Lesson 3 : Geometries of space : flat, spherical, hyperbolic

Notes from Prof. Susskind video lectures publicly available on YouTube

Introduction

We haven't talked much about geometry. We have supposed that space is flat. And we haven't even talked about spacetime geometry at all in this course.

In fact, as a matter of observation, space *is* very flat. But it is not a principle. And it is very important to understand what cosmology would be like if space were not flat.

As we emphasized over and over, it is not so much that we know that space is flat, but we know it is very big. Whether it is flat, positively curved, or negatively curved, we really don't know.

The idea is easy to understand with a two-dimensional world. Imagine we lived in a very big two-dimensional variety, some very big surface. It would be possible that at the same time, around us things looked pretty flat like a plane, and we lived in reality on a huge sphere without being aware of it. Indeed, as you know, until about the sixth century b.c. humans thought the Earth was a huge plane. They even thought it had borders, was supported by a big turtle, etc.

So it is important to investigate the various possible shapes of the universe.

Now we don't know a lot on scales of space much larger than 10 billion light years, or 20 billion light years. So we will make some assumptions. We won't assume that space is flat but, at least for the time being, we will continue to assume that it is isotropic and even homogeneous. Is that true ? Maybe it is not. But we will never know until we find

out its consequences. Therefore we will assume it is homogeneous and investigate consequences. Then we should assume the opposite and see what it says.

Remember the way science always proceeds. From initial observations we make assumptions and build models [1]. If in turn a model leads to predictions contrary to further observations we can make, we discard it. As a consequence, we never know for sure that a hypothesis is right. We are only able to know for sure when a hypothesis is wrong.

Because almost all cosmology is based on that assumption, it is a good thing to explore that space is not flat but homogeneous. Now what does that mean ? Space not being flat means that it is curved.

What kind of space is curved ? Thinking about two-dimensional spaces, that is surfaces, a sphere is curved, a paraboloid is curved, an ellipsoid is curved. An ellipsoid with a bump on it is curved. So all kinds of surfaces that we can think of are curved. In a higher number of dimensions, we can also imagine many curved spaces.

But only a very small number of curved spaces are also homogeneous. Speaking casually, homogeneity means the space is everywhere the same. Somebody located at a particular place in the space looks around, sees everything around, and sees exactly the same thing that somebody

1. It may happen that our initial observations presuppose or make sense themselves only in an implicit general model called a paradigm, which we are not always aware of. If we are lead to contradictions from whatever assumptions we make, then it is our whole paradigm that must be changed. A paradigm may also prevent us from seeing things which are there but cannot be observed with our ways of thinking.

else would see from another other position. Such a space is called homogeneous.

In two dimensions, a paraboloid is certainly not homogeneous. It is more curved near the tip, and less curved far away. A long pointy ellipsoid is different near its poles than near its equator or girth.

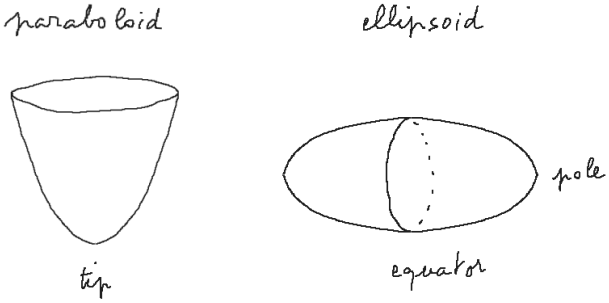

Figure 1 : Curved two-dimensional spaces, not homogeneous.

If we walked on these surfaces, we would notice the difference. With surveying tools, it would be easy to distinguish where it is more curved and where it is less. We would not be fooled, unless of course it was huge and we couldn't detect any curvature.

So ellipsoids, two-dimensional spaces with bumps on them, the surface of the Earth if we take into account the mountains and the valleys, are all curved surfaces. But they are not homogeneous.

What kind of spaces can be curved and homogeneous ? Answer : irrespective of the number of dimensions, there are only three kinds of homogeneous curved spaces. And in each

kind, when they differ it is only by their overall size or scale. The three kinds correspond to flat spaces, positively curved spaces, and negatively curved spaces.

The first kind is flat space. In two dimensions, it is the familiar plane, the top of your desk for instance, or a wall, or any slanting plane. In three dimensions it is the familiar 3D Euclidean space, etc.

Let's be a bit more abstract. To study conveniently the shape of a space, and not rely only on "obvious" geometry, we begin with a metric. We are not talking about spacetime now, just space. The metric of a space is a formula defining and allowing us to compute the distance between any two *neighboring* points.

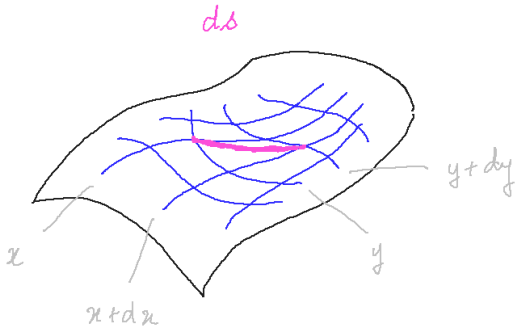

Figure 2 : Most general system of coordinates on a 2-dimensional variety. As usual we show it embedded in 3D, which has the advantage of making it easy to perceive visually the 2D shape, and the drawback of suggesting an intrinsic importance to the embedding 3D space, which it has not.

On a two-dimensional surface, also called a two-dimensional variety, points are labeled with a system of two coordinates. The system can be very general, see figure 2. Axes don't have to be straight in any sense.

Two neighboring points have coordinates (x, y) and $(x + dx, y + dy)$. And the metric has the general form

$$ds^2 = g_{11}\ dx^2 + g_{12}\ dxdy + g_{22}\ dy^2 \tag{1}$$

g is called the metric tensor. Its components g_{ij} depend on which system of coordinates we are using to label all the points on the surface, and they also depend on (x, y). At different points the g_{ij} components of g are different. If we change coordinates, the geometry or shape of the space doesn't change, but the points on it will receive new labels. The actual expression of the metric will change too. We talk of the same tensor g, but with new components. This is analogous to the concept of vector which is an intrinsic geometric object, but has components dependent of the basis used to express it.

We have extensively treated the subject in volume 4 of the collection *The Theoretical Minimum* on general relativity and we assume that the reader has some familiarity with these geometric considerations.

Sometimes, but – remember – not always, with an appropriate choice of coordinate system on the surface, the metric can take the following simple form

$$ds^2 = dx^2 + dy^2 \tag{2}$$

Then of course the space is a plane. In fact, that is the mathematical definition of a plane and of flatness. A plane is a two-dimensional variety where the metric can be expressed by equation (2), or equivalently where the matrix of components of the metric tensor can be the unit matrix.

Notice that on a sphere, whatever system of coordinates we use, there is no way for the metric to have the form given by equation (2). Any change of coordinates will be of no avail. We shall review in a moment the metric on a sphere.

In three dimensions, a flat space is a space whose metric, in an appropriately chosen system of coordinates, has the form

$$ds^2 = dx^2 + dy^2 + dz^2 \tag{3}$$

That is, straightforwardly we add dz^2.

That is the way we describe spaces : by giving the metric tensor, in other words by specifying how to calculate the distance between any two neighboring points. If we know that, we know everything about the shape of the space.

When the metric corresponds to the unitary matrix, we say that the space is flat, or also Euclidean. And the system of coordinates is called Cartesian. But, as the reader should by now clearly be aware of, the space can be flat and the metric actually used not be that corresponding to the unitary matrix.

Let's continue to investigate the flat cases defined by equation (2) or equation (3). And instead of working in Car-

tesian coordinates, let's work in polar coordinates. Polar coordinates have some nice features well adapted to cosmology. They have a center. And if you think of the center as where you are, when you look at the sky you are looking around at angles. Your visual field is a field of angles. In 3D the direction you look at is defined by two angles, but in a plane you are literally looking around you at one angle.

In flat two dimensions, in the polar coordinate system, after having selected a center point O, we introduce two coordinates : an angle θ and a radial variable r.

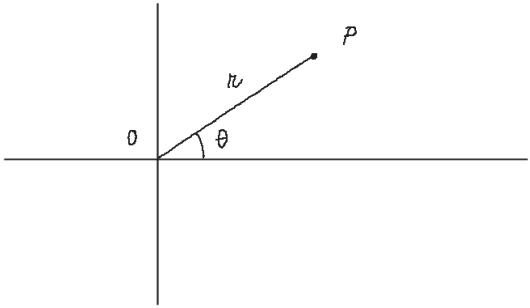

Figure 3 : Flat two-dimensional space with polar coordinates.

Any point P in the plane, instead of being located with two Cartesian coordinates x and y, is located using two polar coordinates r and θ. Of course there is a simple transformation to go from Cartesian coordinates to polar coordinates. We have done this many times in the previous volumes of the collection *The Theoretical Minimum.*

The metric of the space, which in Cartesian coordinates had the form given by equation (2), has now the form

$$ds^2 = dr^2 + r^2\ d\theta^2 \tag{4}$$

It is important to understand that it is the *same metric*, in the sense that the distance between two neighboring points has the same value in both cases – because it is an intrinsic feature of the space –, but the metric has a different expression in each system of coordinates.

The square of the metric ds is the square of the distance between the two points at coordinates $(r,\ \theta)$ and $(r+dr,\ \theta+d\theta)$. It is computed using Pythagoras theorem. It is the sum of the squares of the lengths of the sides of the right triangle shown in figure 4.

The second term on the right-hand side of equation (4) is just a statement that for a given incremental angle $d\theta$, the distance interval between the two points at coordinates $(r,\ \theta)$ and $(r,\ \theta + d\theta)$ gets bigger and bigger as we move away. And it grows linearly with the distance r.

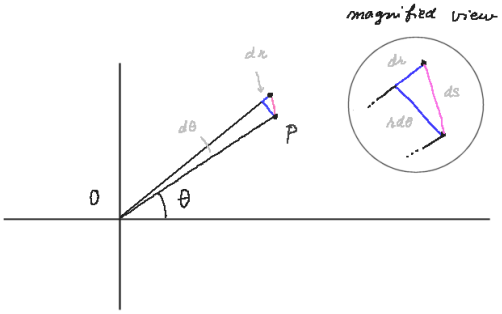

Figure 4 : Metric of the plane in polar coordinates.

That is the metric of the ordinary plane, and we are going to give it another name. We are going to invent a new term for $d\theta^2$.

To understand what we are going to do, let's go back for a moment to one-dimensional spaces, and more specifically to those which close on themselves. In fact $d\theta$ is itself a metric. It is the metric of a circle of radius 1. Think of a circle of radius 1 represented on a plane, and centered at the origin of the plane. Any point on the circle can be located using only one coordinate since the circle is a one dimensional space. The coordinate is for instance simply θ. The square of the distance between two neighboring points on the circle is $d\theta^2$.

Incidentally, a circle is a one-dimensional sphere. Indeed it is a set of points on the plane at a fixed distance of a center.

Notice, however, that to call it a circle or a one-dimensional sphere we have to view it in at least two dimensions. We have to view it *embedded* in a larger space. Topologists point out that the property of being a set of points at a fixed distance of a center point is not an intrinsic feature of the circle itself.

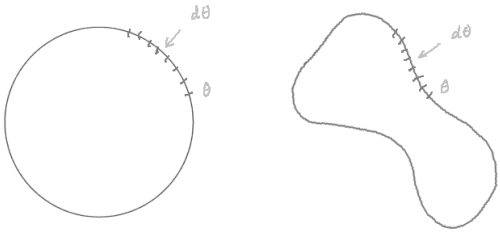

Figure 5 : A circle and one-dimensional loop of same length are topologically equivalent.

By this we mean that some creature living on the circle, and not being able to see outside it, would not be aware that we see its space as a circle. It would just know that it is living in a one-dimensional space, which, instead of being an infinite line, closes back on itself. In other words, for the creature it does not make any difference. That is the topology of a loop [2], see figure 5.

Anyway we call the circle centered at O of radius one, Ω_1. And we will denote $d\theta^2$ as $d\Omega_1^2$. Equation (4) is rewritten

$$ds^2 = dr^2 + r^2 \, d\Omega_1^2 \tag{5}$$

where $d\Omega_1^2$ is the metric of the unit circle. And let's keep in the back of our mind that it is also the metric of any loop of length 2π. From now on, circles will often be called *1-spheres*, and this will sometimes even include loops.

We are going to adopt generally the notation $d\Omega^2$, because we will use it over and over, not only with the index 1 as in equation (5). We don't want to always have to write the details of the metric. So we have a name for the metric of the unit 1-sphere. It is $d\Omega_1^2$. And it enables us to nicely relate, in polar coordinates, the metric of the flat two-dimensional space and that of the unit 1-sphere. We shall generalize this fact.

2. After having taught us in primary school that the circle is the set of points at a fixed distance of a center point, mathematicians like to say that it is the least important of its properties. Another defining property of the circle is that it is a set of points invariant under a certain Lie group of transfomations. It is the starting point of symplectic geometry which simplifies further classical mechanics, like the Lagrangian approach simplified the Newtonian approach.

2-spheres and 3-spheres

The next step is, staying in polar coordinates, and staying first of all in two-dimensions, to go from the plane to the surface of a sphere.

Let's draw the plane on its edge, so to speak, with a series of concentric circles on it, see figure 6. And notice that the center point can be put anywhere, because the flat plane is indeed homogeneous.

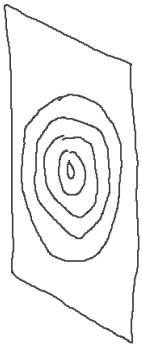

Figure 6 : Series of concentric circles covering the whole plane.

So we can think of the flat space as a kind of nested sequence of circles of increasing size. Better yet, let's see it as a sequence of nested 1-spheres of increasing radius r.

Now, instead of a plane, let's consider the surface of a ball, that is what we call a sphere, or more precisely a *2-sphere*. The surface of the Earth for instance can be idealised as a sphere or 2-sphere. The sphere is also a homogeneous surface. Every place on the surface of the sphere is exactly the same as any other place.

A sphere of course has a rich collection of properties. But for us its main one is that it has uniform curvature. It is everywhere the same. Moreover if you walk around it, you come back to the same place, whatever point you started from and direction you took. These are the properties of the 2-sphere that we care about.

Let's discuss its metric.

Again, like the plane, the 2-sphere can also be charted with a collection of nested circles covering it entirely. And we can put the center of the nested family anywhere, see figure 7.

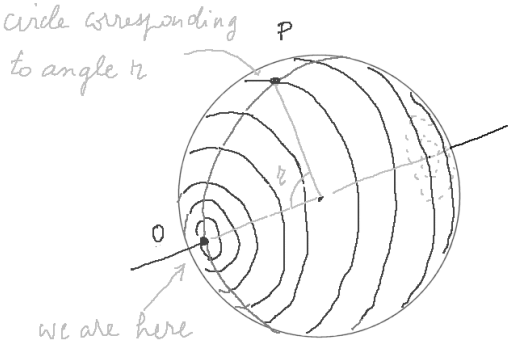

Figure 7 : Series of concentric circles covering the whole 2-sphere.

Let's position ourselves at the center of the nested family of circles – or 1-spheres –, as shown on figure 7. And suppose we are living on the two-dimensional sphere. We don't see any other dimension. We don't see out of the sphere.

Don't be mistaken : as usual we have represented the 2-sphere plunged into 3D, but it is important to forget the

third dimension, and think of ourselves as creatures living in and knowing only the two-dimensional surface.

Suppose I happen to be an astronomer – a "two-dimensional astronomer" in a two-dimensional space –, and I look out into my universe. What do I see ? At one light year I see everything arranged on a 1-sphere. At two light-years I see everything arranged on another 1-sphere. At three light-years same pattern, and so forth.

But there is something different from the plane about this series of nested more and more distant 1-spheres. As I look out farther and farther, at first for a while the radius of the 1-sphere I see increases – though not like the distance at which I look –, then it reaches a maximum, after which it begins to decrease until at some most distant point the 1-sphere has diminished again to a point. On the Earth we call it of course the antipode of where we are.

Let's stress that this increase followed by a decrease is not with time. It is with distance. Time plays no role yet in these considerations about geometry.

The collection of circles – i.e. 1-spheres – covering this 2-sphere are parameterized by the angle which we call r and which ranges from 0 to π. Let's suppose we are considering the unit 2-sphere, that is the 2-sphere of radius of one, then the value r also happens to be the distance from the point O to the point P. But defining the parameter as the angle is fine.

As usual, the parameter at first simply labels a collection of objects. It is an important idea in physics. Sometimes mea-

ningful calculations are possible with the parameter, sometimes not. Generally speaking to view coordinates first of all as labelling parameters, before we make any calculations with them, is useful. We already met this idea for instance, in volume 4, with the time coordinate on various kinds of trajectories when studying the horizon of a black hole. Here of course it is possible to do algebra involving r. In particular the circle corresponding to the angle r has radius $\sin r$.

What is the metric of the unit 2-sphere in terms of r ? It looks a lot like equation (5) but is slightly different. At point P in figure 7, the metric is similar to that at point P in figure 4, except that before applying Pythagoras theorem we must replace $r^2\ d\Omega_1^2$ by $\sin^2 r\ d\Omega_1^2$, because on the circle at angle r an increment $d\theta$ from the center of the sphere produces only a differential displacement $\sin r\ d\theta$ on its circonference. So the metric of the unit 2-sphere is

$$ds^2 = dr^2 + \sin^2 r\ d\Omega_1^2 \tag{6}$$

In the same manner we gave the name $d\Omega_1$ to the metric of the 1-sphere, we now give a new name to the metric of the 2-sphere : we call it $d\Omega_2$. Equation (6) can be rewritten

$$d\Omega_2^2 = dr^2 + \sin^2 r\ d\Omega_1^2 \tag{7}$$

It is important to perfectly understand equation (7). Ω_1 is the unit 1-sphere, that is a circle of radius one ; Ω_2 is the unit 2-sphere, in other words the surface of an ordinary sphere of radius one. And the metric of the latter can be nicely expressed with the metric of the former. This is simply because, as shown in figure 7, the 2-sphere can be built

from a collection of nested 1-spheres sweeping its surface.

This pattern just continues. We now want to make a three-dimensional sphere, something that we will call a 3-sphere. A 3-sphere is a three-dimensional space that is everywhere the same. If you go out in any direction you come back to yourself.

A 3-sphere is not an ordinary ball. It is not an ordinary sphere viewed in its entire volume. It is something different. It is to the usual three-dimensional Euclidean space what the 2-sphere is to the usual Euclidean plane, see figure 9.

Think about looking out in the sky again. Now we are in our real universe, which is three-dimensional. Suppose furthermore that it is not a 3D Euclidean space but a 3-sphere. Then our observations would go as follows. We see things at a certain distance. They form a 2-sphere around us, not a circle as when we were "two-dimensional astronomers" on a 2-sphere. We look further : another 2-sphere bigger. We look still further : another 2-sphere even bigger. At some point comes a largest 2-sphere, after which the radii of the more distant 2-spheres begin to decrease. Finally an ultimate 2-sphere shrinks again to a point.

Our real universe might be a 3-sphere. If it is a very big 3-sphere, its local topology won't reveal it, because locally it would look like the ordinary 3D space – just like on a very big 2-sphere, locally the space looks flat like a plane.

3-spheres are just as good a space as 2-spheres. But they are harder to visualize. Our visual cortex doesn't have the machinery to be able to visualize directly 3-spheres. But we

can use as a bootstrap the construction processes we went through in lower dimensions. We saw that the plane can be constructed from a collection of 1-spheres, figure 6. And the 2-sphere can also be constructed from a collection of 1-spheres, figure 7.

Similarly, the ordinary 3D Euclidean space can be built from a collection of nested 2-spheres. And the newfangled 3-sphere can also be constructed from a collection of nested 2-spheres.

Figure 8 suggests the sequence of nested 2-spheres we can see out in the universe, from the point P where we stand. Now we can no longer show them "really" like we did in figure 7 for the nested 1-spheres.

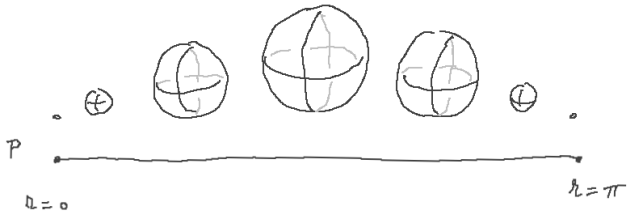

Figure 8 : Series of concentric 2-spheres, which we can see when we look out at distance r. Thay sweep the whole 3-sphere.

Notice that the series of concentric nested 2-spheres can be imagined around any point P in the universe, because the 3-sphere, like the 2-sphere, and also like their Euclidean counterparts, is homogeneous, that is it is everywhere the same.

Let's insist that the series of spheres on figure 8 refers to

the spheres that we see around us. The closest one around me is my head. Then I look at a distance of 1 meter, things are arranged on a 2-sphere. Then I look farther, say at 10 meters, or at 10 billion light-years, I see things arranged on a bigger 2-sphere. Then if I look even farther away, at some distance the 2-sphere on which things are arranged has a maximum radius. After that, going farther again the radius becomes smaller. At some extreme distance it is the end of the 3-sphere universe. But this antipodal point presents nothing dramatic or special in any way. It is not like the end of a rope. It is like – in 2D again – the antipode of San Francisco or of any other point on Earth.

Furthermore this 3-sphere universe is such that if you start out on what appears to be a straight line (mathematically, a geodesic line) in any direction you eventually come back to where you were. Again we are familiar with that on the surface of the Earth.

Figure 9 shows the various spaces we have encountered so far, which have lead us to the 3-sphere

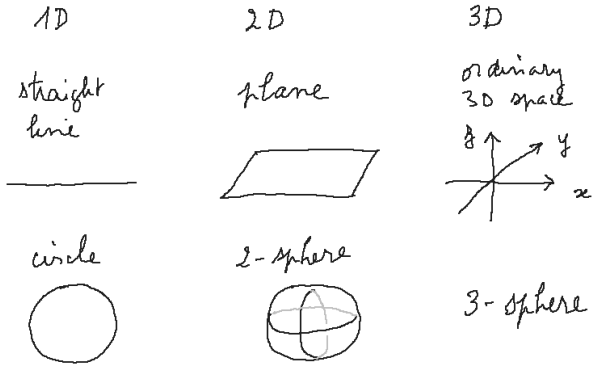

Figure 9 : Flat spaces and their positively curved counterparts.

So we explained what is the 3-sphere and why we cannot represent it in a natural way. The next question is : what is the metric of the 3-sphere ?

We may not be able to represent it, but we can straightforwardly write the metric of the unit 3-sphere. That is the beauty of the construction process with a series of nested spheres of lower dimension. We just extend equation (7) to the next dimension :

$$d\Omega_3^2 = dr^2 + \sin^2 r \; d\Omega_2^2 \tag{8}$$

There is always a dr^2. It is the incremental distance away from us. And there is the angular part. It no longer involves circles, that is Ω_1, of growing radius, but 2-spheres of growing radius. We have to imagine the extension of figure 4. Moreover the radius doesn't grow like r – that would produce the metric of the usual 3D Euclidean space – but like $\sin r$. So the angular part is $\sin^2 r \; d\Omega_2^2$.

Figure 10 summarizes the metric, in polar coordinates, of the various geometries.

As said, we may be living in a 3-sphere. Our universe may be a three-dimensional space whose metric is given by equation (8) – like we finally realized in Antiquity that the surface of the Earth was not a plane but rather the surface of a sphere. Of course in Antiquity this innocent looking suggestion had to fight the counter argument saying : "This is ridiculous, if the Earth was a sphere, there would be people living upside down. They would fall." etc. It lead to a profound change of paradigm. But that took centuries to feel at ease with it. Only in 1492 did Columbus decide to sail

to Cipangu (i.e. Japan) starting boldly straight west.

Since the XVIIth we became somewhat accustomed to changing paradigm once in a while. The last very big instances in physics were the quantum revolution, see volume 2 of the collection *The Theoretical Minimum*, and the special and general relativity revolution, see volumes 3 and 4, which took place in the first third of last century. Despite their spectacular successes, there are still difficult problems in them which suggest that we are in for new big changes. And we even understood that this is forever the fate of science.

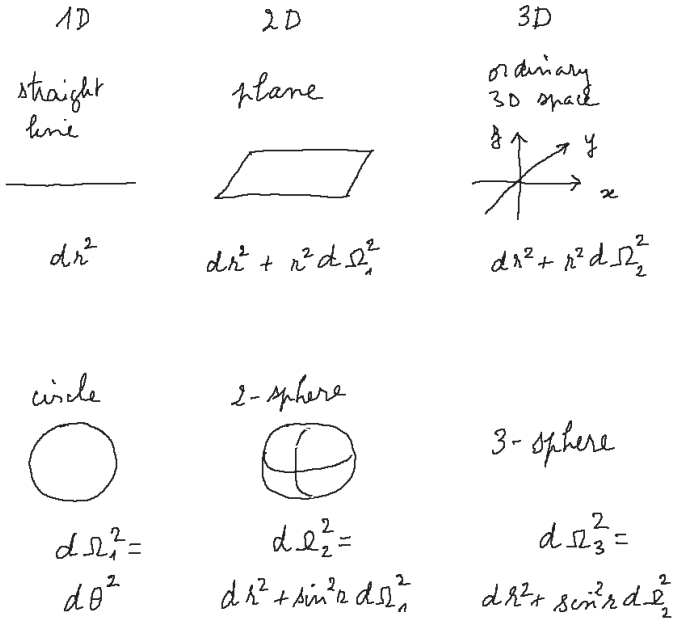

Figure 10 : Metric ds^2, in polar coordinates, in the various geometries.

The first row on figure 10 lists flat spaces, and the second row homogeneous positively curved spaces. Notice in particular that flat 3D Euclidean space has a natural metric in polar coordinates whose expression uses Ω_2. It is

$$ds^2 = dr^2 + r^2 \, d\Omega_2^2 \tag{9}$$

It corresponds to nested 2-spheres without the weird twist of $\sin r$ growth instead of r growth. The growth of the 2-spheres with r, to sweep the ordinary 3D space, is easy for us to visualize.

We see the power of the standard notation $d\Omega_n^2$ for the metric of a homogeneous positively curved space. The index n indicates how many dimensions it has.

Embedding

There is another way to view the spheres. It may be a little more intuitive. But it requires some extra tools.

Suppose you were a little ant living on the circle. And you couldn't look off the circle. All you could do is receive light from along the circle. You could communicate with your neighbors, but you would have no way of telling whether it was truly a circle or some funny shape, see figure 11.

Suppose you couldn't even tell if there was another dimension. Perhaps all there is, is the space along the line with no sense of moving perpendicular to the line. That is what

a creature living on the line, who couldn't see off the line, would experience.

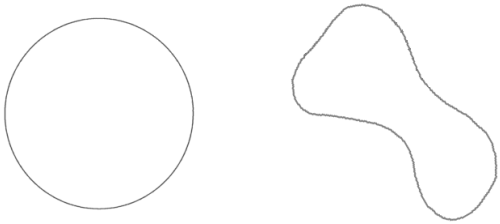

Figure 11 : Ant living in a 1-dimensional space making a loop.

For instance, somebody who lived in an optical fiber, who could receive no light except from along the fiber, would have no way of telling whether that fiber, in three dimensions, was truly circular or it had some other shape. And probably he or she wouldn't care because it would make no sense.

Nevertheless we can, if we like, describe the circle by embedding it in two dimensions. It is only one dimensional but we can embed it in a space with more dimensions.

How do we do that ? We write that the circle is the locus of points in the plane satisfying

$$x^2 + y^2 = 1 \tag{10}$$

That is the unit circle, see figure 12.

If we want to do the same for the unit 2-sphere, we have to introduce a third direction z. And the equation becomes of course

$$x^2 + y^2 + z^2 = 1 \tag{11}$$

In other words, to describe in this way a 2-sphere, even though it is a two-dimensional space, we have no choice but to introduce a fake third dimension.

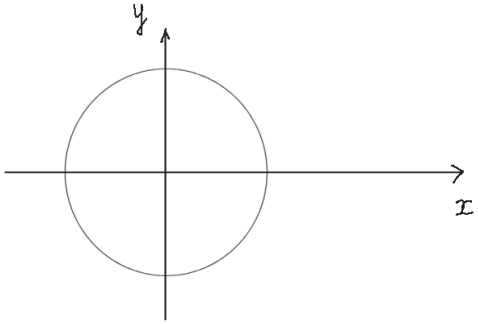

Figure 12 : Unit circle on the plane. To describe a 2-sphere, we would have to introduce a third dimension z.

In the case of the surface of the Earth, the third dimension is real. We can move in the direction perpendicular to the Earth where we are. But again if you thought about the sphere as a sort of flatland, where creatures could only receive light from within the surface itself, then the extra dimension would just be a trick for describing the sphere.

It is described by the Cartesian equation (11), but it is really a two-dimensional space, so it can be charted with only two coordinates. We charted it with 2 angles, r and θ, see figure 7. We did not represent θ ; it is the angle to move *on* the circle at r. These are nothing more than the usual latitude and longitude we are accustomed to, except

that on figure 7 the Earth axis is not represented vertically. Remember that we can place the point O wherever we like on the 2-sphere. We get the same system of angular coordinates, because the sphere is a homogeneous space.

Staying in Cartesian coordinates, we can go another step. We can say let us construct a 3-sphere. We would have to introduce a fourth coordinate on top of x, y and z. Now the four-dimensional space may really be a fake. Maybe only the three-dimensional variety makes any sense. But we would add one more letter.

$$x^2 + y^2 + z^2 + w^2 = 1 \tag{12}$$

In a 4D Euclidean space, the variety [3] defined by equation (12) *is the unit 3-sphere.*

Again if we coordinatize it by distance r from some point selected in it, and the position on the 2-sphere at that distance, the 3-sphere's metric has the form

$$d\Omega_3^2 = dr^2 + \sin^2 r \; d\Omega_2^2$$

In short, embedding a variety in a higher-dimensional space may or may not make real sense, that is really have physical significance. As we said the surface of the Earth is embedded in three-dimensional space.

But if we live in a 3-sphere, chances are it is not embedded in the same way in a four-dimensional space.

3. Sometimes the word *surface* is used for a subspace of any dimension in a Euclidean space of n dimensions. In this book, however, we tend to restrict the word surface to subspaces with 2 dimensions. For subspaces with more dimensions, we use the term *variety.*

Checking if the universe is spherical

Let's go back to a 2-dimensional universe. It is not fundamental for what is coming ; the reasonings can be immediately generalized. But, since our mind is wired to view things in 3D at the most, if we want to embed things into a higher dimension to look at them comfortably, we must start with a 2-dimensional universe.

What difference does it make if we live on a sphere or if we live on an infinite flat plane ? And how can we tell ? [4]

Let us suppose we have at our disposal telescopes which allow us to determine the distance of the objects we observe far away. In other words, our telescopes enable us to know the r of the galaxies in our universe.

Real telescopes don't offer this possibility, but we have various tricks to tell how far away things are. The standard trick of course is spectroscopy and using the Hubble law. So it is a combination of means. But in this particular case what could we use as a trick to tell how far a galaxy was ? One way is to look at the luminosity, how bright the galaxy is, how much light do we receive from it. A light bulb far away looks less luminous than a light bulb close up. So let's assume we can tell the distance.

Let's look at a galaxy in the sky. Remember we assume that our sky is filled homogeneously with galaxies. And furthermore they are all the same. Of course this is not true.

4. The Ancients with a good sight had noticed that the masts of boats or galleys disappeared below the horizon when they sailed away. But this method is not sufficient if we live in a 3-sphere and want to observe galaxies a few million light-years away.

There are several types of galaxies, and we can tell the type when we observe one. If we are looking at a galaxy like our own, we may assume it has about the same size, that is a hundred thousand light-years across approximately, and it emits about the same quantity of light. So when we are looking at galaxies like the Milky Way we can tell that way how far they are. And for simplicity let's just say they are all of the same.

We can ask : what angle do they subtend in the sky ? When we observe a galaxy, obviously the further away it is, the smaller the angle it subtends in the sky.

Let's begin with flat space. We assume all galaxies have the same diameter D. Think of the diameter of the standard spiral galaxy, see figure 13. And, as said, we assume all galaxies are cookie cutter copies.

Figure 13 : Standard spiral galaxy. We assume all galaxies are cookie cutter copies of this. (Source : artistic rendering of the Milky Way, from Nick Risinger / NASA.)

For simplicity we also assume that the universe is two-dimensional, that is a plane since we assumed we are in a flat universe. Limiting the number of dimensions to two is not important. It just makes the geometric reasonings simpler.

So we can write the metric as

$$ds^2 = dr^2 + r^2\ d\theta^2 \tag{13}$$

Around us, we observe several galaxies at different distances, and we can measure the various angles they subtend in the sky. Let's concentrate on one of them, see figure 14.

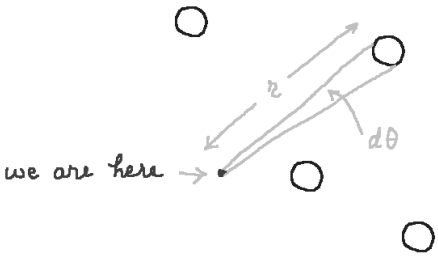

Figure 14 : A galaxy at distance r, subtending an angle $d\theta$.

The galaxy has size D. It is its true size. That means, looking at the distance separating the edges of the galaxy at radial distance r, that ds^2 in equation (13) is D^2. And, since in this case $dr = 0$, we have

$$D^2 = r^2\ d\theta^2 \tag{14}$$

or more simply

$$d\theta = \frac{D}{r} \tag{15}$$

So, using the metric equation (13), we established something which is actually obvious in a flat plane : the angle subtended by the galaxy, from our point of view, is equal to its actual size divided by its distance from us.

We could check this. We can measure the distances to the galaxies, we can see how big they look in the sky. If we lived in a two-dimensional flat universe homogeneously filled with galaxies, all of the same actual size, we could check if the relationship expressed by equation (15) is satisfied on average.

If it was not satisfied something would have to be changed in our model, either the homogeneity or the metric of the universe. But we saw that relinquishing the homogeneity of the universe amounts to assuming that we are at the center of the universe – a hypothesis abandoned since Copernic [5]. So it is the flatness of the universe that would have to be questioned.

The above reasoning can be done in three or more dimensions in the same way to check the flatness of the universe.

Now let's do exactly the same thing on the 2-sphere. And again, it could be extended straightforwardly to 3-spheres

5. Nicolaus Copernicus (1473 - 1543), Polish mathematician and astronomer. In truth, Copernic removed humankind only from the center of the solar system. But it is the same idea : we are not at the center of the universe.

and more.

Our 2-sphere universe is filled homogeneously with galaxies, all about the same size. We look at a galaxy at distance r from us, see figure 15. It subtends an angle $d\theta$. We are going to do the same calculation relating D, r and $d\theta$ as we did in the flat Euclidean case. But this time we use the metric of the 2-sphere, which is

$$ds^2 = dr^2 + \sin^2 r \; d\theta^2 \tag{16}$$

We assume that the units we use are such that we are on the *unit* 2-sphere. The nice consequence is that r is at the same time the distance to the galaxy, and the angle at which it is located in the universe, see figure 7. Remember that r varies from 0 to π. It is easy to extend the reasoning to a 2-sphere universe of any radius a. In equation (16) we just have to multiply the right-hand side by a^2.

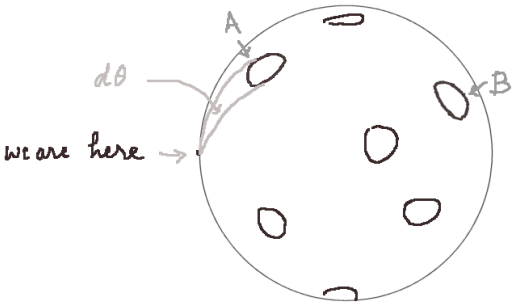

Figure 15 : Galaxy A at distance r, subtending an angle $d\theta$, on a unit 2-sphere.

The analog of equation (14) is

$$D^2 = \sin^2 r \; d\theta^2 \tag{17}$$

or

$$d\theta = \frac{D}{\sin r} \tag{18}$$

Let's compare this result with the one we obtained for the flat universe, see equation (15). Which is bigger at a given distance ? Well, $\sin r$ is smaller than r. Therefore $d\theta$ is bigger than it would have been in the flat case.

If we lived in a 2-sphere, a galaxy a few million or billion light-years away from us would look bigger than if we lived in a flat universe. As said, this also true in a 3-sphere, or a sphere of any dimensions.

At first their apparent size would decrease as r increases, because, before $\pi/2$, $\sin r$ is an increasing function.

But then there is an even more surprising phenomenon : if we looked at galaxies farther away than $r = \pi/2$ they would then start to grow bigger the farther away we looked. For instance, galaxy B, in figure 15, would appear to us bigger than galaxy A. That is because B is closer to π than A is to 0.

Yet B would appear fainter than A. So we would not mix up distances before $\pi/2$ and after $\pi/2$. A galaxy at exactly the other end of the universe from us, that is at the antipode, would appear of infinite size, because we would see in all directions. In other words, it would fill the sky.

So again, since we can measure distances and angles subtended by galaxies, we could check from observations if equation (18) is satisfied on average. We could also check the weird phenomenon of apparent sizes increasing again after a certain distance. In short, we could check whether the universe appears positively curved like a sphere.

Question : Is this also something we can check with the Cosmic Microwave Background ? Answer : Yes, it is analogous to determining the curvature of space by looking at the CMB. More precisely, in the CMB there are features whose apparent size depend on their distance and that we can decipher. It is the size of certain acoustic lumps. We don't look at galaxies in the microwave, we look at oscillating lumps of stuff. But basically it is the same idea.

We could also use other measures to check whether the universe is spherical. We could simply count the number of galaxies at different distances. It is obvious on a 2-sphere that if you look out at some distance, you see fewer galaxies than if you look at the same distance on the plane.

The plane sort of opens up without limits, whereas the sphere opens and then contracts. A flat universe is infinite, whereas a positively curved universe is finite. In fact way out at the maximum distance, you basically have a chance of seeing one galaxy at the most, whereas at the same distance in flat space you would see lots of galaxies. So counting galaxies is another way to tell.

Exercise 1 : Consider a flat plane and a unit 2-sphere with the same density of galaxies ρ on their surface. Compare for different values of the distance r the number of galaxies is the round ribbon between r and $r + dr$, in the two cases.

We remind the reader that we have been reasoning with *unit* 2-spheres, and *unit* 3-spheres, because the formula for their metric is simple, r being at the same time a distance on the sphere, and one of the angular parameters. As mentioned, it is easy to adapt the calculations to a sphere of any radius a by multiplying the right-hand side of the definition of ds^2 by a^2, see equation (16). The conclusions would be exactly the same, because the radius of the sphere doesn't play any fundamental role.

Let's insist, too, that we have been considering 2- or 3-spheres of *fixed size*. At the moment we are just studying their geometry. We are not looking at expanding or contracting universes. Nor are we asking how the spherical universe – if it is spherical – reached the size it has.

Of course the size of the real world does change with time. It grows. But before we talk about the time dependence, we have to talk about another possible geometry for a homogeneous universe.

Hyperbolic universe and stereographic projection

So far we have met two kinds of homogeneous geometries : flat geometries, and spherical geometries. We recapitulated their various metrics in figure 10.

There is a third kind of homogeneous geometry [6]. It has various names ; we are going to call it the hyperbolic geometry. And the corresponding space will be called the hyperbolic space.

It is not as easy to imagine as the sphere. So to help imagination we shall first of all study a geometric transformation called the stereographic projection, and see how it is used in the spherical geometry.

Figure 16 : Illustration of analogous reasoning.

6. We are talking here about *simply connected* homogeneous spaces, where any two paths linking two points P and Q can be continuously transformed into each other. If we include other geometries, like that of the torus where this condition is not satisfied, there are more homogeneous geometries.

Then we will do something that is frequently used in mathematics and physics : analogous reasoning, to help understand a new idea or picture, see figure 16.

The natural representation of 2D spherical geometry is easy to understand. The stereographic projection, which we shall explain, is also easy to understand. It produces a planar representation that is still easy. Then we go to the hyperbolic geometry. We explain what is its planar representation, obtained from a stereographic projection. It is easy too. Finally this helps figure out what is the natural representation of hyperbolic geometry.

So stereographic projection is a creative and useful way of thinking about the sphere. It provides some more insight on spheres. But, as said, more importantly it provides a useful way of thinking about the hyperbolic geometry.

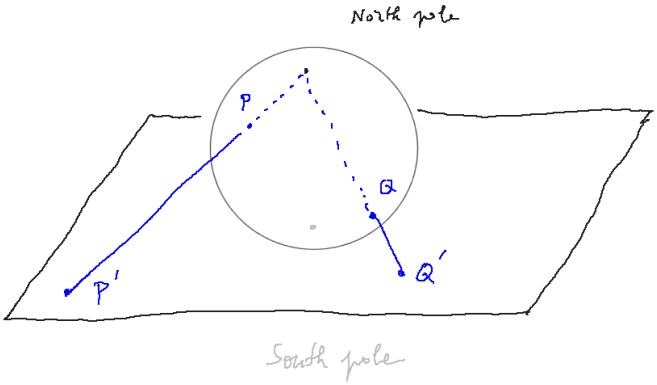

Figure 17 : Stereographic projection of a sphere onto a plane.

We start from a sphere and draw the horizontal plane tan-

gent to the south pole. Then any point P on the sphere is transformed, or "mapped", into a point P' on the plane as follows : draw the straight line going through the north pole of the sphere and point P. Where it intersects the plane is the point P'. In figure 17, we represented the transformed points, P' and Q', of two points P and Q.

So every point on the sphere is mapped into a unique point on the plane. There is a bit of trouble about the north pole, but geometers solve it elegantly : think of a special extra point I added to the plane, and which is at infinity in every direction. Then the north pole itself is mapped into I [7].

Now let's see how the stereographic projection transforms figures on the sphere into figures on the plane. Around the south pole, figures on the sphere are not deformed very much. Whereas figures higher up on the sphere are more deformed. The closer to the north pole, the more elongated their transformed version gets.

It is convenient to think of ourselves as located at the south pole, because then for us the space locally looks about the same on the sphere and on the plane.

The stereographic projection has some properties that seem like magic : circles are transformed into circles (if we count straight lines as circles of infinite radius). And it is *conformal*, meaning that it preserves angles.

7. The enriched geometry of the plane, thus obtained, is called projective geometry. It is an interesting subject on its own. But we don't need to go into it.

Exercise 2 (easy) : **Show that circles going through the north pole are transformed into straight lines.**

Exercise 3 (harder) : **Show that the stereographic projection is conformal.**

Exercise 4 (harder) : **Show that any circle is transformed into a circle.**

However, as pointed out, circles are deformed, in the sense that circles of equal radius on the sphere are not transformed into circles of the equal radius on the plane. It depends on where they are on the sphere.

So that is a way to think about the sphere : by mapping it onto the plane. And when it is mapped onto the plane it has the bizarre property that the further toward the north pole we go, the bigger things look on the plane – the north pole itself corresponding to the point at infinity.

For instance, in the plane, galaxies that had the same radius in the 2D spherical universe no longer have the same radius. Does that mean the universe is not homogeneous ? No, this is just a way of representing the universe, which doesn't geometrically preserve length, although it preserves angles.

We can do the same kind of stereographic transformation with a 3-sphere. We start from a 3-sphere and project it stereographically onto a flat 3D space. Similar things happen. Spheres are transformed into spheres. When they are near us, they transform into spheres of almost the same size. When they are farther away, they transform into larger ones.

Now that we have prepared the ground with stereographic projections, we are ready to examine hyperbolic spaces. We will proceed in the same way : start with a metric and investigate the geometric consequences.

Let's consider first a two-dimensional hyperbolic space, and the *unit* one to boot. Instead of calling it Ω_2, like the unit 2-sphere, we call it $\mathcal{H}_2$. The letter $\mathcal{H}$ stands for hyperbolic. Its metric will be

$$d\mathcal{H}_2^2 = dr^2 + \sinh^2 r \; d\Omega_1^2 \tag{19}$$

It is almost the same as the metric of the 2-sphere, given by equation (7), except that $\sin r$ is replaced by $\sinh r$. In other words, the ordinary trigonometric sine of r is replaced by the hyperbolic sine of r. Professor Susskind likes to pronounce sinh as "cinch", probably to stress that hyperbolic sine is easy :-)

Let's recall what the hyperbolic sine is. First of all, we remember that the ordinary sine of r is

$$\sin r = \frac{e^{ir} - e^{-ir}}{2i} \tag{20}$$

It is the well known formula for sine in terms of complex

exponentials. Hyperbolic sine has a similar expression, even easier

$$\sinh r = \frac{e^r - e^{-r}}{2} \tag{21}$$

In particular, when r is large, the numerator of expression (21) is dominated by e^r. The term e^{-r} becomes indeed negligible. So for large r we have

$$\sinh r \approx \frac{e^r}{2} \tag{22}$$

Hyperbolic sine is a function which grows very rapidly, like half the exponential of r. It is still called an exponential growth of course.

Compare that with ordinary sine. The function $\sin r$ oscillates between $+1$ and -1. In fact, on the 2-sphere, r only goes from 0 to π, so it is only the positive half of one oscillation : it grows from 0 to 1, then shrinks again from 1 to 0.

On the contrary, for the unit two-dimensional hyperbolic space, the range of r goes from 0 to $+\infty$. And $\sinh r$ blows up exponentially without limit.

From equation (19), we see that $\mathcal{H}_2$ is a two-dimensional space still made out of a collection of concentric circles somehow sweeping its surface. But the size of the circles grows very rapidly. We shall see in a moment, however, that, due to the new form of the metric, the homogeneity of the space won't be geometrically obvious as for the sphere.

To go to $d\mathcal{H}_3^2$, which is a candidate geometry for the universe we live in, we straightforwardy extend equation (19). It is now

$$d\mathcal{H}_3^2 = dr^2 + \sinh^2 r \; d\Omega_2^2 \tag{23}$$

In other words, $\mathcal{H}_3$ is also made out of a collection of 2-spheres. If we are at $r = 0$, we are surrounded by concentric spheres like Russian dolls, but which are growing very rapidly. And the metric again has to be handled with care to see the homogeneity.

In order to familiarize ourselves with hyperbolic geometry, let's ask a first question : what happens to the angle subtended by a galaxy as we go further and further away ? We can work it out in the hyperbolic geometry to get the answer.

We are again at the center, that is at $r = 0$, and we are looking at an object which has size D. It is its *intrinsic* size. From our point of view, it apparently occupies an angle $d\theta$ in the sky. As we did before, we can write

$$D^2 = \sinh^2 r \; d\theta^2 \tag{24}$$

or

$$d\theta = \frac{D}{\sinh r} \tag{25}$$

These are the hyperbolic analogs of equations (17) and (18). Now let's plug in what $\sinh r$ is. For large r we use approximation (22). We find

$$d\theta \approx \frac{2D}{e^r} \tag{25}$$

e^r blows up very quicky as we look far away. Therefore the subtended angle $d\theta$ shrinks very fast. Furthermore the number of galaxies grows quite fast too, the size of them shrinking to match. If we lived in a hyperbolic world and looked out in the sky, we would notice that distant galaxies look anomalously small and are too numerous.

There is no antipode in the hyperbolic space, unlike in the case of the 2-sphere. We don't have this phenomenon where after a last point there is nothing left but coming back. And remember, on the sphere, after a certain distance corresponding to $r = \pi/2$, galaxies appear to grow again as we progress further toward the antipode.

So we see that geometries have meanings. Things are quite different on a flat plane, on a 2-sphere, or on a two-dimensional homogeneous hyperbolic space.

The natural representation of a hyperbolic space is a bit trickier than for a sphere. We are going to see it. But let's first talk about its stereographic projection. Figure 18 shows the stereographic projection of a two-dimensional hyperbolic space on which had been drawn homogeneously angels and devils. After the projection they appear on a plane – and on a limited part of it : on a disk. It is a famous picture by M.C. Escher [8]

The angels and devils are background of each other. This striking artistic feature has nothing to do with the hyperbolic geometry. It is only there to show the effect of the transformation. The creatures were homogeneously spread on the hyperbolic surface equipped with the hyperbolic me-

8. Maurits Cornelis Escher (1898 - 1972), Dutch graphic artist.

tric. And the figure shows the distortion operated by the stereographic projection.

Figure 18 : Stereographic projection of a 2-dimensional hyperbolic space onto a plane. Picture by M.C. Escher.

A stereographic projection maps a surface onto a plane, and creates enormous distorsions. In the case of the sphere, the initial space is finite and the end result covers the whole plane, see figure 17. Figures near the north pole of the sphere get very elongated and widened, and sent very far away. In the case of the hyperbolic space, we will discover that it is the opposite. The hyperbolic space is infinite, but its map is finite. It is a disk on the plane.

So let's explain what a two-dimensional hyperbolic space looks like. To this end, as usual, we will embed it in 3D, because that affords us a feeling for its shape and metric. The homogeneity won't be as obvious as for the sphere. And we will also explain how we do a stereographic projection of it onto the plane. It is not done exactly as with the sphere, simply because Ω_2 and $\mathcal{H}_2$ are quite different.

A two-dimensional sphere is

$$x^2 + y^2 + z^2 = 1 \tag{26}$$

To draw a two-dimensional hyperbolic space, again we start with three coordinates. We could call them x, y and z, but we will call them x, y and t. It is not really time ; it is just a trick for drawing the space, as we explained in the section on embedding. The equation of the unit two-dimensional hyperbolic space is

$$t^2 - x^2 - y^2 = 1 \tag{27}$$

So, instead of considering $t^2 + x^2 + y^2 = 1$, which would produce a sphere, we consider $t^2 - x^2 - y^2 = 1$. The surface that it produces is a hyperboloid. The positive value on the right-hand side makes it of the two-sheet kind.

In a 2D embedding space, $t^2 - x^2 = 1$ would produce a hyperbola. But in a 3D embedding space, $t^2 - x^2 - y^2 = 1$ produces a hyperboloid of revolution.

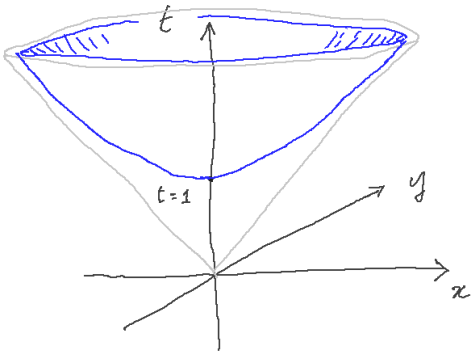

Figure 19 : Upper sheet of the unit two-sheet hyperboloid.

The hyperboloid has the asymptotic cone $t^2 - x^2 - y^2 = 0$, which can be used to conveniently draw the hyperboloid, see figure 19. When we put $+1$ instead of 0 on the right-hand side we get the unit two-sheet hyperboloid. Let's concentrate on its upper sheet only.

We are not doing relativity, yet the equation and figure we work with have a lot in common with relativity. The cone in figure 19 would correspond to the light cone of the Minkowski space, see volume 3 of the collection *The Theoretical Minimum* on special relativity.

It does not look like every point on the hyperboloid in figure 19 is the same as every other point. In other words, it does not look like it is a homogeneous two-dimensional space. It looks like the curvature is maximum at the point $x = 0,\ y = 0,\ t = 1$ and flattens out when we move away from it. But this is true only if we mistakenly use the metric $ds^2 = dt^2 + dx^2 + dy^2$.

If we use correctly the metric

$$ds^2 = -dt^2 + dx^2 + dy^2 \tag{28}$$

then in fact at every point we have exactly the same shape, the same curvature, and the two-dimensional hyperbolic space of figure 19 is perfectly homogeneous. In other words, on our surface, we must measure distances using the relativistic metric, pretending that t was time.

Here the coordinate t is not time though, it is just a third fake coordinate which we introduced for the sake of creating an embedded representation, because it is easier for us to

view things in 3D. Yet this does not enable us to really view things as we are accustomed to in 3D because the metric in the embedding space is the Minkowski metric given by equation (28).

When we work in two dimensions on the hyperboloid itself, with the two hyperbolic polar coordinates r and θ, forgetting about any third dimension, the metric is

$$ds^2 = dr^2 + \sinh^2 r \; d\theta^2 \tag{29}$$

It is still relatively easy to figure out, because it remains an application of Pythagoras theorem analogous to what we did in figure 4. But when we work with three Cartesian coordinates x, y and t in a 3D embedding space, then the metric becomes

$$ds^2 = dx^2 + dy^2 - dt^2 \tag{30}$$

which is less easy to naturally figure out.

Exercise 5 : Show that when embedding the unit two-dimensional hyperboloid in a three-dimensional Cartesian space as above, the metric that was given by equation (29) is now given by equation (30).

How do we know that the hyperboloid is uniform with the metric given by equation (30) ? One proof for instance is to point out that going from one point to another point on

the hyperboloid is equivalent to a Lorentz transformation which moves the time axis in various ways.

If the reader is not comfortable with that. We will just take it for granted that the hyperboloid, when distances are measured with the metric $ds^2 = dx^2 + dy^2 - dt^2$, is completely uniform.

On the sphere, even though the metric is not exactly Euclidean, we perceive it easily. The sphere is locally flat. We are not too much disturbed by the curvature to perceive distances. The homogeneity poses no problem. The same is true in fact of the hyperboloid with the Minkowski metric : we should be able to figure it out without too much trouble. It is locally flat too. But homogeneity surely is not as obvious.

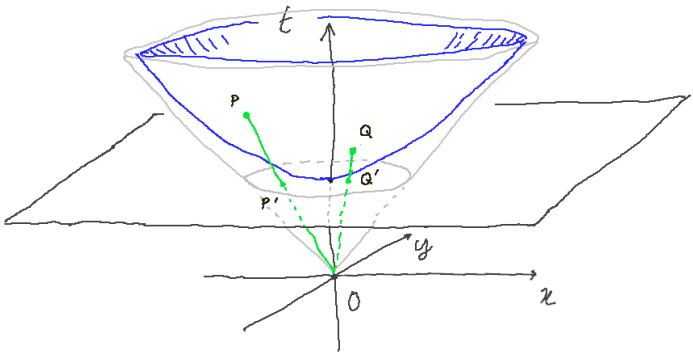

Figure 20 : Stereographic projection of a hyperboloid. The entire hyperboloid is mapped onto the disk on the plane.

Now let's see how we do a stereographic projection of the

hyperboloid. Again we draw a plane on which points from the hyperboloid will be mapped. It is the tangent plane to the bottom point $x = 0$, $y = 0$, $t = 1$, see figure 20. And we will draw straight lines linking points on the sheet and the origin. For instance, we draw a line from point P to O, it intersects the tangent plane at P'. That is the transformed of P. Similarly point Q in figure 20 is mapped onto point Q'.

We see that the entire hyperboloid is mapped onto the disk shown on the tangent plane.

Again think of us as living at the point of contact of the hyperboloid and the tangent plane. Then figures close to us on the hyperboloid will not be distorted very much when projected on the plane. On the other hand figures far away on the hyperboloid will be squished close to the border of the disk. The asymptotic cone of the hyperboloid intersects the plane on the circle bordering the disk.

Now we understand Escher's picture, in figure 18, better : the angels and devils were initially drawn with a regular homogeneous repartition on the hyperbolic sheet – equipped with the hyperbolic metric (29) or (30). But then we applied a stereographic projection, and Escher's picture is the result.

As we move out on the hyperboloid we get closer and closer to the asymptotes of the hyperboloid, and on the map closer and closer to the disk's boundary. So there is a lot of distortion again. Things near the center don't get distorted much. But things, or angels and devils, far away from us on the hyperboloid are squished close to the edge of the disk, and become very small. They become even more numerous,

and the angle they subtend is smaller, than on a flat plane at the same distance.

This is exactly the opposite distorsion from that produced by the stereographic projection of the sphere. On the sphere near the north pole – that is, near the other end of the universe from us – things got sent very far toward infinity on the plane. On the hyperbolic space, on the other hand, things near infinity on the curved sheet – the end of the universe on the hyperbolic space is really at infinity – get squished, and cram near the boundary of the disk.

In summary : on the map of the sphere, everything far away was too big, and they were too few. On the map of the hyperbolic space, everything far away is too small and they are anomalously numerous.

In both cases, on the other hand, things close to us are not much distorted. This is the consequence of the fact that both spaces are locally flat, and in each case the stereographic projection is done in such a way that our location is mapped onto itself.

One more comment on Escher's picture : when looking at it, we are supposed to think in our mind of the hyperbolic space itself, and to view every angel and devil as exactly the same size as every other one. In fact there are coordinate transformations that allow us to move the devils and angels around, analogous to the rotation of a sphere. Like the sphere, the hyperbolic space is homogeneous.

If we look carefully at figure 18, we see that every devil or angel essentially sees exactly the same features around him.

Let's not pay a lot of attention to the size but rather focus our attention for example on how many angels and devils each one sees neighboring him. Then we see that the space is really homogeneous. Every point is the same as every other point, as far as the shape and metric of the space is concerned. When we view the stereographic projection of the hyperbolic space, we should think somehow that we are viewing the hyperbolic space itself.

This geometry has various names : it is the Poincaré [9] space, or the disk version of it ; it is the Lobachevsky [10] plane ; it is the hyperbolic geometry. It is also the uniformly negatively curved space.

The sphere is the uniformly positively curved space, and the hyperbolic space the uniformly negatively one. This actually means something technical. There is a certain component of the geometry which is really a positive number in the case of spherical geometry, and which has the opposite sign in the case of hyperbolic geometry. See volume 4 of the collection *The Theoretical Minimum* on general relativity, in which we studied extensively Riemannian geometry, curvature, the Riemann tensor, and the curvature scalar.

Incidentally, just as the sphere has a natural radius, the hyperbolic geometry also has a radius. Its square is the real number on the right-hand side of equation (27). There it was 1, and the equation was that of the unit two-sheet hyperboloid. This number 1 has a geometric meaning which

9. Henri Poincaré (1854 - 1912), French mathematician, theoretical physicist and philosopher of science.

10. Nikolai Lobachevsky (1792 - 1856), Russian mathematician and geometer.

can be viewed in figure 19. It is the lowest ordinate point. Just as we can write the equation and draw the sphere of radius 2, we can do the same for the hyperboloid. Equation (27) becomes

$$t^2 - x^2 - y^2 = 4 \tag{31}$$

In figure 20, the lowest point moves to $t = 2$. We don't obtain simply the translated version of the hyperboloid of radius 1 of course. It still has the same asymptotic cone. But it is flatter. And in fact its negative curvature has a smaller absolute value.

The point is that each of the curved geometries we have been considering – the 2-sphere, the 3-sphere, the 2-hyperboloid, the 3-hyperboloid, etc. – have a radius associated with them. So far we considered geometries of radius 1. It was implicitely expressed in all the metrics

$$\begin{aligned} ds^2 &= dr^2 + \sin^2 r \; d\Omega_1^2 \\ ds^2 &= dr^2 + \sin^2 r \; d\Omega_2^2 \\ ds^2 &= dr^2 + \sinh^2 r \; d\Omega_1^2 \\ ds^2 &= dr^2 + \sinh^2 r \; d\Omega_2^2 \end{aligned}$$

To change the radius from 1 to any number a, we just have to multiply the right-hand sides by a^2. Thus, for instance, the metric in polar coordinates of the ordinary sphere of radius a is

$$ds^2 = a^2 \; (dr^2 + \sin^2 r \; d\Omega_1^2 \;) \tag{32}$$

When the metric is expressed in Cartesian coordinates in an embedding space, we do the same thing : we multiply

the definition of ds^2 by a^2. For instance the metric of the two-dimensional hyperbolic space of radius a is

$$ds^2 = a^2 \ (dx^2 + dy^2 - dt^2) \tag{33}$$

It is assumed in most of cosmology that the space that we live in is one of these three spaces [11] :

a) a flat 3D Euclidean space, or

b) a 3-sphere, or

c) a three-dimensional hyperbolic space, we can also call a 3-hyperboloid.

Remember that in the last two cases, since the universe is very big, in other words its radius is very large, locally it looks the same as a flat 3D space. The two stereographic projections we studied illustrated this for two-dimensional spaces : locally around us the mapping onto a plane did not much distort the initial space. "Locally" means a distance small in comparison with the radius of the space. But on very large scales, a curved universe is definitely not the same as a flat one. Galaxies very far away display weird phenomena, they are too few or too many, to large or too small, etc.

If we increase the size of a sphere what happens to its curvature ? It becomes smaller. The surface of the Earth is less curved than a basketball. In other words, the smaller the radius of the geometry the higher its curvature. Conversely, the larger the radius of the geometry, the flatter the space

11. There are even other possible geometries for our universe, see below. They are no longer simply connected topologies ; they have holes. We describe in some details the torus, which happens to also be a flat geometry.

appears up to some distance.

Astronomers have not been able to observe any curvature yet. On distances as far as our detecting instruments can go, the actual geometry of our universe looks flat. The study of the Cosmic Microwave Background enables us to go up to 20 billion light-years, and things still look flat. Because of this, we know that if the universe is curved it is at least ten times larger in radius, and a thousand times larger in volume. Even the factor 10 is very unlikely for reasons we will come to. It is almost certainly much more than that.

For simply connected geometries – that is geometries with no holes –, a flat universe is necessarily infinite. But there are other kinds of homogeneous geometries, even flat ones that are also finite. There is no good reason to speculate that our universe could be one of those. However for the sake of completeness and to provide more insight into the geometries we have been studying, let's say a word about these other geometries.

Other possible shapes

There are other homogeneous geometries, even flat ones. The universe could for instance be a *torus*. In two dimensions that is the mathematical name for the shape of a donut, see figure 21. But there are tori of any dimensions, just like there are 2-spheres, 3-spheres, and more.

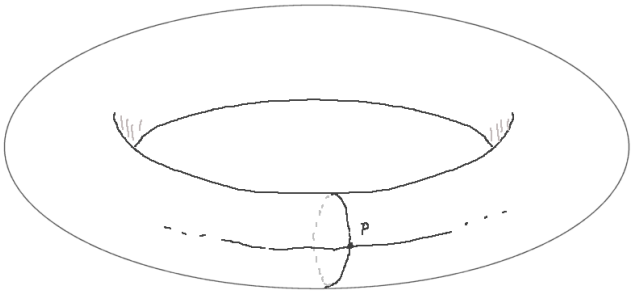

Figure 21 : Ordinary torus, also called casually a donut. It is a two-dimensional surface, which can be equipped with a flat metric.

The surface of a donut does not look flat. But that is only because we look at it embedded into our usual three-dimensional space and think of its usual metric. Remember that embedding is very useful for various purposes – for viewing, or for expressing the metric with certain coordinates –, but it also has drawbacks. It may wrongly suggest a shape. We encountered that with the 2D hyperbolic space, which does not look like it has uniform curvature, let alone a negative one. The same is true when we embed an ordinary torus into our usual 3D space.

A donut has the topological property that from any point P we can go around the torus along different kinds of paths, see figure 21. One category of paths is around the donut hole. Another category of paths is around the donut tube, as shown in the picture. There are other categories combining the two. They wind around the donut in various spiral ways. This is not counting small loops simply in the vici-

nity of P. By definition of a category, within it two paths can be transformed into each other continuously. All these explanations rely on geometric perception and intuition. But there is a more rigorous and useful definition of a two-dimensional torus.

When mathematicians speak of a torus they are usually speaking of a certain topology defined as follows : start from a rectange, with the ordinary Euclidean metric, and sort of paste or link together the opposite sides to think of a finite space without borders [12], see figure 22.

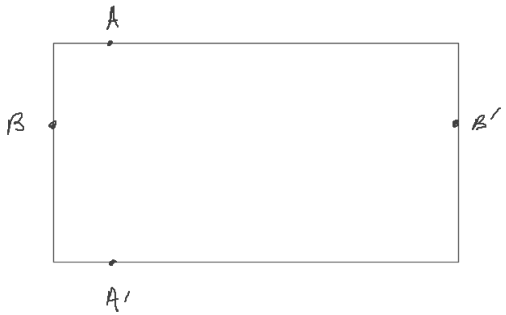

Figure 22 : Mathematical definition of a torus. Link the upper side to the lower side in such a way that A and A' become the same point. Similarly link the left side to the right side. B and B' become the same point. We get a topology equivalent to that shown in figure 21, and with a perfectly flat metric.

After this operation, the torus we obtain is finite and ho-

12. Surprisingly enough, with this way of viewing, a billiard table displays some aspects of a torus (complicated with symetries) ! Players do use this property to hit a ball after one or several rebounds along borders.

mogeneous. It no longer has borders or any other special points. They are all equivalent. We can go through what was a border in the rectangle, for instance through B' toward the right : that means just continuing from B.

The ordinary torus is a two-dimensional surface. But we can think of tori of any dimensions. We can start from a 3-dimensional rectangular parallelepiped (a milk carton), and link together opposite sides, which are now 2-dimensional, as we did above. We obtain a three-dimensional torus, which we can call a 3-torus.

What is a one-dimensional torus ? It is a circle, or Ω_1.

If the universe was a torus, the only thing strange that would be observable is that when looking far enough in one direction we would eventually see our fanny. This is obviously true on the circle, but it is also true on a torus of any dimension. You could also think of the infinite universe as a simple tessellation of identical rectangles replicated ad infinitum. And you would think of yourself as comprising an infinite number of replications – one per rectangle.

Our universe could be a 3-torus. But if the basic parallelepiped block is bigger than the farthest distance we can look to, we won't be able to see this sort of folding or replicating property. It would be equivalent, for all practical purposes, to the usual 3D Euclidean space, just like a big enough billiard table is a good model of the entire plane. For these reasons the model universe as a torus is not a very popular idea, although it does have some interesting features.

Question : Is it possible to have a topology at the same time spherical and with this replication feature ? Answer : Not really. A sphere is not like a rectangle. From any point it grows and then shrinks. How would we paste two spheres ? We would not have an obvious homogeneity. The contact point would be quite special, and probably not very pleasant a place to pass through.

We are finished with the study of the various possible geometries of a fixed universe. The most important ones are the first three, homogeneous and simply connected, that we described in some detail : flat, spherical, and hyperbolic.

It is now time to turn our attention to a universe evolving with time.

Space and time

Let us review the metric of spacetime. We studied it in special relativity, that is the subject of volume 3 in the collection *The Theoretical Minimum*, and we will use it in cosmology.

We use the ordinary Minkowski space. It is a flat spacetime. Its metric has two pieces : a time piece and a space piece.

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2 \tag{34}$$

Notice that $dx^2 + dy^2 + dz^2$ is just the metric of flat space. So we have taken time and space and put them together. As always in special relativity the time component of the

metric has a negative sign in front, whereas the space components have a positive sign.

The reader may want to brush up his or her knowledge of the basic concepts of special relativity : events in spacetime, proper distance between two events, proper time (= the opposite of proper distance), Galilean referentials, Lorentz transforms, etc. The squared element ds^2 can be positive or negative. When it is positive the interval ds is said to be space-like. When it is negative, it is time-like.

Do you remember how a light ray moves ? What is its trajectory ? It is a trajectory along which ds is always equal to 0. For obvious reasons such a trajectory is also called a null ray. Let's consider that it moves along the x-axis. We may drop y and z which are not necessarily for the analysis. So equation (34) becomes

$$0 = -dt^2 + dx^2 \tag{35}$$

or equivalently

$$dx = \pm dt \tag{36}$$

$dx = dt$ characterizes the trajectory of a light ray moving to the right with unit velocity. In the Minkowski plane (usually with t-axis vertical and x-axis horizontal) it is a straight interval with 45° slope. And $dx = -dt$ corresponds to a light ray moving to the left with unit velocity. It is a straight interval with -45° slope.

In a spacetime, there is always time and always space. But we can consider a more general spacetime than that given by equation (34). We can consider a spacetime whose

spatial part of the metric has another form, and moreover which may depend on time.

So we will keep the time part just as it is, but substitute for the three-dimensional fixed flat space, corresponding to $dx^2+dy^2+dz^2$, one of the three geometries we studied. One possibility is flat space. Another possibility is the sphere. The third possibility is the hyperbolic geometry. And we are going to include one more thing : a scale factor.

Let's consider the case of the 2-sphere. It is easy to visualize [13]. So we are considering a spacetime with one time dimension, and a spatial part with two dimensions. It is no longer the Minkowski flat space we are accustomed to. Now the metric is

$$ds^2 = -dt^2 + a(t)^2 \, d\Omega_2^2 \tag{37}$$

This spacetime is curved, but not with the very general kind of curvature mixing space and time which we studied in general relativity. It is a spacetime, or cosmology, where the time dimension is simple and straight – no black holes interchanging time and space... –, and the space part is curved. Moreover its shape depends on time. The radius of the 2-sphere varies like $a(t)$.

One speaks of a 2 + 1 dimensional cosmology. "2 + 1" means two dimensions of space plus one dimension of time. Furthermore, in this one, space is finite and has a time-

13. By this we mean that we can embed the spatial part in 3D to view it comfortably, and imagine its evolution over time. In such a visualization, unlike in the Minkowski space, we don't represent geometrically the time dimension.

dependent radius.

Again we can write down how light rays move. They have trajectories along which $ds = 0$. We will work out some examples later.

What does the inflating 2-sphere look like ? Well we can visualize it as the surface of a basketball whose size increases. When visualizing like that, we plunge the 2-sphere in 3D. It is convenient, but it should not wrongly suggest that there are three spatial dimensions. There are only two. And such natural visualization is impossible in higher spatial dimensions.

We can indeed consider a spacetime where space is a 3-sphere. The metric is

$$ds^2 = -dt^2 + a(t)^2 \; d\Omega_3^2 \tag{38}$$

Same kind of world, with three spatial dimensions, and a size increasing if the scale factor $a(t)$ increases.

Let's keep to the 2-sphere, however, because it is easier to think about and draw pictures. Now take any pair of points on the sphere, separated by a given angular distance θ. What is the actual distance between the two points ? It is

$$D = a\theta \tag{39}$$

What is the relative velocity of the two points ?

$$V = \dot{a}\theta \tag{40}$$

We are keeping the angular positions of the two points, therefore their angular distance too, fixed on the sphere. But we are just letting the radius of the sphere change with time. What about the velocity as a function of distance ? It is

$$V = \frac{\dot{a}}{a}\, D \tag{41}$$

We have seen this formula before. It is the Hubble law, with the ratio $\dot{a}/a$ being the Hubble constant, which we denoted H. Remember : it doesn't depend on space, but it depends on time. So we are talking about something quite similar to what we talked about in chapter 1.

Notice that equation (41) doesn't depend on whether we are in a flat space, a 2-sphere, a 3-sphere or even a hyperboloid of any dimension. It is simply the same equation we had with the inflating grid, accompanying the expanding flat 3D universe, we studied in chapter 1. The change of scale of the geometry, and the subsequent Hubble law, doesn't much depend on the particular shape of the geometry. Of course on a hyperboloid we have to be careful to work with its specific metric.

Let us write now the spacetime metric of the three cases, and let's write them for 3D spaces.

a) Flat space :

$$ds^2 = -dt^2 + a(t)^2\ (dx^2 + dy^2 + dz^2)$$

b) Spherical space :

$$ds^2 = -dt^2 + a(t)^2\ d\Omega_3^2$$

c) Hyperbolic space :

$$ds^2 = -dt^2 + a(t)^2 \; d\mathcal{H}_3^2$$

All this means is that, at any fixed time t, the geometry is either a flat space, a sphere, or a hyperboloid. But the scale of the space, and the distance between any two galaxies, frozen in the moving grid, changes according to a and $\dot{a}$. That is our cosmology.

The next step will be to give it some dynamics. What do we need for that ? We need equations for how the scale factor a changes with time. How does the universe change ? Does it expand ? Will it continue to expand ? At what rate does it expand ? Is it like $t^{2/3}$, or $t^{1/2}$, or e^t, or whatever.

Now we can no longer use Newton's equations, like we did in chapters 1 and 2, although we will find that the equations we will reach are Newton's equations. We cannot start with Newton's equations because we are talking about curved spacetime. We are not talking about anything Newton would have written down. The geometries are curved in spacetime – here only in space, but all the same – we have to use more elaborate tools.

What are the rules for such geometries ? How does physics work ? The answer is given by general relativity. So we will write down the equations of general relativity for the three cases of the metrics listed above. They will translate into equations of motion for $a(t)$.

We will discover that the three cases lead to the same equations respectively as the Newtonian equations for zero

energy, positive energy, and negative energy. They will correspond to a universe at escape velocity, above escape velocity, or below escape velocity. That is our objective for the next chapter.

We will not spend a lot of time going through the equations, calculating all of the symbols in Einstein's general relativity. I will just sketch out what are the equations, and what they lead to in terms of equations for $a(t)$.

The equations are simple, we have seen before, they are the Friedmann equations. Once we have them and know what they mean, we will explore the dynamic cosmology of the universe they entail.

Questions / answers session

Q. : Are these the only cosmologies possible ?

A. : No. By no means are they the only ones. We can consider more complicated cosmologies. And you could ask : why don't we study them too. The fact is that a lot of them will naturally evolve toward one of the three cosmologies, flat, spherical, or hyperbolic. But, undoubtedly we can have more complicated cosmologies. These three are the easy ones. Let us call them popular.

Q. : Is luminosity the only way to determine how far is a

galaxy ?

A. : No. There is a whole range of different methods. Another one is to use the Hubble law to relate distance to velocity, and then measure the velocity by the redshift, or by the Doppler shift of spectral lines. But it is true that luminosity is the simplest and is quite convenient. Supernovae, for instance, have a definite absolute luminosity. So it is easy to figure out their distance. Astronomers use several methods to ascertain their measurements.

Q. : How the expansion of the universe manifests itself at our scale ?

A. : At our scale it is not measurable. If you were holding your friend's hand out in space, the distance between you and your friend would not expand with the general expansion.

The general expansion does produces a kind of very very mild repulsive force between everything and everything else, basically proportional to the Hubble constant. But this very tiny repulsion between you and your friend is vastly more than made up for by just the attractive force of you holding on to your friend's hand.

The same it true for the solar system, or even for atoms. Within an atom the electrostatic forces more than overwhelm, by an enormous amount, the mild repulsive force due to the universe expansion. Similarly the solar system is held together by gravity. The gravitational pull of the Sun is sim-

ply much larger than the tendency to to expand.

Q. : How is the expansion of the universe consistent with the principle of conservation of energy ?

A. : Energy conservation in an expanding universe is different than energy conservation in a static universe. There are several ways to think about energy conservation. Let me give you the simplest way to think about it. Energy conservation is a consequence of time translation invariance. In other words, if everything is time independent, meaning to say space and time and all experiments would reproduce exactly the same effect if they were considered or performed later or earlier, then the consequence of that is energy conservation.

On the other hand, if the basic setup, that is the background spacetime itself, is not static but expanding, or changing with time in whatever way, then energy conservation doesn't apply. There is no energy conservation in a world where the parameters of the world are time-dependent.

In the case of an expanding 3-sphere universe, for instance, its radius is time dependent. Then energy conservation is not quite what we are accustomed to. Basically changes of energy in the universe translate into kinetic energy of expansion. We are going to study it.