. Instead of 2gy in equation (42), what is the gravitational potential due to a gravitating object ? It is $-G/y$[11].

So we can expect, when we study the metric of a real gravitational field, that we will have something like

$$d\tau^2 = (1 - \frac{2GM}{y})\ dt^2 - dy^2 \tag{50}$$

where G is Newton's constant, and M is the mass of the gravitating object.

[11] By convention, for a uniform gravitational field, the gravitational potential is taken to be zero at ground level and increases to $+\infty$ when the height increases, while for the radial field created by an object, the gravitational potential is taken to be zero infinitely far away and goes to $-\infty$ when the radius goes to zero. That is why y is now in the denominator, and there is a minus sign.

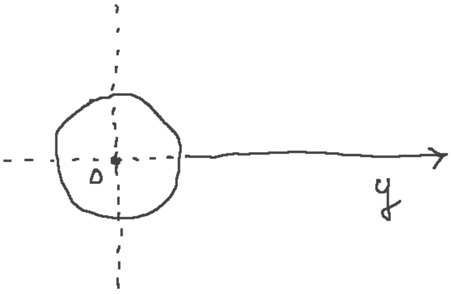

Figure 18 : Gravitating object of mass M
and gravitational potential $-G/y$.

That is almost the Schwarzschild metric, but not quite. We will work out what is the Schwarzschild metric of a gravitating object.

Equation (50) will lead to a weird phenomenon. When y is large, the term $2GM/y$ is small. That is good because $(1 - 2GM/y)$ is positive. But something crazy happens at the point where y is equal to $2GM$. The coefficient in front of dt^2 becomes zero. That point y where the coefficient changes sign is called the horizon of the black hole[12].

Real gravitational fields and the Schwarzschild metric will be the subject of next chapter. We are not going to derive the metric entirely from what we already know. To derive it, we need fields equations. We haven't discussed them yet,

[12] For the Earth, supposing all its mass was almost point-like, the horizon would be 9 millimeters.

and we won't do it until chapter 9.

So far we have only discussed geometry, flatness, curvature, geodesics, etc. And in this chapter, when we finally arrived at the space-time of relativity and its peculiar geometry, we ended up with a little demonstration of how, in a uniformly accelerated reference frame in space-time, movement along a geodesic gives rise to Newton's equation.