General Relativity Without Calculus

A Concise Introduction to the Geometry of Relativity

The idea here is to make you think about geometrical things more deeply before we proceed to a rigorous derivation of General Relativity

1. Special Relativity

1.1 Relativity of motion

Motion is relative.

As you read this you probably think that you are not moving.

To be precise, however, you should think that you are not moving *with respect* to the Earth's surface.

But the Earth spins - at the latitude of Lisbon, one spins at approximately 1,300 kilometers per hour.

Moreover, the Earth moves around the Sun, at about 30 kilometers per second, and the Sun moves around the center of the galaxy, at about 220 kilometers per second.

Therefore I am moving at 1,300 kilometers per hour with respect to the center of the Earth, at 30 kilometers per second with respect to the Sun, and at 220 kilometers per second with respect to the center of the galaxy!

1.2 Inertial frames

To study any motion we must first choose what is called a frame of reference.

A *frame of reference* is simply a system of coordinate axes with respect to which the coordinates of each point in space can be specified.

It is often attached to a solid object (for instance the Earth), but that is not strictly necessary.

When I said I was moving at 1,300 kilometers per hour with respect to the center of the Earth I wasn't being entirely accurate.

What I meant was that I am moving with this speed in the frame whose center is the center of the Earth but which does not spin.

This frame is (approximately) what is called an inertial frame, that is, a frame where the law of inertia holds: *any free particle moves in a straight line with constant speed*.

The frame attached to the surface of the Earth is not inertial because of the Earth's rotation, which prevents the law of inertia from holding exactly (this is revealed in certain experiments, as for instance the Foucault pendulum - see https://www.youtube.com/watch?v=pKGgICawAKc).

However, this frame can be considered an inertial frame for most purposes.

In reality, the non-rotating frame centered at the center of the Earth is also not an exact inertial frame (despite being a better approximation), because of the Earth's motion around the Sun.

Progressively better approximations are the non-rotating frame centered at the Sun and the non-rotating frame centered at the center of the galaxy.

In practice, "non-rotating" means "non-rotating with respect to the distant stars".

The suggestion that the matter in the Universe as a whole somehow determines the inertial frames is called the *Mach principle*.

The big 16th century controversy between geocentrism and heliocentrism was in part a discussion about frames.

In a way both parties were right: it is as accurate to say that the Earth moves around the Sun as saying that the Sun moves around the Earth.

In the first case we using the Sun's frame, while in the second case we are adopting the frame attached to the surface of the Earth.

However, since the Sun's frame is (more approximately) inertial, it allows a much simpler description of motion.

1.3 Galileo transformations

If S is an inertial frame then any other frame S' with parallel axes moving with constant velocity with respect to S is also an inertial frame.

Suppose that S' moves along the x-axis with constant velocity $\mathbf{v} = v\,\hat{e}_x$, and that the two frames coincide at time t = 0.

If at time t a given point P has coordinates (x, y, z) in S, then it should be clear from Figure 1.1
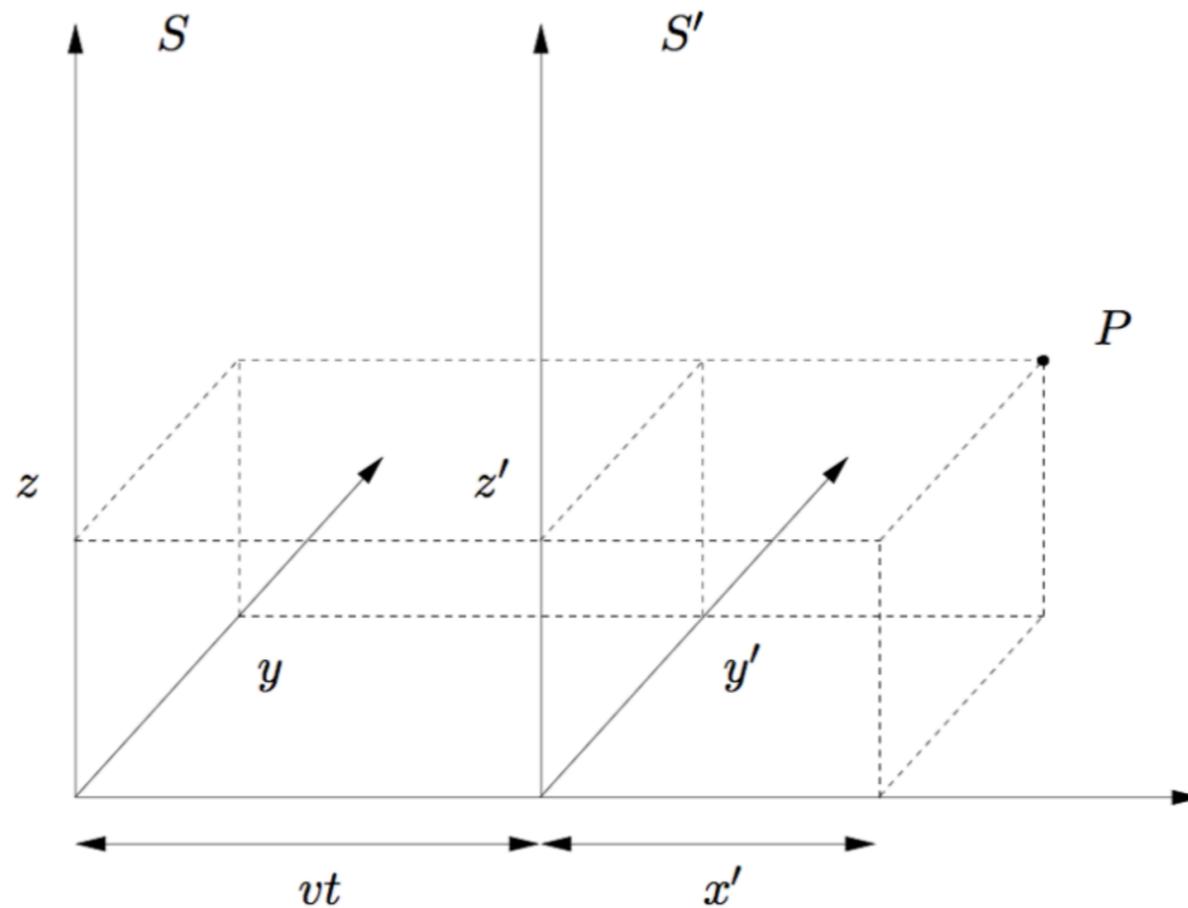
Figure: 1.1 Galileo transformation.

that its coordinates (x', y', z') in S' are

$$x' = x - vt$$

$$y' = y$$

$$z' = z$$

$$(1.1)$$

We should add to these equations

$$t' = t \qquad (1.2)$$

that is: the time measured in S' is the *same* as the time measured in S.

This seems so obvious that it is almost not worth writing down (and indeed for a long time no one did).

We shall soon see that it is actually wrong: as incredible as it may seem, the time measured in the two frames is not exactly the same.

The formulas above are said to define a *Galileo transformation* (which is nothing more than a change of inertial frame).

The inverse transformation is very simple:

$$
\begin{aligned}
t &= t' \\
x &= x' + vt' \\
y &= y' \\
z &= z'
\end{aligned}
\qquad (1.3)
$$

In other words, we *just* have to change the sign of **v**.

This is what one would expect, as S is moving with respect to S' with velocity -**v**.

1.4 Velocity addition formula

A consequence of the Galileo transformations is the velocity addition formula.

Suppose that the point P is moving along the x-axis.

Let u be the instantaneous velocity of P in the frame S.

This means that

$$u = \frac{\Delta x}{\Delta t} \qquad (1.4)$$

where $\Delta x = x_2 - x_1$ is the distance travelled by P between the positions x_1 and x_2 (as measured in S) in a very small time interval $\Delta t = t_2 - t_1$ between times t_1 and t_2 (also measured in S).

In S', P moves by $\Delta x'$ between positions x'_1 and x'_2 in the time interval $\Delta t'$ between times t'_1 and t'_2.

The values of t'_1, t'_2, x'_1, x'_2 are related to the values of t_1, t_2, x_1, x_2 by a Galileo transformation:

$$t'_1 = t_1$$

$$t'_2 = t_2$$

$$x'_1 = x_1 - vt_1 \tag{1.5}$$

$$x'_2 = x_2 - vt_2$$

Therefore

$$\Delta t' = \Delta t$$

$$\Delta x' = \Delta x - v\Delta t \tag{1.6}$$

and the instantaneous velocity of P in S′ is

$$u' = \frac{\Delta x'}{\Delta t'} = \frac{\Delta x - v\Delta t}{\Delta t} = \frac{\Delta x}{\Delta t} - v = u - v \tag{1.7}$$

In other words, the velocity u of P in S is simply the sum of the velocity u′ of P in S′ with the velocity v of P with respect to S.

1.5 Lorentz transformations

In 1887, to their great surprise, Michelson and Morley discovered that the speed of light is the same in all inertial frames.

This violates the velocity addition formula, according to which any object or signal is at rest in the inertial frame which moves with the same velocity.

It was Einstein who first understood that this meant that the Galileo transformations formulas could not be entirely correct, and had to be replaced by the Lorentz transformation formulas (previously discovered by Lorentz and Poincare, who had failed to correctly interpret them):

$$t' = \gamma \left(t - \frac{vx}{c^2} \right)$$
$$x' = \gamma(x - vt) \tag{1.8}$$

where c represents the speed of light (about 300,000 kilometers per second) and

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{1.9}$$

The special theory of relativity, developed by Einstein in 1905, boils down to analyzing the consequences of these transformations.

The velocities with which we usually deal are much smaller than the speed of light, $|v| \ll c$.

In this case γ is almost equal to 1 and vx/c^2 is almost equal to zero.

Therefore for most applications the Lorentz transformation formulas reduce to the Galileo transformation formulas.

It is only when the velocities involved become comparable to the speed of light that the Lorentz transformations become important.

It is easy to check that the inverse transformation formulas are obtained (as one would expect) replacing v by -v

$$t = \gamma \left(t' + \frac{vx'}{c^2} \right)$$
$$x = \gamma(x' + vt') \qquad\qquad (1.10)$$

1.6 Relativistic velocity addition formula

Notice that the Lorentz transformation formulas require $|v| < c$: given two inertial frames, the velocity of one of them with respect to the other must be less than the speed of light.

Therefore it is never possible for a light signal to be at rest in a given inertial frame.

More generally, the Lorentz transformations imply that light moves with the same speed in all inertial frames.

To check this fact we need the relativistic velocity addition formula.

Again assume that the point P is moving with instantaneous velocity u in S, traveling a distance Δx (as measured in S).

Then the displacement $\Delta x'$ measured in S' and the corresponding time interval $\Delta t'$ are given by

$$\Delta t' = \gamma \left(\Delta t - \frac{v \Delta x}{c^2} \right)$$

$$\Delta x' = \gamma (\Delta x - v \Delta t) \qquad (1.11)$$

Consequently, the instantaneous velocity of P in S' is

$$u' = \frac{\Delta x'}{\Delta t'} = \frac{\Delta x - \Delta t}{\Delta t - \frac{v \Delta x}{c^2}} = \frac{u - v}{1 - \frac{uv}{c^2}} \qquad (1.12)$$

In the special case when u = c we obtain

$$u' = \frac{c - v}{1 - \frac{v}{c}} = c \cdot \frac{c - v}{c - v} = c \qquad (1.13)$$

If on the other hand u = -c we have

$$u' = \frac{-c - v}{1 + \frac{v}{c}} = -c \cdot \frac{c + v}{c + v} = -c \qquad (1.14)$$

Therefore whenever P moves at the speed of light in S it also moves at the speed of light in S'.

1.7 Time Dilation

One of the most counter-intuitive consequences of the Lorentz transformations is the fact that the time interval between two events depends on the inertial frame in which it is measured.

Suppose that an observer at rest in the inertial frame S′ (Δx′ = 0) measures a time interval Δt′.

Then the corresponding time interval measured in the inertial frame S is

$$\Delta t = \gamma \left(\Delta t' + \frac{v \Delta x'}{c^2} \right) = \gamma \Delta t' > \Delta t' \qquad (1.15)$$

(as $\gamma > 1$ whenever v ≠ 0).

This phenomenon is known as **time dilation**.

1.8 Derivation of the Lorentz transformation formulas

In what follows we give a derivation of the Lorentz transformation formulas, due to Einstein.

Einstein started with two postulates:

1. **Relativity principle:** Any two inertial frames are equivalent.

2. **Invariance of the speed of light:** The speed of light is the same in all inertial frames.

Since the Galileo transformations are not compatible with the second postulate, we cannot expect the "obvious" formula x′ = x - vt to work.

Suppose however that x′ is *proportional* to x - vt, that is

$$x' = \gamma(x - vt) \tag{1.16}$$

for some constant γ (to be determined). Since S moves with respect to S′ with velocity -v, the first postulate requires an analogous formula for the inverse transformation

$$x = \gamma(x + vt) \tag{1.17}$$

Solving for t′ yields

$$t' = \frac{x}{v\gamma} - \frac{x'}{v} \tag{1.18}$$

Substituting in this formula the initial expression for x′ gives

$$t' = \left(\frac{1}{\gamma} - \gamma\right)\frac{x}{v} + \gamma t \tag{1.19}$$

We now use the second postulate.

Consider a light signal propagating along the x-axis in S, passing through x = 0 at time t = 0.

The position of the signal at time t will then be x = ct.

On the other hand, the second postulate requires that the position of the signal in S' be x' = ct'.

Therefore

$$c = \frac{x'}{t'} = \frac{\gamma(x - vt)}{\left(\frac{1}{\gamma} - \gamma\right)\frac{x}{v} + \gamma t} = \frac{\frac{x}{t} - v}{\left(\frac{1}{\gamma^2} - 1\right)\frac{x}{vt} + 1} = \frac{c - v}{\left(\frac{1}{\gamma^2} - 1\right)\frac{c}{v} + 1} \qquad (1.20)$$

implying that

$$\left(\frac{1}{\gamma^2} - 1\right)\frac{c}{v} + 1 = 1 - \frac{v}{c} \Leftrightarrow \frac{1}{\gamma^2} = 1 - \frac{v^2}{c^2} \Leftrightarrow \gamma = \pm \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad (1.21)$$

Since we must have $\gamma = 1$ for v = 0, we must take the positive sign.

Thus

$$\frac{1}{\gamma} - \gamma = \gamma\left(\frac{1}{\gamma^2} - 1\right) = \gamma\left(1 - \frac{v^2}{c^2} - 1\right) = -\gamma\frac{v^2}{c^2} \qquad (1.22)$$

and we finally obtain

$$\Delta t' = \gamma \left(\Delta t - \frac{v\Delta x}{c^2} \right)$$

$$\Delta x' = \gamma(\Delta x - v\Delta t) \hspace{4cm} (1.23)$$

with

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \hspace{4cm} (1.24)$$

2. Minkowski geometry

2.1 Units

Since the speed of light is the same for all observers, we may *without ambiguity* choose units in which $c = 1$.

For instance, we can measure time in years and distances in light-years (a light-year is the distance travelled by light during one year, approximately 9.5×10^{12} kilometers.

Alternatively, we can measure distances in meters and time in light-meters (a light-meter is the time it takes light to travel 1 meter, approximately 3.3 nanoseconds.

In these units speeds do not have dimensions: they are simply given by the fraction of the speed of light which they represent.

2.2 Space-time diagrams

To formulate the special theory of relativity geometrically we pick an inertial frame S.

Each event can be specified in this frame by giving the instant t and the position x in which it happened.

A space-time diagram consists in representing events as points in the plane with Cartesian coordinates (t, x).

It is traditional to represent the coordinate t vertically.

The motion of a particle can be represented in a space-time diagram by plotting its position x at each instant t; we thus obtain a curve, which we call the *history* of the particle.

We now consider a few examples (see Figure 2.1):

Figure: 2.1 Space-time diagram containing the histories of: (a) a particle at rest; (b) a particle moving with constant velocity; (c) a light ray; (d) an accelerating particle.
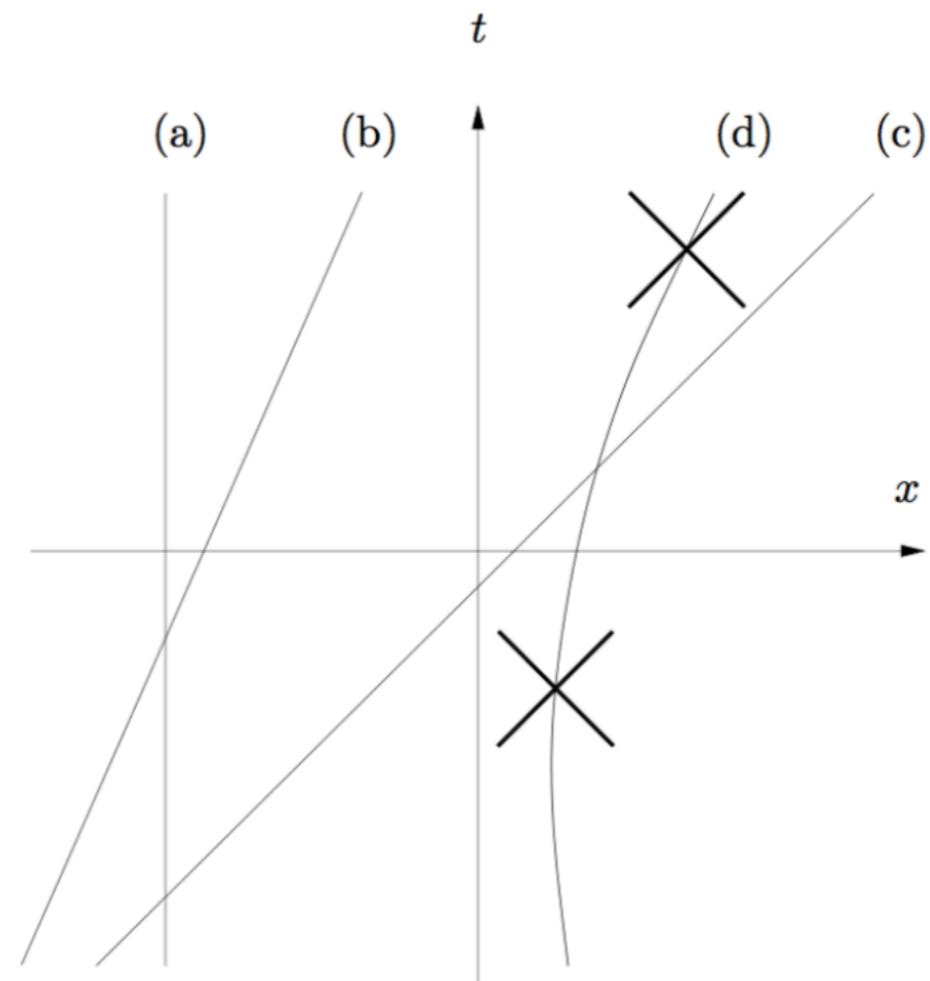
(a) If the particle is at rest in the frame S then its position x does not change with t : $x = x_0$, where x_0 is a constant.

Therefore the history of this particle is a vertical line.

(b) If the particle is moving with constant velocity v then its position at time t is $x = x_0 + vt$, where x_0 is a constant (representing the particle's position at $t = 0$).

The history of this particle is then a line with slope $1/v$ (since its equation can be re-written as $t = (1/v)(x - x_0)$).

(c) A light signal moves with constant velocity $\pm c = \pm 1$, and so its position at time t is $x = x_0 \pm t$, where x_0 is a constant (representing the signal's position at $t = 0$).

The history of the signal is a line with slope ± 1.

(d) If the particle moves with changing velocity then its history is a curve.

As we have seen, the Lorentz transformation formulas force the particle's speed to be smaller than $c = 1$.

Therefore if we imagine two light signals being emitted in opposite directions at each event in the particle's history (lines with slopes ± 1 through that event) then the particle's history cannot intersect the histories of those light signals in any other event.

2.3 Interval between events

To represent events on a space-time diagram we must pick an inertial frame S.

Clearly the representation will change if we choose a different inertial frame S', as the coordinates (t', x') of a given event in S' are in general different from its coordinates (t, x) in S.

This situation is analogous to what happens when we introduce Cartesian coordinates in the Euclidean plane.

To do so we must fix a system S of orthogonal axes.

However, the choice of these axes is not unique: for instance, we can choose a different system S' of orthogonal axes which are rotated by an angle α with respect to S (Figure 2.2).
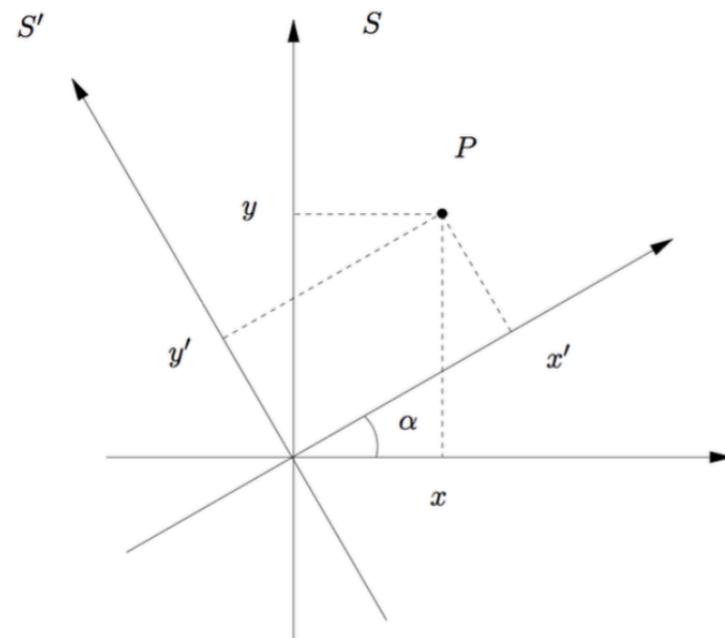

Figure: 2.2: Two systems of orthogonal axes.

If a given point P has coordinates (x, y) in S, its coordinates (x′, y′) in S′ are in general different.

Indeed, it is not hard to show that

$$x' = x \cos \alpha + y \sin \alpha$$
$$y' = -x \sin \alpha + y \cos \alpha \qquad (2.1)$$

The coordinates of point P are thus devoid of intrinsic geometric meaning, since they depend on the choice of the axes.

However, the introduction of a system of orthogonal axes allows us to compute quantities with intrinsic geometric meaning, such as the distance between two points.

Let P_1 and P_2 be two points with coordinates (x_1, y_1) and (x_2, y_2) in S.

The coordinates of these points in S′ will be (x'_1, y'_1) and (x'_2, y'_2), where

$$x'_1 = x_1 \cos \alpha + y_1 \sin \alpha$$
$$x'_2 = x_2 \cos \alpha + y_2 \sin \alpha \qquad (2.2)$$
$$y'_1 = -x_1 \sin \alpha + y_1 \cos \alpha$$
$$y'_2 = -x_2 \sin \alpha + y_2 \cos \alpha$$

If $\Delta x = x_2 - x_1$, $\Delta y = y_2 - y_1$, $\Delta x' = x'_2 - x'_1$, $\Delta y' = y'_2 - y'_1$ are the differences between the coordinates of P_2 and P_1 in each of the two systems, we have

$$\Delta x' = \Delta x \cos \alpha + \Delta y \sin \alpha$$
$$\Delta y' = -\Delta x \sin \alpha + \Delta y \cos \alpha \qquad (2.3)$$

The distance Δs between P_1 and P_2 can be computed in S from the Pythagorean theorem:

$$\Delta s^2 = \Delta x^2 + \Delta y^2 \qquad (2.4)$$

In S' this distance is given by

$$\Delta s'^2 = \Delta x'^2 + \Delta y'^2 \qquad (2.5)$$

Since this distance is a geometric property, it cannot depend on the system of orthogonal axes chosen to perform the calculation.

Indeed:
$$\begin{aligned}
\Delta s'^2 &= \Delta x'^2 + \Delta y'^2 \\
&= (\Delta x \cos \alpha + \Delta y \sin \alpha)^2 + (-\Delta x \sin \alpha + \Delta y \cos \alpha)^2 \\
&= \Delta x^2 \cos^2 \alpha + \Delta y^2 \sin^2 \alpha + 2\Delta x \Delta y \sin \alpha \cos \alpha \\
&\quad + \Delta x^2 \sin^2 \alpha + \Delta y^2 \cos^2 \alpha - 2\Delta x \Delta y \sin \alpha \cos \alpha \\
&= \Delta x^2 (\sin^2 \alpha + \cos^2 \alpha) + \Delta y^2 (\sin^2 \alpha + \cos^2 \alpha) \\
&= \Delta x^2 + \Delta y^2 = \Delta s^2 \qquad (2.6)
\end{aligned}$$

By analogy, we define the "distance" $\Delta\tau$ between two events P_1 and P_2 with coordinates (t_1, x_1) and (t_2, x_2) in S as

$$\Delta\tau^2 = \Delta t^2 - \Delta x^2 \qquad\qquad (2.7)$$

where $\Delta t = t_2 - t_1$ and $\Delta x = x_2 - x_1$.

Note that $\Delta\tau$ is *not* the Euclidean distance between the two events in the space-time diagram, because of the minus sign.

Also note that we can only define the distance between two events such that $|\Delta x| \leq |\Delta t|$ or $v \leq c$.

Pairs of events satisfying this relation are said to be **causally related**, since only in this case can one of them cause the other: the maximum speed of propagation of any signal is the speed of light, and so $\left|\frac{\Delta x}{\Delta t}\right| \leq 1$ along the history of any signal.

To the "distance" $\Delta\tau$ we call the interval between the causally related events P_1 and P_2.

Surprisingly, the interval does not depend on the inertial frame in which it is computed:

$$\Delta\tau'^2 = \Delta t'^2 - \Delta x'^2 = \gamma^2(\Delta t - v\Delta x)^2 - \gamma^2(\Delta x - v\Delta t)^2$$
$$= \gamma^2(\Delta t^2 + v^2\Delta x^2 - 2v\Delta t\Delta x - \Delta x^2 - v^2\Delta t^2 + 2v\Delta t\Delta x)$$
$$= \gamma^2(1 - v^2)\Delta t^2 - \gamma^2(1 - v^2)\Delta x^2$$
$$= \Delta t^2 - \Delta x^2 = \Delta\tau^2 \tag{2.8}$$

(where we used the formulas for the Lorentz transformations with c = 1.)

We can therefore regard special relativity as the study of a geometry, different from the usual Euclidean geometry, in which the *distance* between two points is replaced by the *interval* between two events.

This new geometry is called **Minkowski geometry**.

What is the physical meaning of the interval between two events?

If $\Delta\tau^2 > 0$, we have $|\Delta x| \leq |\Delta t|$

Therefore there exists an observer with constant velocity v which is present at both events, since

$$|v| = \left|\frac{\Delta x}{\Delta t}\right| < 1 \tag{2.9}$$

Suppose, in this observer's inertial frame S′, the two events happen in the same location, $\Delta x' = 0$.

Therefore in this frame
$$\Delta\tau = |\Delta t'| \tag{2.10}$$

We conclude that the **interval between two events represents the time measured between them by an inertial observer which is present at both events** (if nonzero).

If $\Delta\tau = 0$, we have $|\Delta x| = |\Delta t|$, and hence **if the interval between two events is zero then these events are placed on the history(path) of a light signal**.

2.4 Generalized twin paradox

The twin paradox: Two twins, Alice and Bob, part on their 20th birthday.

While Alice remains on Earth (which is an inertial frame to a very good approximation), Bob departs at 80% of the speed of light towards Planet X, 8 light-years away from Earth.

Therefore Bob reaches his destination 10 years later (as measured in the Earth's frame).

After a short stay, he returns to Earth, again at 80% of the speed of light.

Consequently Alice is 40 years old when she sees Bob again.

How old is Bob?

Using the time dilation formula, one can show that Bob is only 32 years old.

Let us see how to reach the same conclusion by using Minkowski geometry.

We start by picking an inertial frame.

The simplest choice is the Earth's frame.

Since in this frame the Earth is at rest, its history is a vertical line, say the t-axis ($x = 0$).

The history of Planet X, which is also at rest in this frame, is another vertical line, say the line $x = 8$ (using years and light-years as units).

Suppose that Bob departs from Earth at $t = 0$, corresponding to the event O with coordinates $(0, 0)$.

Since it takes Bob 10 years to reach Planet X (in the Earth's frame), his arrival is the event P with coordinates $(10, 8)$.

Finally, the twins' reunion is the event Q with coordinates $(20, 0)$ (Figure 2.4).
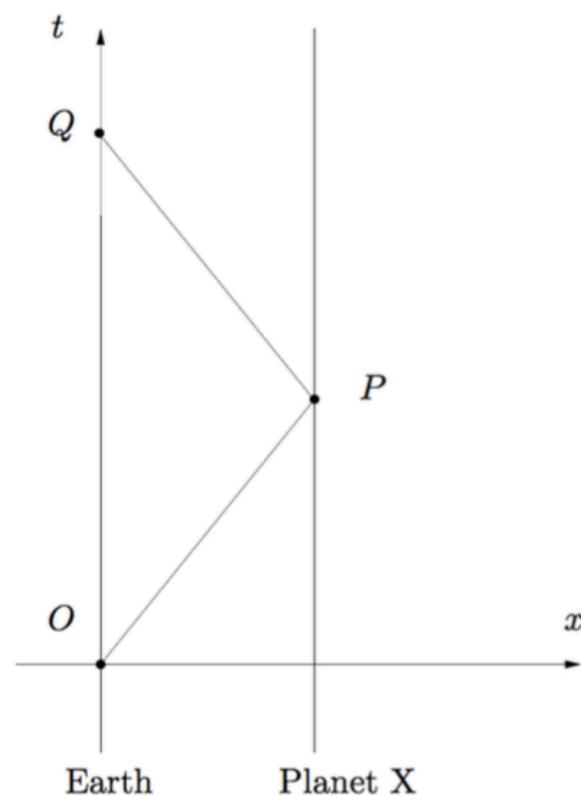

Figure: 2.3: Space-time diagram for the twin paradox.

The time interval measured by Bob on the first leg of the journey is then the interval $\overline{OP}$ between events O and P, given by

$$\overline{OP}^2 = 10^2 - 8^2 = 100 - 64 = 36 \qquad (2.11)$$

that is, $\overline{OP} = 6$.

The time interval measured by Bob in the return leg of the journey is the interval $\overline{PQ}$ between events P and Q, given by

$$\overline{PQ}^2 = (20 - 10)^2 - (0 - 8)^2 = 10^2 - 8^2 = 100 - 64 = 36 \quad (2.12)$$

that is, $\overline{PQ} = 6$.

Therefore the total journey takes Bob $\overline{OP} + \overline{PQ} = 6 + 6 = 12$ years.

The fact that Bob is younger at the twins' reunion can be geometrically reformulated as the statement that

$$\overline{OQ} > \overline{OP} + \overline{PQ} \qquad (2.13)$$

that is: the interval corresponding to the side OQ of the triangle OP Q is bigger than the sum of the intervals corresponding to the other two sides.

This is exactly the opposite of what happens in Euclidean geometry, where the length of one side of a triangle is always smaller than the sum of the other two (triangle inequality).

The triangle inequality can be used to show that the minimum length curve between two points in the Euclidean plane is a line segment.

Indeed, any curve connecting two points in the plane can be approximated by a broken line joining intermediate points along the curve (Figure 2.5).
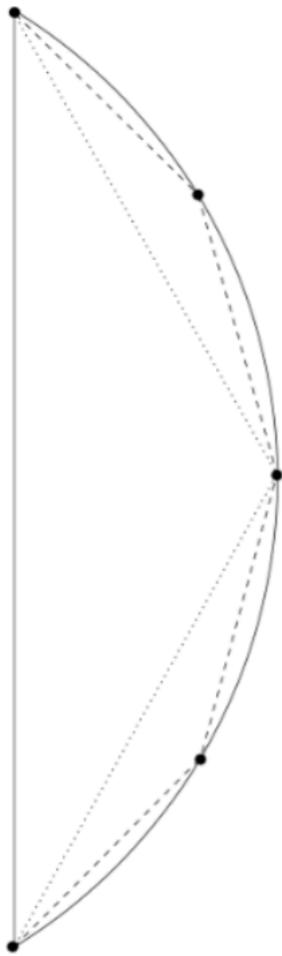
Figure: 2.4: The minimum length curve between two points in the Euclidean plane is a line segment. The same diagram can be used to show that the maximum length causal curve between two events is a line segment

By repeatedly applying the triangle inequality it is clear that the broken line's length is bigger than the length of the line segment joining the two points.

Since we can make the broken line's length as close to the curve's length as we like (by increasing the number of intermediate points), we conclude that the curve's length is necessarily bigger than the length of the line segment.

2.5 More dimensions

For simplicity's sake, we have so far considered space-time diagrams with two dimensions only (coordinates (t, x)).

However, a complete space-time diagram has four dimensions, corresponding to the coordinates (t, x, y, z) of events in an inertial frame.

In this case, the interval between causally related events is

$$\Delta \tau^2 = \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2 \qquad (2.14)$$

Unfortunately it is not easy to visualize a 4-dimensional space.

For this reason we will use space-time diagrams with at most 3 dimensions, corresponding to the coordinates (t, x, y) in some inertial frame.

This allows us to consider particles and light signals which move in the (x, y)-plane.

The interval between two causally related events is in this case

$$\Delta \tau^2 = \Delta t^2 - \Delta x^2 - \Delta y^2 \qquad (2.15)$$

Notice that in 3 dimensions the set of all points with zero interval with respect to a given event O (that is, the set of all events along light signals through O) is now a cone, called the *light cone* of O (Figure 2.5).
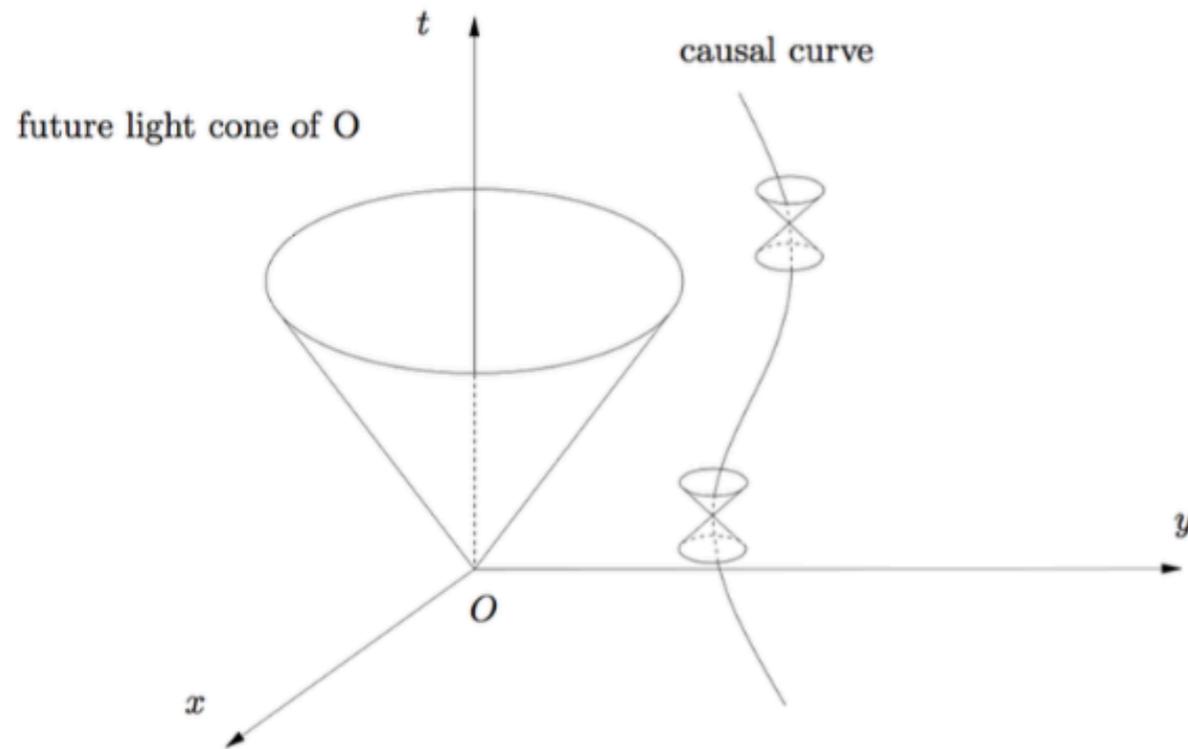

Figure: 2.5: Space-time diagram in 3 dimensions.

Correspondingly, the condition for a curve to be causal is now that it be *inside* the light cone of each of its points.

3. Non-Euclidean geometry

3.1 Curvilinear coordinates

Besides the usual Cartesian coordinates, there are many other possible choices of coordinates in the plane (traditionally called *curvilinear* coordinates).

An example that naturally occurs in many situations (e.g. radars) are the so-called polar coordinates (r, θ), in which each point is identified by its distance r to the origin and by the angle θ between its position vector and the x-axis (Figures 3.1 and 3.2).
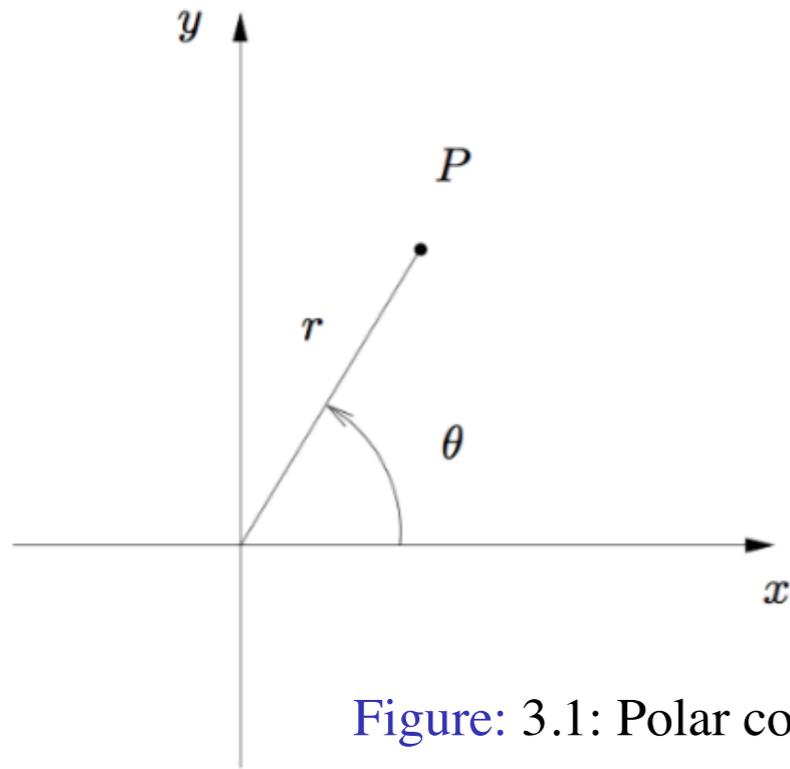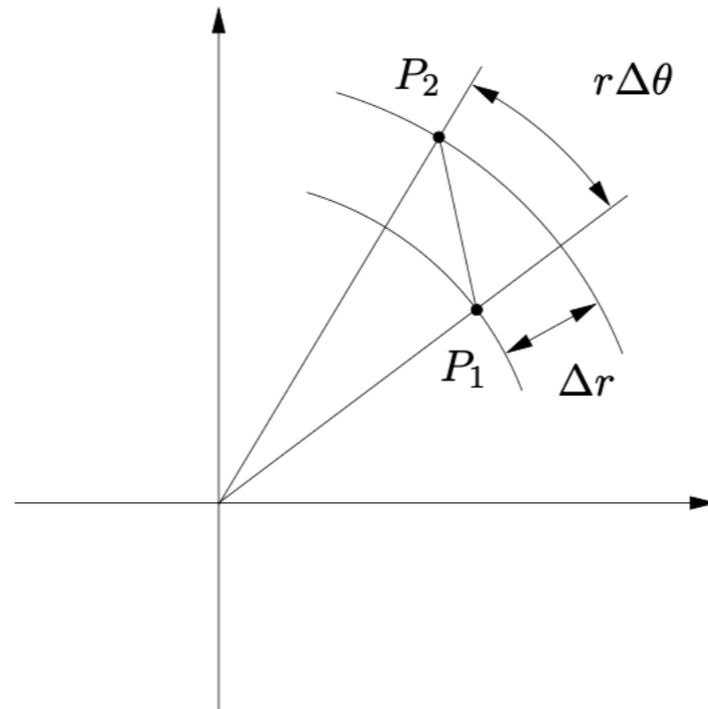

Figure: 3.1: Polar coordinates.

Figure 3.2: Distance in polar coordinates.

Similarly, fixing r and varying θ by $\Delta\theta$ corresponds to traveling a distance $r\,\Delta\theta$ along the circle with center at the origin and radius r.

Since the ray is perpendicular to the circle, the Pythagorean theorem tells us that if two points are very close (so that the bit of circle is almost straight) then the distance between them is approximately

$$\Delta s^2 = \Delta r^2 + r^2 \Delta\theta^2 \qquad\qquad (3.2)$$

(with increasing accuracy as the points become closer).

We can use this formula to compute the length of a curve in polar coordinates approximating it by a broken line.

The distance between two points which are not close can then be computed by computing the length of the line segment(s) connecting them.

3.2 The sphere

Euclidean geometry describes the geometry of the plane.

However, it is often useful to understand the geometry of other surfaces.

For instance, to plan a long sea or air voyage it is necessary to understand the geometry of the sphere.

Just as in the plane, the first thing to do is to choose a system of coordinates.

To do so, we fix a *great circle* which we call the equator.

Remember that a great circle is simply the intersection of the sphere with a plane through its center.

In the case of the Earth's surface, the natural choice for equator is the great circle defined by the plane perpendicular to the rotation axis.

The circles obtained by intersecting the sphere with planes parallel to the equator's plane are called parallels.

Note that the parallels are not great circles.

The points of intersection of the sphere with the line through the center of the sphere which is perpendicular to the equator's plane are called the poles.

The great circle arcs between the poles are called the meridians.

Among these we choose one which we call the principal meridian.

Unlike the equator, there is no natural choice of principal meridian in the case of the Earth, the convention being to choose the meridian through the Greenwich astronomical observatory (near London).

Any meridian can be identified by the angle φ with the principal meridian.

To this angle we define the meridian's longitude.

Similarly, any parallel can be identified by the angle θ with the equator, measured along any meridian.

To this angle we define the parallel's latitude.

Any point P on the sphere's surface can be specified by indicating the parallel and the meridian to which it belongs (Figure 3.3).
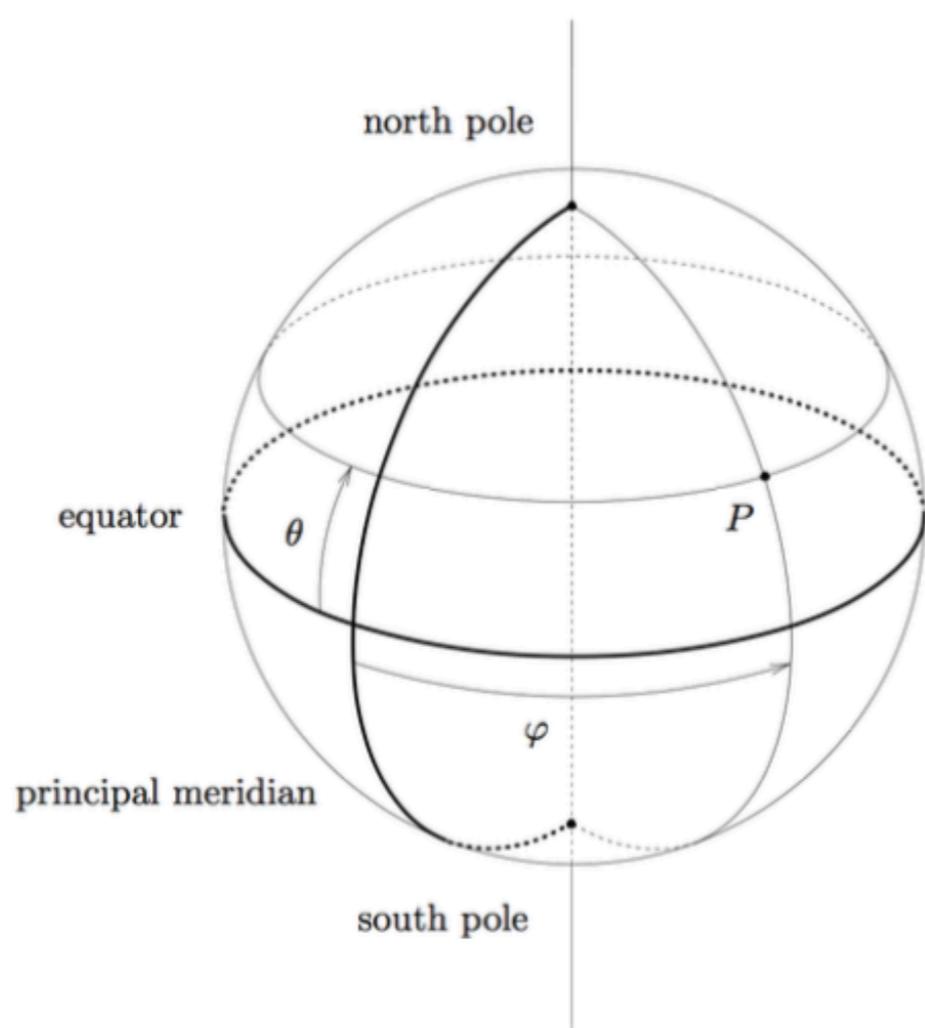
Figure: 3.3: Coordinates on the sphere.

We can then use (θ, φ) as coordinates on the surface of the sphere.

Taking $-\pi/2 \leq \theta \leq \pi/2$ and $-\pi \leq \varphi \leq \pi$, we can then represent the sphere as a rectangle in the plane (Figure 3.4).
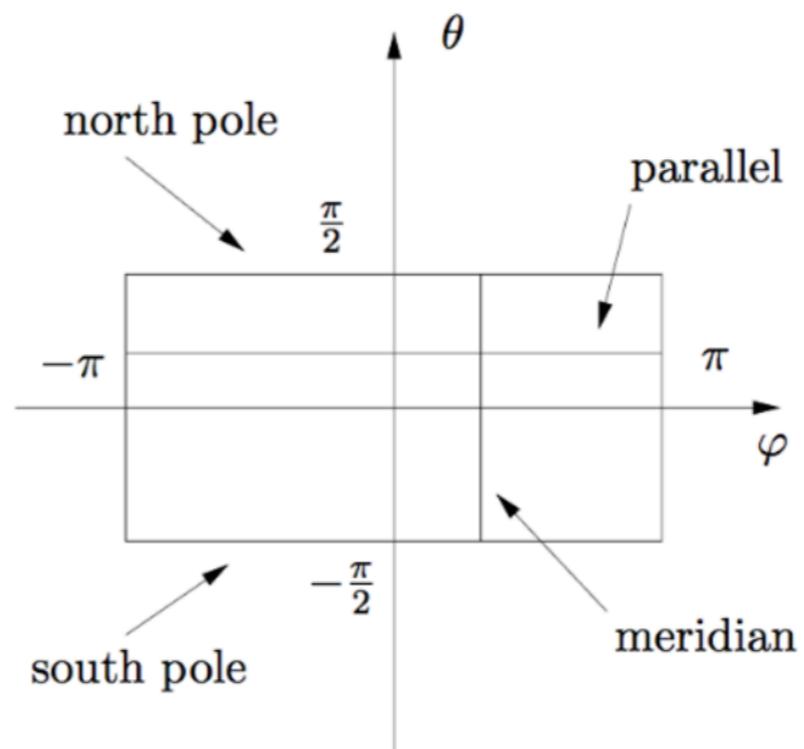
Figure: 3.4: Coordinates on the sphere.

In this representation, the horizontal axis corresponds to the equator, the vertical axis corresponds to the principal meridian and, more generally, horizontal lines represent parallels and vertical lines represent meridians.

We then say that the rectangle is a **map** (or a **chart**) for the sphere.

Since the sphere is very different from a rectangle, it is no big surprise that the map given by the coordinates (θ, φ) is not entirely accurate, representing certain points more than once.

Namely, the lines $\varphi = -\pi$ and $\varphi = \pi$ represent the same meridian, whereas the lines $\theta = \pi$ and $\theta = -\pi$ correspond each to a single point (each of the poles).

Moreover, this map distorts distances: for instance, all parallels are represented by line segments with the same length, when in fact that have different lengths.

More precisely, if R is the radius of the sphere then it is
easy to see that the radius of the parallel of latitude θ is $R\cos\theta$ (Figure 3.5),

Figure: 3.5: Radius of a parallel.
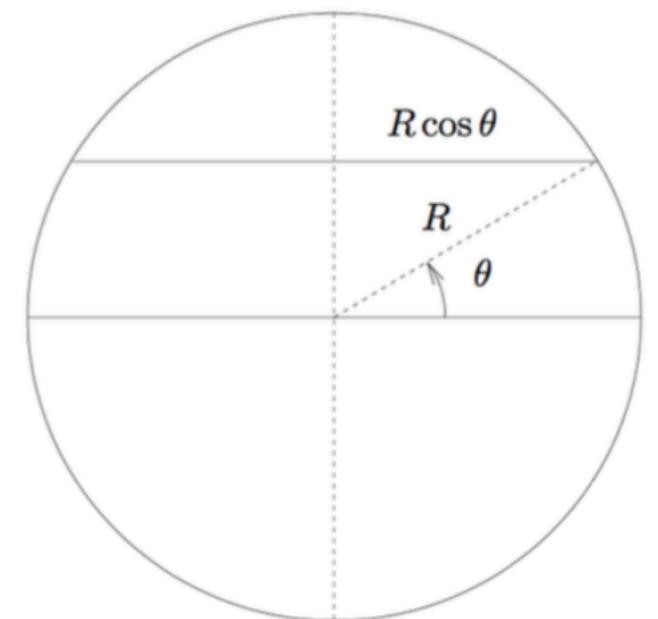
and its length is therefore $2\pi R \cos\theta$.

To compute distances on the sphere we start by observing that fixing φ and varying θ by $\Delta\theta$ corresponds to travelling a distance $R\,\Delta\theta$ along the meridian of longitude φ (Figure 3.6).
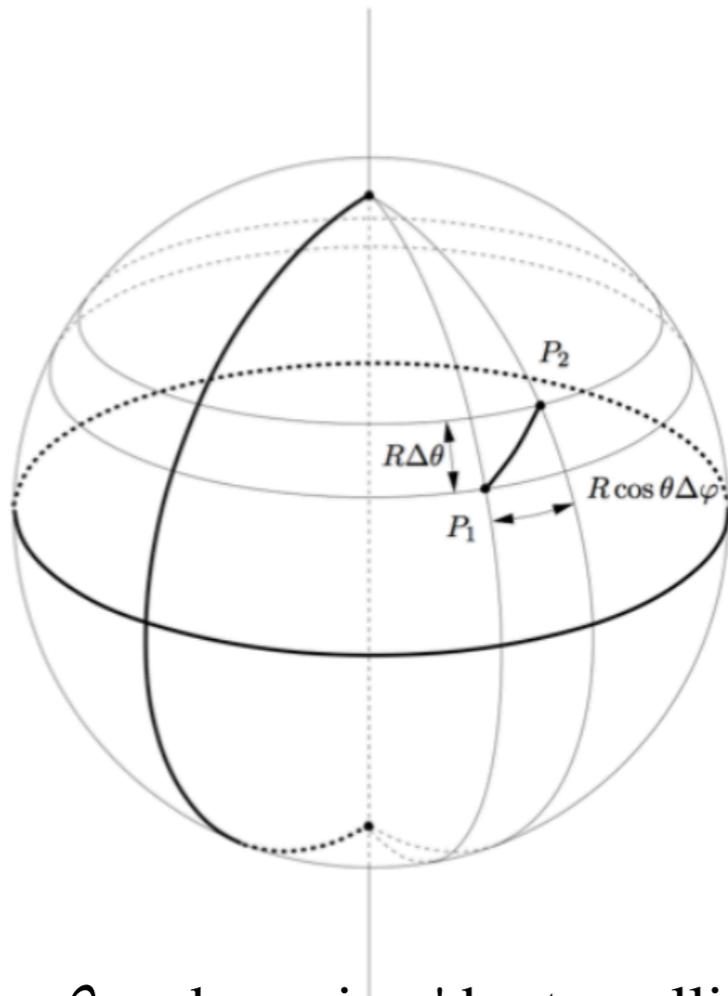

Figure: 3.6: Distance on the sphere.

Similarly, fixing θ and varying ' by travelling a distance $R\cos\Delta\theta$ along the parallel of latitude θ.

Since the meridians are orthogonal to the parallels, the Pythagorean theorem tells us that if the points are very close (so that the curve segments are approximately straight) then the distance between them is approximately

$$\Delta s^2 = +R^2 \Delta\theta^2 + R^2 \cos^2\theta \Delta\varphi^2 \qquad (3.3)$$

(with increasing accuracy as the points become closer).

The expression for the distance Δs^2 between two nearby points on a given surface is called the *metric* of that surface.

For instance, we have already seen that the metric of the plane is written $\Delta s^2 = \Delta x^2 + \Delta y^2$ in Cartesian coordinates, and $\Delta s^2 = \Delta r^2 + r^2\Delta\theta^2$ in polar coordinates.

It is important to stress that although these two expressions are different, they contain the same information: the distance between two nearby points will have the same value whether it is computed in Cartesian or in polar coordinates.

3.3 Geodesics

Now that we have the expression for the metric of a sphere of radius R, we would like to be able to compute the distance between two points on the sphere (not necessarily close).

Recall that in the plane this is done by computing the length of the line segment (that is, the curve with minimum length) joining the two points.

Thus we need to identify the sphere's *geodesics* (that is, the minimum length curves).

These curves will play the *same* role in the geometry of the sphere as straight lines do in the geometry of the plane.

Let P and Q be two points on the sphere.

It is always possible to choose coordinates (θ, φ) such that P and Q are on the meridian $\varphi = 0$.

Consider an arbitrary curve connecting P to Q.

We approximate this curve by a broken line by choosing nearby points

$$P_0 = P, P_1, P_2, ..., P_{N-1}, P_N = Q \qquad (3.4)$$

along the curve.

These points have coordinates

$$(\theta_0, 0), (\theta_1, \varphi_1), (\theta_2, \varphi_2), ..., (\theta_{N-1}, \varphi_{N-1}), (\theta_N, 0), \qquad (3.5)$$

If Δs_i is the distance between P_{i-1} and P_i (with i = 2, 3, ..., N), then the curve's length is

$$\ell = \Delta s_0 + \Delta s_1 + \cdots + \Delta s_N \qquad (3.6)$$

According to the expression for the metric, we have *approximately*

$$\Delta s_i^2 = R^2(\theta_i - \theta_{i-1})^2 + R^2 \cos^2 \theta_i (\varphi_i - \varphi_{i-1})^2 \qquad (3.7)$$

and therefore

$$\Delta s_i \geq R(\theta_i - \theta_{i-1}) \qquad (3.8)$$

We conclude that

$$\ell \geq R(\theta_1 - \theta_0 + \theta_2 - \theta_1 + \cdots + \theta_N - \theta_{N-1}) = R(\theta_N - \theta_0) \quad (3.9)$$

which is exactly the length of the meridian arc between P and Q.

Hence the minimum distance between two points on the sphere is measured along the smaller great circle arc joining them, or, the geodesics of the sphere are its *great circles*.

3.4 Curvature

Having identified the sphere's geodesics, we can now begin to study the sphere's geometry, which is quite different from the geometry of the plane.

For instance, the sum of the internal angles of a triangle is always bigger than π: Figure 3.7 depicts a triangle on the sphere with three right angles.
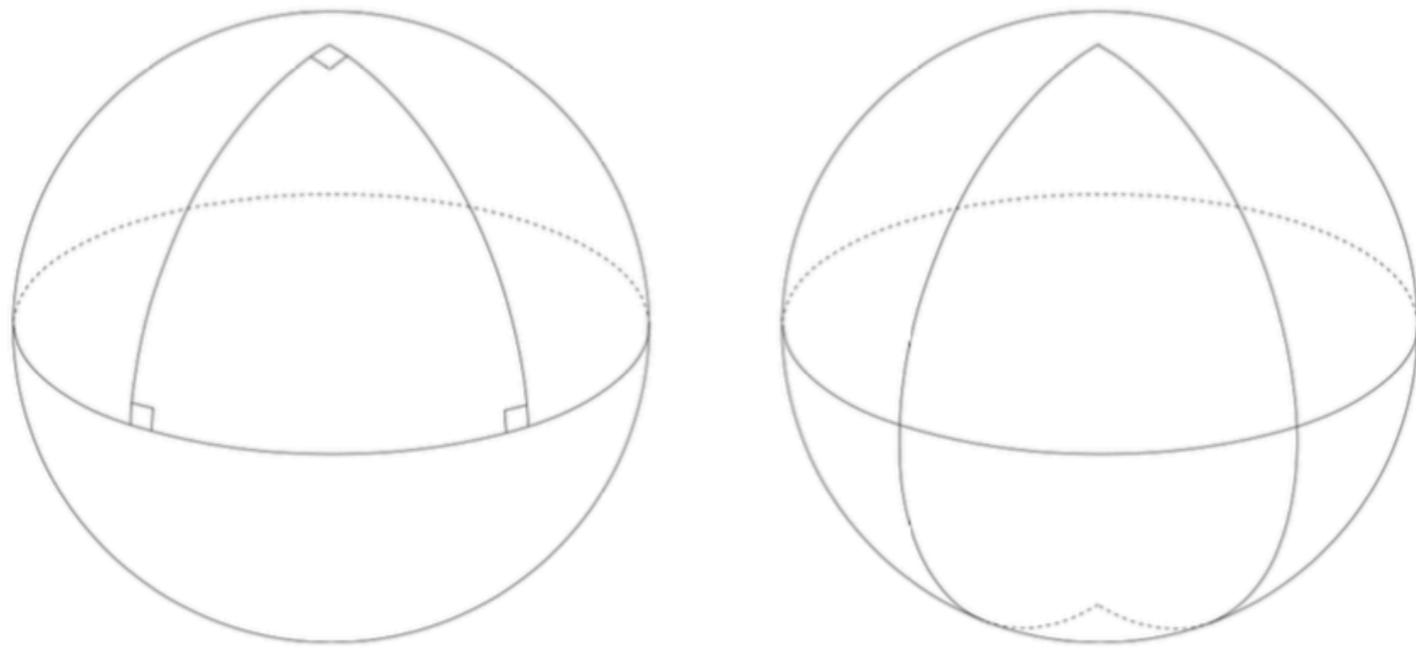
Figure: 3.7: Spherical triangle with three right angles and polygon with two sides.

The difference between the sum of the internal angles of a spherical triangle and π is called the **spherical excess.**

For instance, the spherical excess for the the π triangle in Figure 3.7 is 2 .

The **average curvature** of a spherical triangle is the ratio between the spherical excess and the triangle's area.

Recalling that the area of a sphere of radius R is $4\pi R^2$, and noticing that the complete sphere is formed by 8 triangles like the one in Figure 3.7, we conclude that the average curvature of that triangle is

$$\frac{\frac{\pi}{2}}{\frac{4\pi R^2}{8}} = \frac{1}{R^2} \qquad (3.10)$$

The curvature of a surface at a point is simply the value of the average curvature for a very small triangle around that point.

Hence the curvature at a given point measures how much the local geometry of the surface at that point differs from the geometry of the plane.

The sphere is a constant curvature surface: all triangles have the same average curvature, and so the curvature of the sphere at any point is $1/R^2$.

An interesting consequence of the spherical excess is that there exist polygons on the sphere with only two sides (for instance polygons whose sides are meridians, as depicted in Figure 3.7).

The existence of these polygons is thus a *sign* of the presence of curvature.

3.5 Other maps of the sphere

Just like in the plane, there are many possible choices of coordinates on the sphere.

One possibility is to use the so-called cylindrical projection, which consists on projecting each point P of the sphere to a point Q in the cylindrical surface which is tangent to the sphere at the equator, perpendicularly from the axis (Figure 3.8).
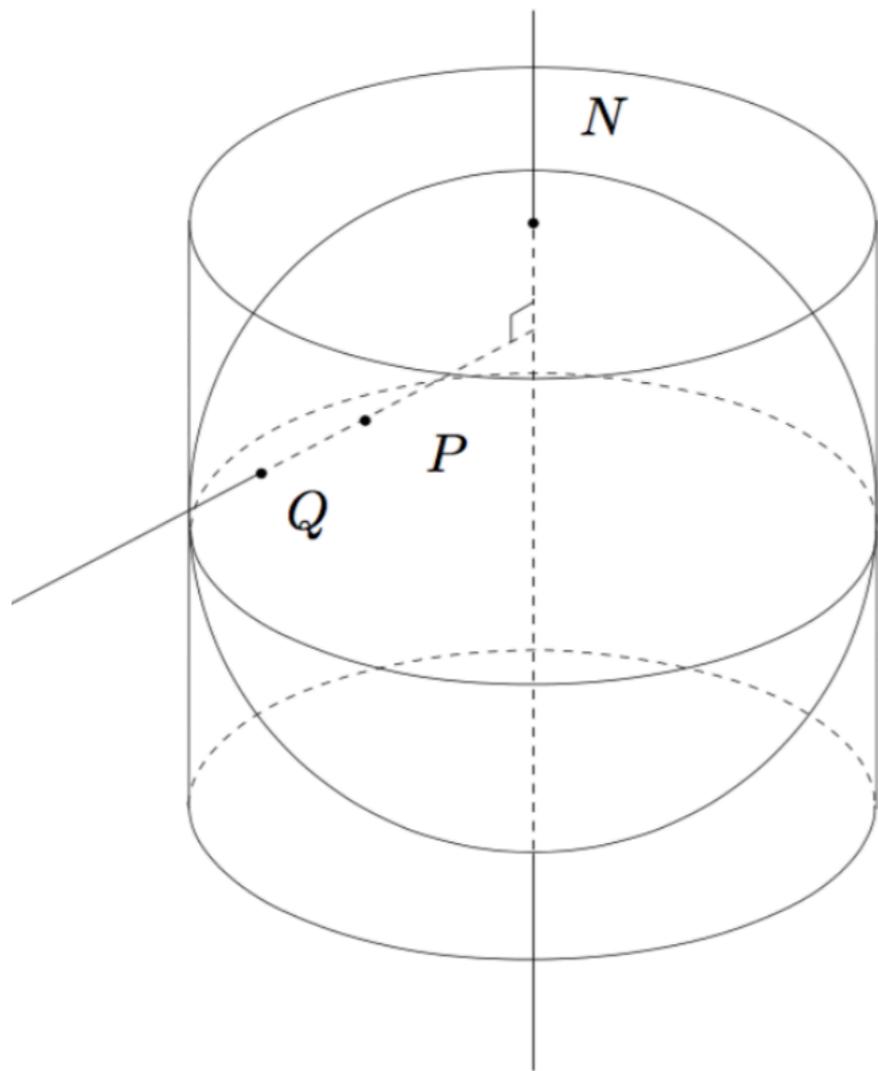
Figure: 3.8: Cylindrical projection

The cylindrical surface can then be unrolled into a rectangle with height 2R and length $2\pi R$, on which we choose Cartesian coordinates (x,y) so that the equator projects to the x-axis, the parallels to horizontal lines and the meridians to vertical lines.

Similarly to what we have done for the coordinates (θ, φ), one can show that the sphere's metric in these coordinates is

$$\Delta s^2 = \left(1 - \frac{y^2}{R^2}\right)\Delta x^2 + \left(1 - \frac{y^2}{R^2}\right)^{-1}\Delta y^2 \qquad (3.11)$$

This expression has the particularity that the coefficients of Δx^2 and Δy^2 are inverse of each other.

One can show that this happens if and only if the projection preserves areas, that is, the area of a given figure on the sphere is equal to the area of its representation on the map.

Another famous projection is the so-called stereographic projection, which projects each point P of the sphere to a point Q on the plane containing the equator from the north pole (Figure 3.9).
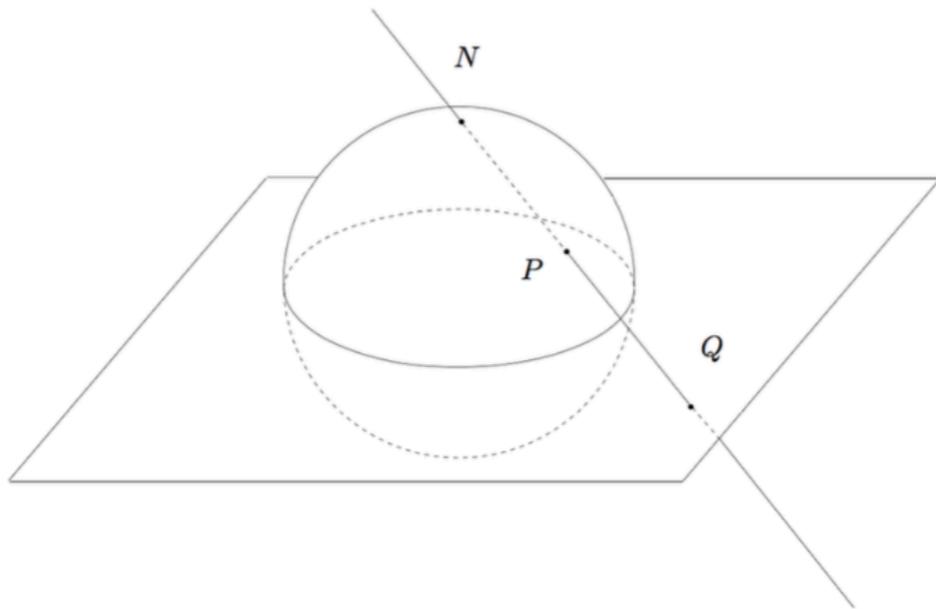

Figure: 3.9: Stereographic projection.

Choosing Cartesian coordinates (x, y) on the plane with origin at the center of the sphere, we see that the equator projects to itself (that is, a circle of radius R and center at the origin), the parallels to concentric circles and the meridians to lines through the origin.

The sphere's metric in these coordinates is

$$\Delta s^2 = \left(1 + \frac{x^2}{R^2} + \frac{y^2}{R^2}\right)^{-2} \Delta x^2 + 4\left(1 + \frac{x^2}{R^2} + \frac{y^2}{R^2}\right)^{-2} \Delta y^2$$

$$(3.12)$$

This expression has the particularity that the coefficients of Δx^2 and Δy^2 are equal.

One can show that this happens if and only if the projection preserves angles, that is, the angle between two intersecting curves on the sphere is equal to the angle between its representations on the map (which is then called a **conformal map**).

The map that uses latitude and longitude as coordinates does not preserve areas nor is conformal.

The most common world maps use the so-called **Mercator projection**, which can be obtained from this map (or the cylindrical projection map) by deforming it vertically until it becomes conformal (this has the advantage that the curves of constant heading are represented by straight lines on the map) (Figure 3.10).
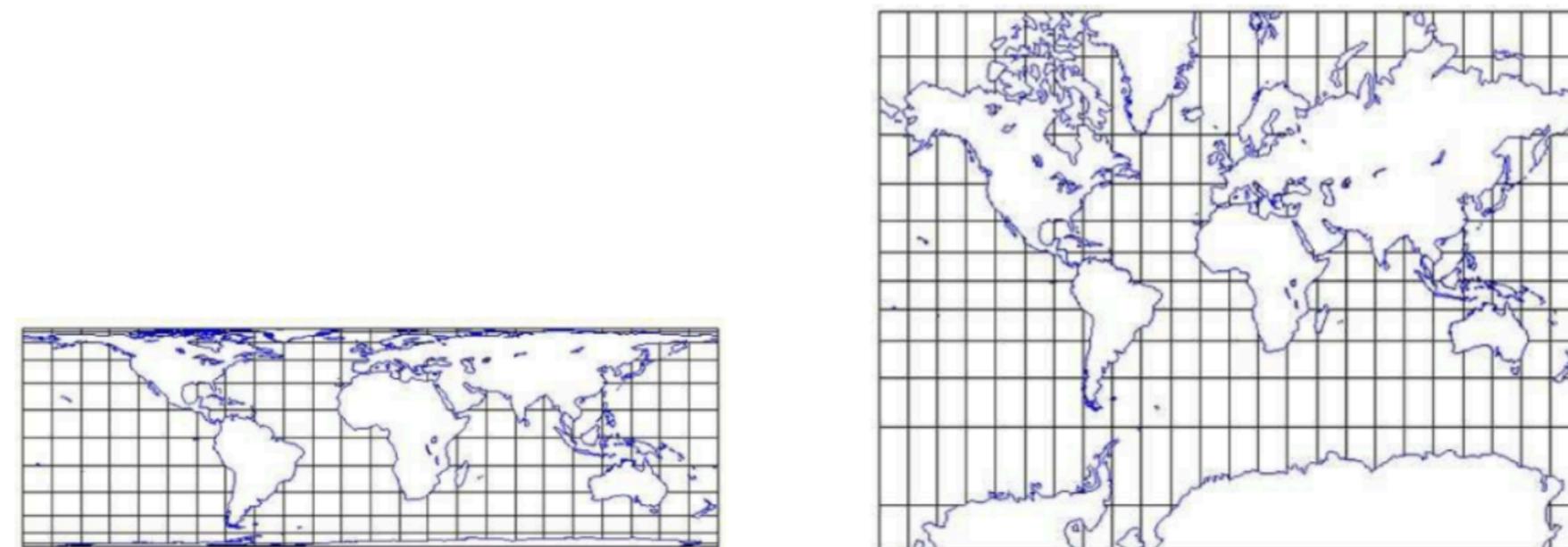
Figure: 3.10: World maps using the cylindrical and Mercator projections.

3.6 Other geometries

The geometry of the sphere is just one example of a non-Euclidean geometry.

For each different surface we will in general have a different geometry, most of which will not have constant curvature.

Many will have negative curvature, that is, the sum of a triangle's internal angles will be smaller than π.

Others will be flat, that is, the sum of a triangle's internal angles will be equal to π.

Examples of the latter are, surprisingly, the cylinder and the cone.

That is why we can roll a sheet of paper into a cylinder or a cone (but not a sphere).

By analogy, we can consider curved spaces with three or more dimensions.

Although it is in general impossible to visualize such spaces, all we need to study their geometry is the expression of the metric in some coordinate system.

The ideas of non-Euclidean geometry were mostly developed by Lobachevsky, Bolyai, Gauss and Riemann, who generalized them to spaces with any number of dimensions.

For this reason, the metric of a surface (or more generally of a curved space with more dimensions) is usually called a **Riemannian metric**.

4. Gravity

4.1 Newton's law of universal gravitation

In Newtonian physics gravity is simply an attractive force between any two bodies along the line that connects them.

If the bodies have masses M and m and are separated by a distance r then the intensity F of the force between them is given by Newton's law of universal gravitation:

$$F = \frac{GMm}{r^2} \tag{4.1}$$

where G is the so-called universal gravitational constant.

Rigorously, this formula holds for point masses only; the force between two bodies with non- negligible dimensions is obtained subdividing them into very small pieces and adding up the contributions of each pair of pieces.

One can show that for a spherically symmetric body the result is the same as if all the mass were placed at the center.

In particular, the gravitational acceleration of a point particle of mass m due to a spherically symmetric body of mass M is

$$g = \frac{F}{m} = \frac{GM}{r^2} \qquad (4.2)$$

where r is the distance to the center of the body.

Note that g does not depend on the value of the mass m; this explains why the gravitational acceleration is the same for all objects.

One can show that the gravitational potential energy of the system formed by the masses M and m is

$$U = -\frac{GMm}{r} \qquad (4.3)$$

(where by convention the potential energy of two masses infinitely far apart is zero).

This energy is negative, since gravity is attractive: we have to spend energy to separate the two masses.

The gravitational potential (gravitational potential energy per unit of mass) due to a spherically symmetric body of mass M is then

$$\phi = -\frac{GM}{r} \qquad (4.4)$$

4.2 Units

To simplify our formulas we will use geometrized units, in which besides $c = 1$ one also has $G = 1$.

In these units masses are measured in meters.

For instance, the mass of the Sun in geometrized units is about 1.5 kilometers, whereas the mass of the Earth is about 4.5 millimeters.

We shall see later that this means that a black hole with the mass of the Sun has a radius of 2 x 1.5 = 3 kilometers, whereas a black hole with the mass of the Earth has a radius of 2 x 4.5 = 9 millimeters.

4.3 Escape velocity

Consider the free-falling motion of a point particle of negligible mass m (for instance the Earth) in the gravitational field of a spherically symmetric body of mass M (for instance the Sun), which we can assume to be fixed at the origin of an inertial frame.

One can show that the particle's mechanical energy (*per unit mass*),

$$E = \frac{1}{2}v^2 - \frac{M}{r} \qquad\qquad (4.5)$$

is conserved along the motion (where v is the particle's velocity).

In particular,

$$\frac{M}{r} \geq -E \qquad (4.6)$$

Consequently, if E < 0 then

$$r \leq -\frac{M}{E} \qquad (4.7)$$

that is, m will never venture more than -M/E away from the center of M.

If m is to move arbitrarily far away from M we must have E ≥ 0.

This means that for m to escape the attraction of M from a point at a distance r from the center of M it must be launched with velocity

$$v \geq \sqrt{\frac{2M}{r}} \qquad (4.8)$$

This minimum velocity is called the **escape velocity**.

4.4 Kepler laws

It is easy to see that the motion of the mass m happens on a plane.

Introducing polar coordinates (r, θ) in that plane, with the origin at the center of the spherically symmetric mass M , we see from the expression of the metric in polar coordinates that the square of the velocity of the particle is

$$v^2 = \left(\frac{\Delta s}{\Delta t}\right)^2 = \left(\frac{\Delta r}{\Delta t}\right)^2 + r^2 \left(\frac{\Delta \theta}{\Delta t}\right)^2 \qquad (4.9)$$

So the mechanical energy is given by

$$E = \frac{1}{2}\left(\frac{\Delta r}{\Delta t}\right)^2 + \frac{r^2}{2}\left(\frac{\Delta \theta}{\Delta t}\right)^2 - \frac{M}{r} \qquad (4.10)$$

Moreover, one can show that the particle's **angular momentum** (per unit mass),

$$L = r^2 \frac{\Delta \theta}{\Delta t} \qquad (4.11)$$

is also conserved alongs its motion.

We can rewrite the conservation laws as

$$\frac{\Delta\theta}{\Delta t} = \frac{L}{r^2}$$

$$\frac{\Delta r}{\Delta t} = \pm\sqrt{2E + \frac{2M}{r} - \frac{L}{r^2}} \qquad (4.12)$$

These equations are examples of differential equations, and can be used to determine all possible motions of m.

The idea is as follows: suppose that we are interested in the motion with certain values of E and L which passes through the point with coordinates (r_0, θ_0) at time t = 0.

We can use the differential equations above to compute approximate values $r_1 = r_0 + \Delta r_0$ and $\theta_1 = \theta_0 + \Delta\theta_0$ for (r, θ) at time t = Δt, where

$$\Delta\theta_0 = \frac{L}{r_0^2}$$

$$\Delta r_0 = \Delta t\sqrt{2E + \frac{2M}{r_0} - \frac{L}{r_0^2}} \qquad (4.13)$$

(we have assumed $\Delta r/\Delta t \geq 0$)

Repeating this procedure with r_1 and θ_1, we can compute approximate values $r_2 = r_2 + \Delta r_1$ and $\theta_2 = \theta_2 + \Delta\theta_1$ for (r, θ) at time $t = 2\Delta t$, and, in general, approximate values $r_n = r_{n-1} + \Delta r_{n-1}$ and $\theta_n = \theta_{n-1} + \theta_{n-1}$ for (r, θ) at time $t = n\Delta t$.

The approximation can be made increasingly better by decreasing Δt.

The differential equations above were solved exactly by Newton, who discovered that the orbits with negative energy are actually ellipses, with the center of M at one of the foci (first Kepler law).

Moreover, the law of conservation of angular momentum has the geometrical interpretation that the line segment joining the center of M to the point particle m sweeps equal areas in equal times (second Kepler law).

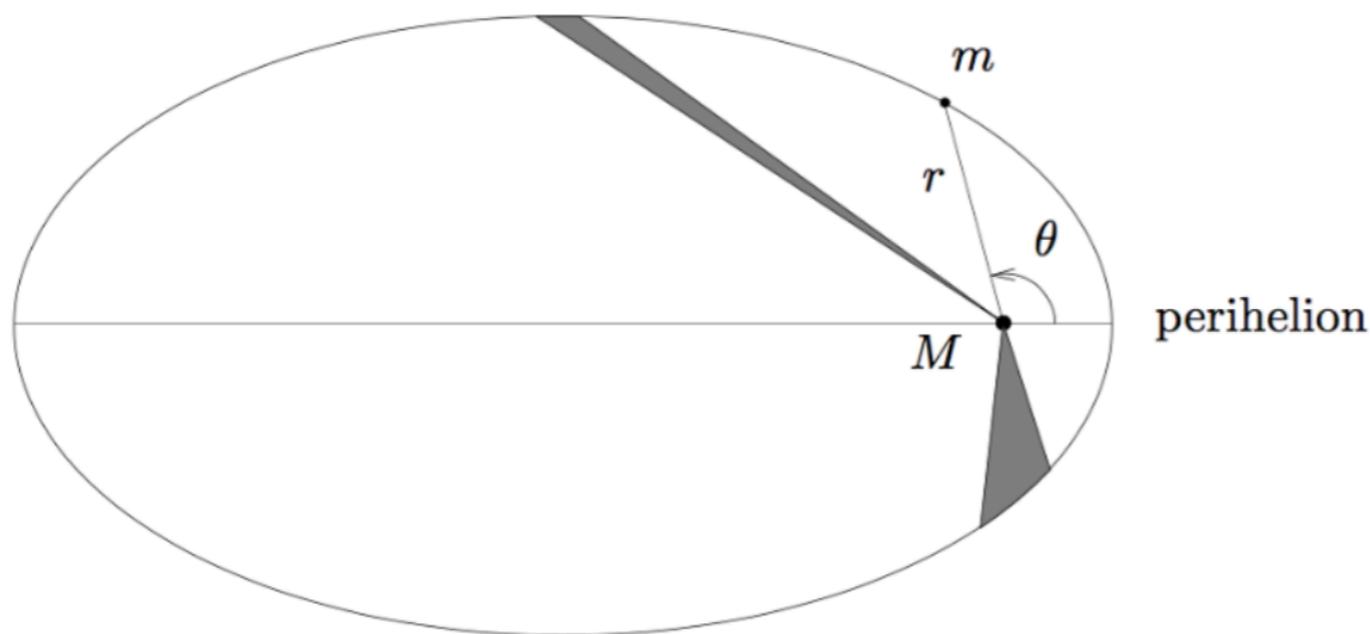In particular, m moves more quickly when it is closer to M (Figure 4.1).

Figure: 4.1: Kepler laws.

Kepler had deduced his laws empirically from astronomical observations of the planets' motions.

Newton showed that these laws were actually mathematical consequences of his universal gravitation law.

4.5 Circular orbits

From the differential equation for r it is clear that we must have

$$2E + \frac{2M}{r} - \frac{L^2}{r^2} \geq 0 \Leftrightarrow 2Er^2 + 2Mr - L^2 \geq 0 \qquad (4.14)$$

If $E < 0$, we see that r can only vary between the two roots of the above polynomial,

$$r_{\pm} = \frac{-M \pm \sqrt{M^2 + 2EL^2}}{2E} \qquad (4.15)$$

corresponding to the two points on the ellipse which are closest and farthest to the center of M (both on the major axis).

The point which is closest to the center of M is called the orbit's **perihelion**.

If r = r+, that is, if

$$E = -\frac{M^2}{2L^2} \qquad (4.16)$$

then the orbit must be a circle of radius

$$r = -\frac{M}{2E} = \frac{L^2}{M} \qquad\qquad (4.17)$$

So we have for circular orbits

$$\left(\frac{\Delta\theta}{\Delta t}\right)^2 = \frac{M}{r^3} \qquad\qquad (4.18)$$

In particular, the orbital speed is given by

$$v = r\frac{\Delta\theta}{\Delta t} = \sqrt{\frac{M}{r}} \qquad\qquad (4.19)$$

Note that the escape velocity at a given point is $\sqrt{2}$ times the speed of the circular orbit through that point.

We can therefore think of the circular orbit speed as a kind of characteristic speed of the gravitational field at that point.

Hence one expects relativistic effects to become important only when this speed becomes comparable to the speed of light.

For familiar gravitational fields (say the Earth's or the Sun's), this speed is much smaller than the speed of light (so M / r << 1).

5. General relativity

5.1 Equivalence principle

In 1907, just two years after publishing the special theory of relativity, Einstein had, in his own words, "the happiest thought of his life".

This idea, which he later called the equivalence principle, was simply this: for a free-falling observer everything happens as if there was no gravity at all.

To understand what Einstein meant we have to remember that the gravitational acceleration is the same for all bodies, regardless of their mass.

This is illustrated by Galileo's legendary experiment of dropping balls of different weights from the leaning tower of Pisa and watching them reach the ground simultaneously.

Should Galileo have jumped together with the balls he would have seen them floating around him, and could momentarily imagine that he was in a gravity-free environment.

This is exactly what happens to astronauts in orbit.

It is sometimes said that there is no gravity in orbit, but this is obviously wrong: if there was no gravity there would be no forces acting on the spacecraft, and it would just move away from Earth on a straight line with constant velocity according to the law of inertia.

Actually, an orbiting spacecraft is free-falling around the Earth; that is why its crew seem to float inside.

5.2 Gravitational redshift

The first thought experiment to which Einstein applied his equivalence principle was the following: suppose that Einstein, from an intermediate floor on the leaning tower of Pisa, sends a light signal with period T towards Galileo, who is on the top floor.

What is the period T′ measured by Galileo?

To answer this question let us suppose that both Einstein and Galileo jump from the tower as the signal is emitted.

Then they would be in a free-falling frame, thus equivalent to an inertial frame (without gravity).

So Galileo would measure the same period T for the light signal (as he would be at rest with respect to Einstein in this inertial frame).

Let Δz be the vertical distance from Einstein to Galileo, and g the gravitational acceleration.

It takes the light signal a time Δz to get to the top floor (as $c = 1$), and in this time interval Galileo acquires a speed $v = g \Delta z$ with respect to the tower (we are assuming that this speed is much smaller than the speed of light, $v \ll 1$, so that this Newtonian formula is approximately correct; we are also assuming that the light signal's period is much smaller than the time taken by the signal to reach the top floor, $T \ll \Delta z$).

According to the Doppler effect formula for $|v| \ll 1$, we must have approximately

$$T' = (1 + g\Delta z)T \qquad\qquad (5.1)$$

that is, should Galileo have remained on the top floor he would have measured a bigger period for the light signal.

Since the period of visible light increases from blue to red, it is usual to call this effect the gravitational redshift.

Recall that the potential energy of a particle on the Earth's surface is U= mgz, and so the corresponding gravitational potential is $\varphi = gz$.

Another way to write the gravitational redshift formula is then

$$T' = (1 + \Delta\phi)T \qquad\qquad (5.2)$$

As long as $|\Delta\varphi| \ll 1$, this formula is still valid even when the field is not uniform (one just has to introduce sufficiently many intermediate observers, for whom the field is approximately uniform).

Since the two observers are at rest with respect to each other, this formula can only be interpreted as meaning that time flows at different rates for observers in different points of a gravitational field: observers in lower positions measure smaller time intervals.

5.3 Curved space-time

Figure 5.1 clearly shows that the gravitational redshift is not compatible with Minkowski geometry
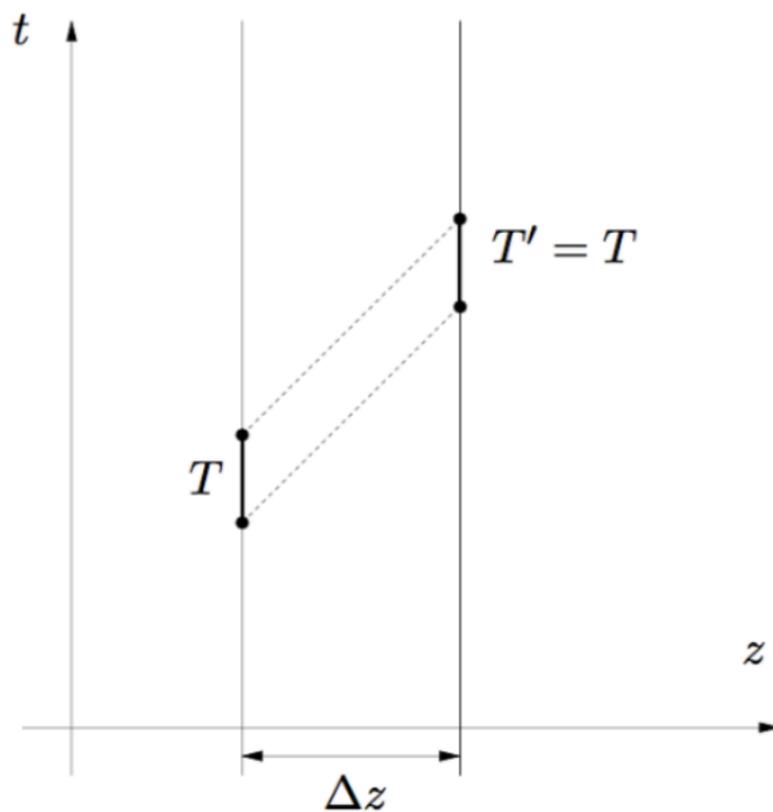

Figure: 5.1: Minkowski geometry is incompatible with the gravitational redshift

according to which any two observers at rest on a given inertial frame must measure the same period for any light signal.

To account for the gravitational redshift we need a space-time diagram which distorts intervals, just like any map of the sphere distorts distances.

This led Einstein to conjecture that the existence of a gravitational field would be *equivalent* to space-time being *curved*.

A small region of a curved surface (say a sphere) looks approximately flat, the approximation becoming better as the region becomes smaller: for example, the sum of a triangle's internal angles tends to π as the triangle becomes smaller and smaller, since the difference is given by the curvature multiplied by the triangle's area.

In the same way, thought Einstein, a small region of a curved space-time must look approximately as flat Minkowski space-time.

But this is exactly the equivalence principle: in a small region of a free-falling frame everything happens as if we were in an inertial frame (the region must be small so that the gravitational field is approximately uniform across that region; otherwise not all bodies would fall with the same acceleration and it would be possible to infer the existence of a gravitational field).

Moreover, any free-falling body in one of these free-falling frames will move on a straight line with constant velocity, and hence, as seen earlier, its history will be a maximum length causal curve.

We conclude that in a curved space-time the histories of free-falling bodies must be maximum length causal curves, that is, geodesics.

In the same way, in a small free-falling frame the history of a light ray must be a causal curve of zero length with the property that it is the unique causal curve joining any two of its points.

We conclude that in a curved space-time the histories of light rays are causal curves of this kind, called *null geodesics*.

These ideas were already clear for Einstein in 1912.

To complete his theory of general relativity he now needed to know how to compute the space-time metric determined by a given matter distribution.

For that, Einstein spent the next three years studying the mathematics of curved spaces (with the help of his mathematician friend Marcel Grossmann).

There was no method for (nor any guarantee of) reaching the right result: he literally needed to *guess the answer*.

After many blind alleys and failed attempts, he finally managed to find the correct equation, by the end of 1915 (Einstein was almost scooped by the famous mathematician Hilbert, who arrived at the same equation by a different method almost at the same time).

The Einstein equation is

$$\mathbf{R} = 8\pi\mathbf{T} \tag{5.3}$$

where $\mathbf{R}$ and $\mathbf{T}$ are mathematical objects called tensors (a generalization of vectors).

The tensor **R** is called the Ricci **curvature tensor**, and describes (part of) the space-time curvature at each point.

It can be computed from the variation of the space-time metric coefficients in each direction.

The tensor **T** is called the trace-reversed energy-momentum tensor, and describes the matter distribution.

Thus the Einstein equation specifies how matter curves space-time.

It turns out that **R** and **T** have ten independent components, and so the Einstein equation is in fact a complicated set of ten differential equations for the space-time metric coefficients.

I note at this point in our first pass (no calculus) that in our second pass we will use mathematical and physical reasoning plus lots more mathematics to fully derive all these results.

6. The Schwarzschild solution

6.1 The Schwarzschild solution

In 1916, just a few months after the publication of the Einstein equation, Schwarzschild (who was then on the Russian front) discovered the solution corresponding to the gravitational field of a spherically symmetric body of mass M .

If we restrict ourselves to events on the equatorial plane, the Schwarzschild metric is

$$\Delta\tau^2 = \left(1 - \frac{2M}{r}\right)\Delta t^2 - \left(1 - \frac{2M}{r}\right)^{-1}\Delta r^2 - r^2\Delta\theta^2 \qquad (6.1)$$

(for r > 2M).

How are we to interpret the coordinates (t, r, θ)?

We notice that when M = 0 the Schwarzschild metric becomes

$$\Delta\tau^2 = \Delta t^2 - \Delta r^2 - r^2\Delta\theta^2 \qquad (6.2)$$

which is just the Minkowski metric in polar coordinates, since, as we have seen, the metric of the Euclidean plane is

$$\Delta s^2 = \Delta x^2 + \Delta y^2 = \Delta r^2 + r^2\Delta\theta^2 \qquad (6.3)$$

This makes sense: when M = 0 there is no gravity and therefore space-time must be flat Minkowski space-time.

We can then think of the Schwarzschild coordinates (t, r, θ) as a generalization of the Minkowski coordinates (t, r, θ).

6.2 Stationary observers

The curves along which the coordinates r and θ are constant are the histories of observers at rest with respect to the spherically symmetric body M , which we shall call *stationary observers*.

Note that these curves **are not** geodesics, since these observers are not in free fall (if they were, their coordinate r should decrease).

Actually, it can be shown that these observers measure a gravitational acceleration

$$g = \frac{\frac{M}{r^2}}{\sqrt{1 - \frac{2M}{r}}} \qquad (6.4)$$

which is approximately the Newtonian result for M / r << 1.

When the coordinate t varies by Δt , a stationary observer measures a proper time interval

$$\Delta\tau = \Delta t\sqrt{1 - \frac{2M}{r}} \qquad (6.5)$$

(since $\Delta r = \Delta\theta = 0$).

Note that when r is very big we have $\Delta\tau = \Delta t$.

Therefore we can interpret the coordinate t as the time measured by a stationary observer *far away* from M ("at infinity").

6.3 Redshift

The Schwarzschild metric is invariant under time translations.

This means that the interval between two nearby events P_1 and P_2, with coordinates (t_1, r_1, θ_1) and (t_2, r_2, θ_2), is equal to the interval between the events Q_1 and Q_2 with coordinates $(t_1 + \Omega, r_1, \theta 1)$ and $(t_2 + \Omega, r_2, \theta_2)$, for any $\Omega \in R$.

So if γ is a geodesic, that is, a curve of maximum length, the curve γ obtained from by moving all of its points to points with the coordinate t increased by Ω is also a geodesic.

The same is true if γ is a null geodesic.

Consider two stationary observers O and O′, with radial coordinates r and r′.

Suppose that O sends a light signal with period T towards O′.

The history of the light ray corresponding to the beginning of the period is a null geodesic γ .

The light ray corresponding to the end of the period is another null geodesic γ', constructed from γ translating by Ω in the t coordinate (Figure 6.1).
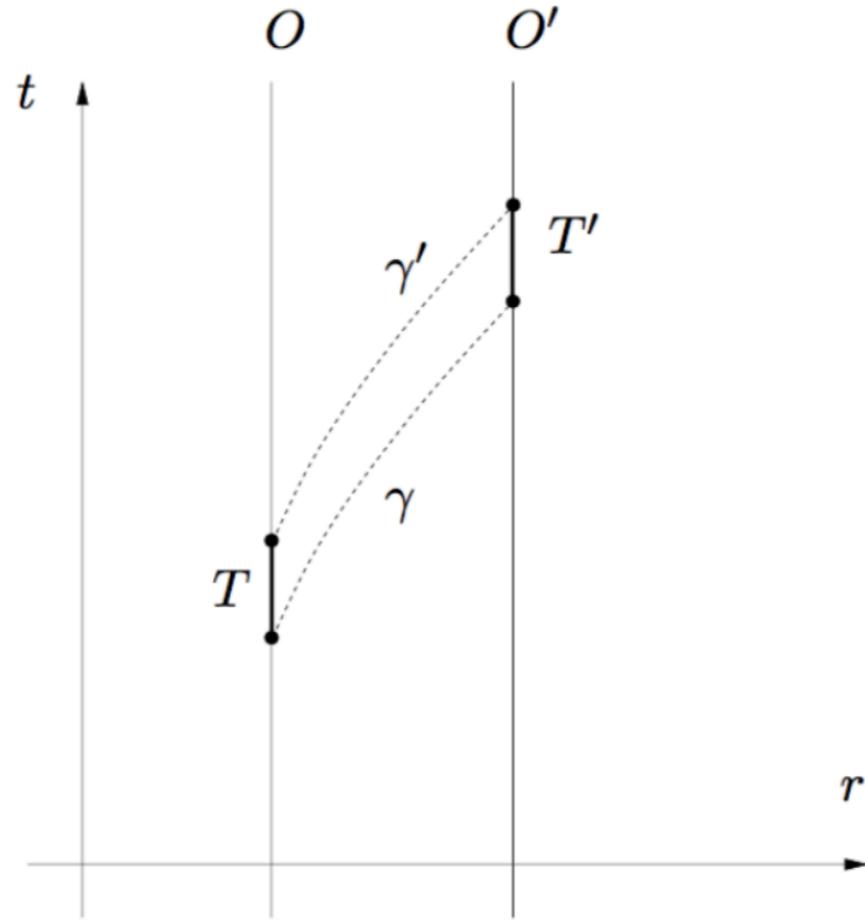

Figure: 6.1: Redshift in the Schwarzschild geometry.

The period T measured by O is related to Ω by

$$T = \Omega\sqrt{1 - \frac{2M}{r}}$$

(6.6)

In the same way, the period T′ measured by O′ satisfies

$$T' = \Omega\sqrt{1 - \frac{2M}{r'}}$$

(6.7)

We conclude that

$$T' = T\sqrt{\frac{1 - \frac{2M}{r'}}{1 - \frac{2M}{r}}}. \qquad (6.8)$$

This is the redshift formula for stationary observers in the Schwarzschild metric.

It reduces to the approximate formula from earlier for M / r , M' / r' << 1.

Note that if r' > r then T' > T , as one would expect.

This statement is sometimes summarized by saying that **gravity delays time**.

6.4 Space curvature

Suppose that two nearby stationary observers, with spatial coordinates (r, θ) and (r+Δr, θ +$\Delta\theta$), want to measure the distance between them.

To do so they just have to measure the time taken by a light signal to travel from one to the other (c = 1).

A light signal propagating between the two observers takes a coordinate time interval Δt such that

$$\left(1 - \frac{2M}{r}\right)\Delta t^2 - \left(1 - \frac{2M}{r}\right)^{-1}\Delta r^2 - r^2\Delta\theta^2 = 0 \qquad (6.9)$$

However, the stationary observers measure a proper time interval

$$\Delta\tau = \Delta t\sqrt{1 - \frac{2M}{r}} \qquad (6.10)$$

and so measures a distance

$$\Delta s = \Delta\tau = \Delta t\sqrt{1 - \frac{2M}{r}} = \sqrt{\left(1 - \frac{2M}{r}\right)^{-1}\Delta r^2 + r^2\Delta\theta^2} \qquad (6.11)$$

that is,

$$\Delta s^2 = \left(1 - \frac{2M}{r}\right)^{-1}\Delta r^2 + r^2\Delta\theta^2 \qquad (6.12)$$

This is not the metric of an Euclidean plane, since the coefficient of Δr^2 is not 1 (although it does approach 1 for large r).

In fact, it can be shown to be the metric of the surface shown as an embedding diagram in Figure 6.2.
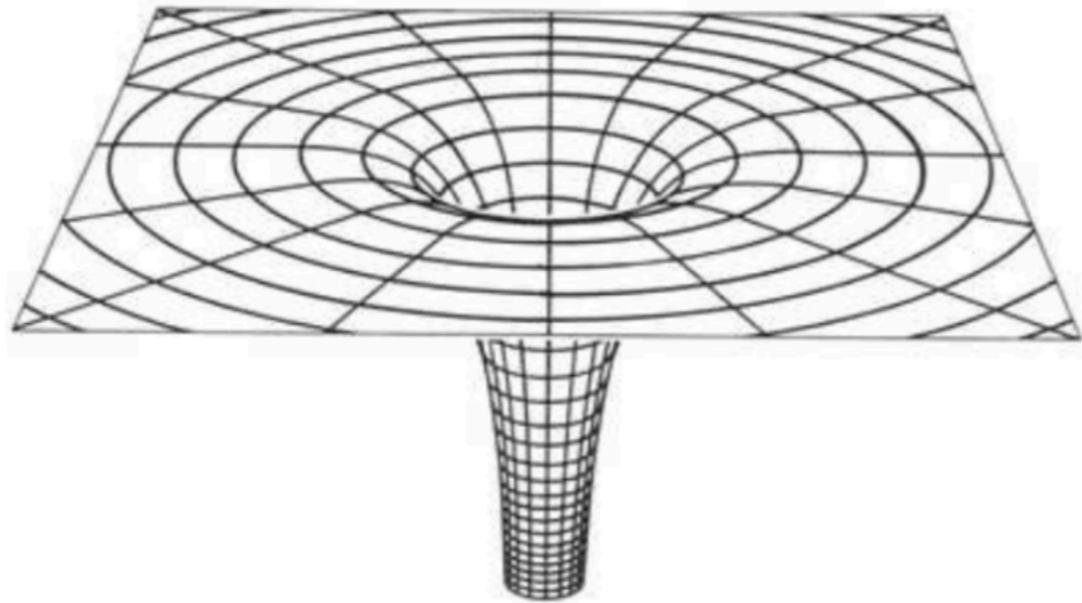
Figure: 6.2: Surface with the metric measured by the stationary observers.

In other words, stationary observers on the equatorial plane deduce from their distance measurements that they are actually living on a curved surface.

This statement is sometimes summarized by saying that **gravity curves space**.

6.5 Orbits

To compute the histories of free-falling particles we have to compute the geodesics of the Schwarzschild metric.

One can show that these satisfy the following differential equations:

$$\frac{\Delta\theta}{\Delta\tau} = \frac{L}{r^2}$$

$$\frac{\Delta r}{\Delta\tau} = \pm\sqrt{2E + \frac{2M}{r} - \left(1 - \frac{2M}{r}\right)\frac{L^2}{r^2}} \qquad (6.13)$$

$$\frac{\Delta t}{\Delta\tau} = \left(1 - \frac{2M}{r}\right)^{-1}\sqrt{1 + 2E}$$

Note that the first two equations are almost identical to the Newtonian equations, with the proper time τ playing the role of the Newtonian time.

This means that if the relativistic effects are small ($M / r \ll 1$), then the trajectories of free-falling particles in the Schwarzschild metric are *approximately* ellipses.

However, they are not exactly ellipses: their perihelion precesses (i.e. rotates), at a rate of about

$$\frac{6\pi M}{r} \qquad (6.14)$$

radians per orbit (for almost circular orbits of radius r).

Actually, the perihelia of the orbits of all planets in the Solar System precess for other reasons.

This is not surprising, since besides the Sun's gravitational force, which is by far the main influence, the planets also feel the gravitational attraction of each other.

When Einstein discovered the general theory of relativity, Newton's theory could explain the precession of the perihelia of all planets except one: Mercury.

The precession of Mercury's perihelion is about 5600 arcseconds per century; Newton's theory predicted 5557.

There was thus an unexplained discrepancy of 43 arcseconds per century.

When Einstein computed the precession of Mercury's perihelion due to general relativistic effects he exactly obtained, to his great satisfaction, the missing 43 arcseconds per century.

By similar methods to those used in the Newtonian theory, one can show that there are circular orbits of any radius $r > 3M$.

By coincidence, the Newtonian result

$$\left(\frac{\Delta\theta}{\Delta t}\right)^2 = \frac{M}{r^3} \qquad (6.15)$$

still holds.

6.6 Light rays

The histories of light rays correspond to the null geodesics of the Schwarzschild metric, which satisfy the following differential equations:

$$\frac{\Delta\theta}{\Delta\lambda} = \frac{L}{r^2}$$

$$\frac{\Delta r}{\Delta\lambda} = \pm\sqrt{2E - \frac{2M}{r} - \left(1 - \frac{2M}{r}\right)\frac{L^2}{r^2}} \qquad (6.13)$$

$$\frac{\Delta t}{\Delta\lambda} = \left(1 - \frac{2M}{r}\right)^{-1}\sqrt{1 + 2E}$$

The parameter λ plays the same role as the proper time on a geodesic (recall that null geodesics have zero length).

These equations can be obtained from the geodesic equations in the limit $E \gg 1$.

This is what one would expect, as particles with very high energies move very close to the speed of light.

It is however clear that these equations are not the equations for a null geodesic in Minkowski's space-time, which can be obtained by setting $M = 0$.

The fact is that M > 0 makes $\Delta t / \Delta \lambda$ bigger than what it would be in Minkowski's spacetime.

Therefore a light ray takes a longer coordinate time t to travel a given path in the Schwarzschild metric.

This is the so-called Shapiro effect, and has been measured by radar experiments in the Solar System starting in 1966.

On the other hand, M > 0 makes the absolute value of $\Delta r / \Delta \lambda$, and hence $\Delta r / \Delta \theta$ bigger than it would be in Minkowski's space-time.

Consequently the light ray follows a curved trajectory, instead of the straight line it would travel for M = 0.

This is the so-called gravitational lens effect, and was (together with the gravitational redshift and the precession of Mercury's perihelion) one of the three experimental tests of general relativity proposed by Einstein.

The experimental confirmation of this prediction, achieved by an English expedition led by Eddington in 1919, made Einstein a celebrity overnight.

The English astronomers had to travel to the remote locations of the Prince's Island (then a Portuguese colony) and Sobral (Brazil) to photograph a total eclipse of the Sun.

The reason for this is illustrated in Figure 6.3: the presence of a body with a large mass M bends the light rays, shifting the images of objects with respect to their usual positions.
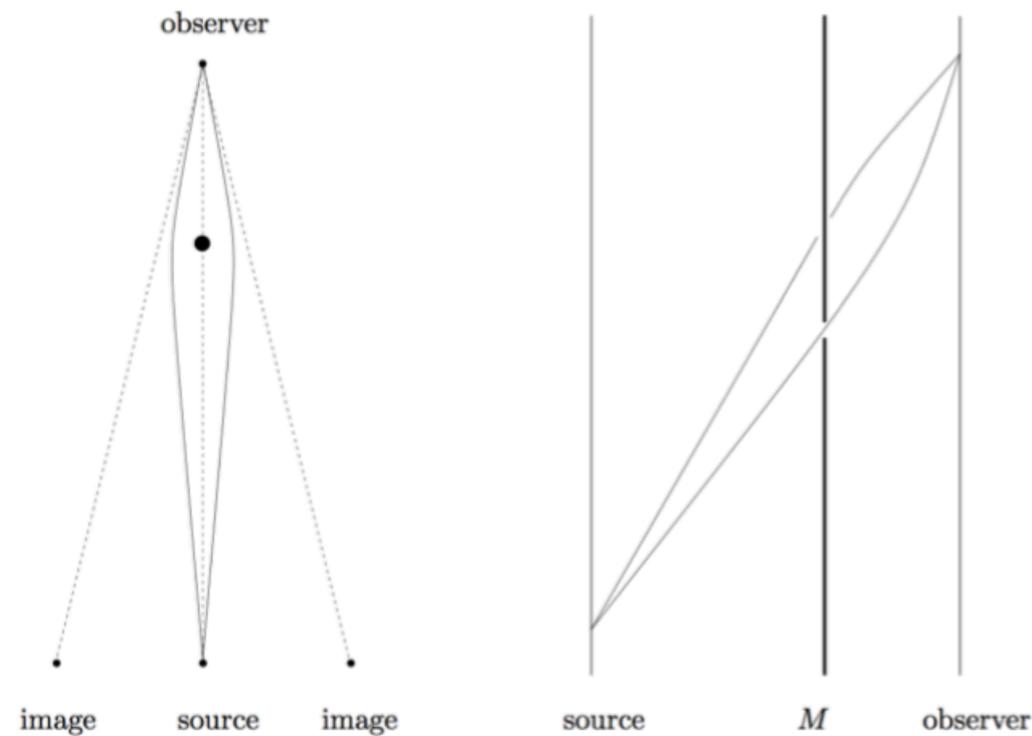

Figure: 6.3: Gravitational lens effect in space and in space-time.

By comparing a picture of the sky around a body of high mass with a picture of the same region at a time when the body is/not there it is then possible to measure the shift in the stars' positions due to the bending of the light rays.

The bending angle computed by Einstein was

$$\frac{4M}{b} \qquad\qquad (6.17)$$

where b is the distance of closest approach of the light ray to the body (for comparison's sake, Newtonian gravity predicts half this value for a particle moving at the speed of light).

This is a very small angle under normal circumstances.

For a light ray just grazing the surface of the Sun (which is by far the body of larger mass in the Solar System) it is only about 1.75 arcseconds.

Hence the need of using the Sun to measure this effect, and consequently of photographing the Sun during a total eclipse (to make the stars visible).

Figure 6.3 illustrates the case in which the mass M is directly between the light source and the observer; in this case she sees two images, on opposite sides of M.

In the space-time diagram this corresponds to the existence of two null geodesics connecting the same events, which, as we saw in earlier, signals the presence of curvature.

A spectacular example of this effect is the so-called Einstein Cross (Figure 6.4).
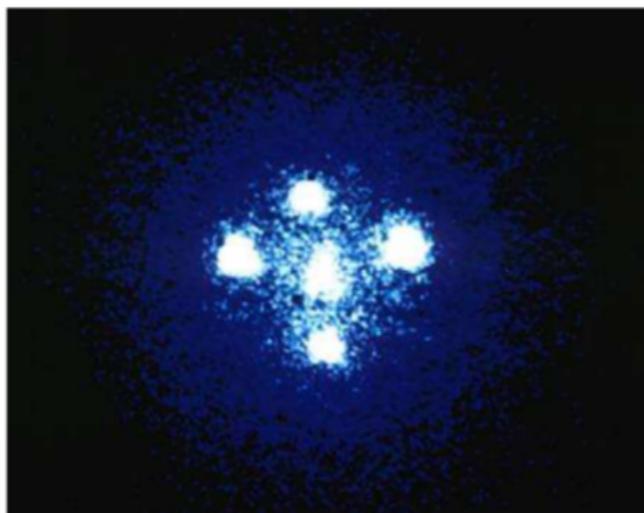

Figure: 6.4: Einstein Cross.

It consists of four images of the same quasar, 8 billion light-years away, surrounding the nucleus of a galaxy "only" half a billion light-years away, which is acting as a gravitational lens.

The reason why the images are known to be of the same quasar is that they show the same exact pattern of brightness variation.

Due to the Shapiro effect, however, the brightness variation of the different images is desynchronized by a few days.

6.7 Black holes

The Schwarzschild metric is not defined for r = 2M (the so-called Schwarzschild radius), since there is a division by zero in the coefficient of Δr^2.

 For a long time it was not clear what to make of this.

In the Newtonian theory, the Schwarzschild radius corresponds to the points where the escape velocity is equal to the speed of light.

In the Schwarzschild metric a similar phenomenon occurs: the escape velocity for a stationary observer approaches the speed of light as r approaches 2M .

Moreover, the redshift and the gravitational field measured by these observers approach infinity as r approaches 2M .

For these reasons it was initially thought that r = 2M was a mathematical singularity, beyond which one could not extend the Schwarzschild space-time.

However, it was noticed that the curvature of this space-time is

$$\frac{M}{r^3} \qquad\qquad (6.18)$$

and so nothing out of the ordinary seems to happen at the Schwarzschild radius.

An analogous situation occurs on the map of the sphere obtained from the cylindrical projection: although the coefficient of Δy^2 is not defined for y = -R (south pole), nothing special happens to the curvature of the sphere as y approaches this value, since it is constant equal to $1 / R^2$.

In this case, it is easy to see that the sphere does not have any singularity at the south pole: the coordinates (x, y) associated to the cylindrical projection just happen to be ill defined at this point.

The problem can be overcome by choosing a different coordinate system (for instance the coordinates (x, y) associated to the stereographic projection).

A coordinate system which is well defined on the region r = 2M was discovered by Painleve in 1921.

Painleve substituted the Schwarzschild coordinate t by a different time coordinate t', in terms of which the metric is written

$$\Delta r^2 = \Delta t'^2 - \left(\Delta r + \sqrt{\frac{2M}{r}}\Delta t'\right)^2 - r^2\Delta\theta^2$$

$$= \left(1 - \frac{2M}{r}\right)\Delta t'^2 - 2\sqrt{\frac{2M}{r}}\Delta t'\Delta r - \Delta r^2 - r^2\Delta\theta^2 \quad (6.19)$$

Note that in these coordinates there is no problem at r = 2M.

However, the coefficient of $\Delta t'^2$ vanishes at the Schwarzschild radius.

This means that the curves on which r = 2M and θ is constant are null geodesics.

No wonder that there cannot be stationary observers at r = 2M (they would have to be moving at the speed of light).

Similarly, there cannot exist stationary observers for r < 2M , since these would correspond to non-causal curves (they would have to be moving faster than the speed of light).

These facts can be geometrically understood by representing the light cones in a space-time diagram (Figure 6.5).
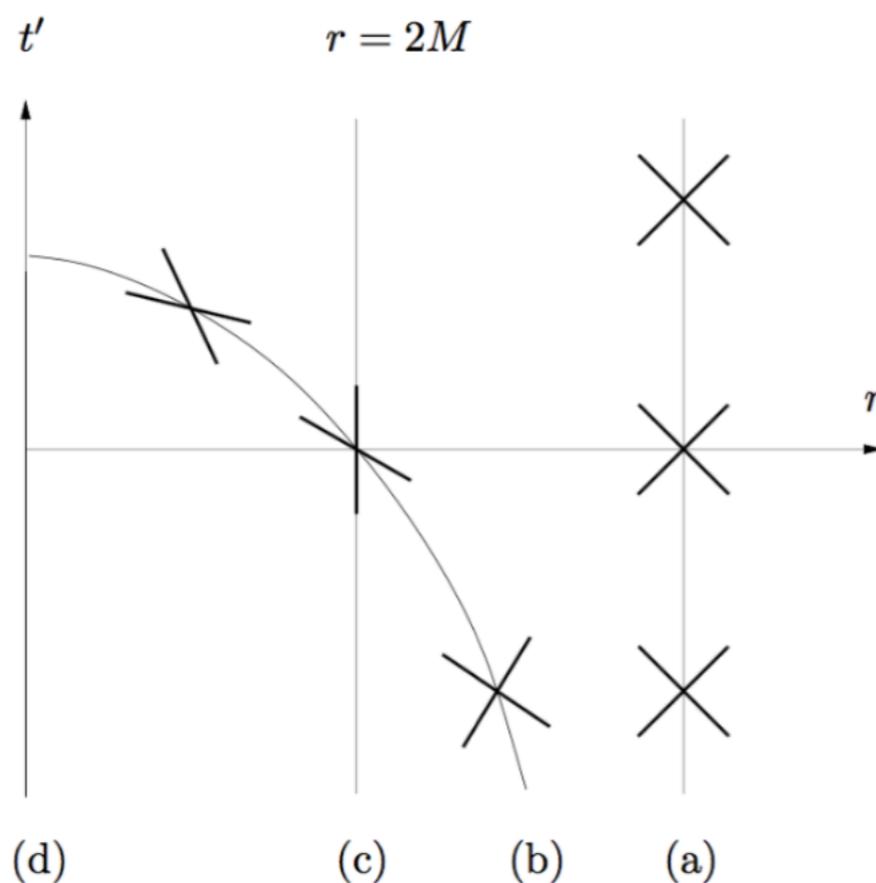
Figure: 6.5: Space-time diagram for the Schwarzschild solution containing the histories of: (a) a stationary observer; (b) a particle falling through the event horizon; (c) the event horizon; (d) the singularity.

These are given at each point of coordinates (t', r) by

$$\Delta \tau^2 = 0 \Leftrightarrow \pm \Delta t' = \Delta r + \sqrt{\frac{2M}{r}} \Delta t' \Leftrightarrow \pm 1 - \sqrt{\frac{2M}{r}} \qquad (6.20)$$

For r >> 2M the light cones approach the Minkowski light cones, with slopes ±1, corresponding to light rays receding from or approaching r = 0.

As r diminishes, however, the light cones tilt towards r = 0.

At the Schwarzschild radius, the light cones are so tilted that the light ray "receding" from r = 0 actually stands still.

For r < 2M the situation is even more dramatic: the light ray "receding" from r = 0 is actually approaching r = 0.

Therefore a particle entering the region r < 2M can never leave; moreover, it is forced to move towards r = 0.

Also, it is clear that observers at infinity can never see any event in the region r ≤ 2M ; this region is then called a **black hole**.

The surface r = 2M , which bounds this region, is called the **event horizon** (since one cannot see beyond it).

Finally, the curve r = 0 is called the **singularity**.

It is indeed a mathematical singularity, beyond where it is not possible to continue the Schwarzschild space-time, since the curvature becomes infinite there.

Physically, the curvature of a space-time measures the so-called tidal force, resulting from the non-uniformity of the gravitational field.

Consider, for instance, an astronaut in orbit, floating upside-down.

Then the Earth's gravitational force on her head is slightly larger than on her feet.

Hence there is a residual stretching force (the tidal force).

At the singularity inside the black hole this force becomes infinite, and thus any object which falls inside a black hole ends up destroyed.

Ordinary spherically symmetric bodies (stars, planets, golf balls) have radii which are much larger than their Schwarzschild radius.

However, it is known that stars which are 3 times more massive than the Sun at the end of their evolutionary process end up forming black holes.

A famous example is the Cygnus X-1 black hole, 9 times more massive than the Sun, which is about 6,000 light-years away.

Besides black holes with masses comparable to that of the Sun, resulting from stellar evolution, it is now known that most galaxies harbor supermassive black holes at their centers.

The black hole at the center of our galaxy, **Sagittarius A***, weighs about 4.3 million solar masses and is about 26,000 light-years away.

Since black holes do not emit light, they cannot be directly observed.

What is in fact observed are X-rays emitted by matter falling in (see Figure 6.6).
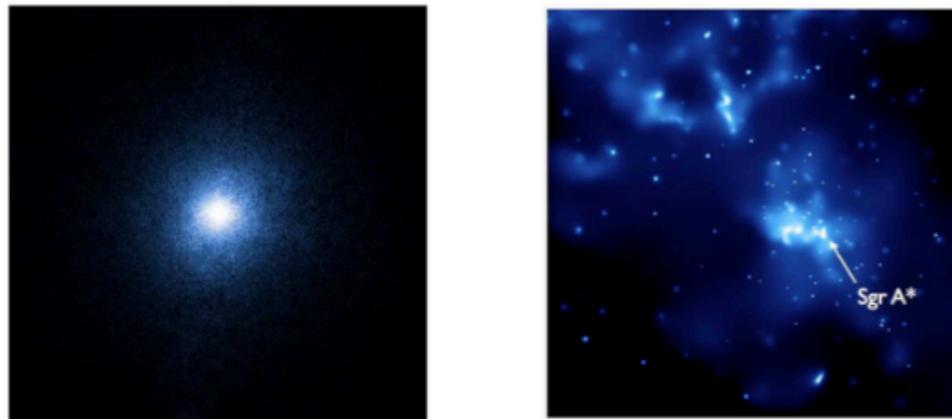

Figure: 6.6: Cygnus X-1 and Sagittarius A*

That was, I hope, a fun, quick tour through a wide range of geometrical ideas that affect physics.

We now go back to the beginning and derive everything rigorously.