Lesson 4 :
Cosmological thermodynamics

Notes from Prof. Susskind video lectures publicly available on YouTube

Introduction

Before we go into cosmological thermodynamics, I would like to clarify two points : first, how is homogeneity of space defined mathematically, secondly, what exactly do we mean by a spherical universe and more generally by the geometry of a space.

Homogeneity of space is a statement about geometry. Loosely speaking, it means "at every point space looks the same". How do we describe the geometry of a space ? We use a metric. Suppose to label points in the space, we use a coordinate system X^m where m ranges from 1 to the number of dimensions. Then at any point P with coordinates X, the square of the length of a little displacement is given by

$$dS^2 = g_{mn}(X)\ dX^m\ dX^n \tag{1}$$

g_{mn} is called the metric tensor. To review the meaning and use of equation (1), the reader is invited to go back to chapters 1 and 2 of volume 4 of the collection *The Theoretical Minimum* on general relativity, which are devoted to geometry, tensors and the metric tensor.

The metric tensor represents some geometry. Now we can make a coordinate transformation. We can use a new coordinate system Y^m to label the points. The origins of the two systems are not necessarily the same points. There is a point corresponding to $X = 0$, meaning all the values X^m are zero, and there is another point corresponding to $Y = 0$.

As said, a coordinate change is a new labelling of the same points in the space. It can also be viewed as a *transfom* which sends the point $X = 0$ into the point $Y = 0$. And similarly each point is sent to another point with the same coordinates values in the new system.

When we do a coordinate transformation, the expression of the metric changes. Remember : just like in a vector space, vectors are intrinsic geometric objects, but the *components of the vectors* depend on the basis used to express them, in an n-dimensional variety the metric is an intrinsic characteristic of the geometry, but the *components of the metric tensor*, that is the expression of the metric, change when we change coordinate system.

Equation (1) changes into an equation involving the new expression of the metric, which we denote g'. A point Q has coordinates Y[1]. Now at Q the square of the length of a little displacement in the Y coordinates is given by

$$dS^2 = g'_{mn}(Y)\ dY^m\ dY^n \tag{2}$$

Typically the metric in the Y coordinates will have a different form than the metric in the X coordinates. Suppose we look at the two origins, and we look at the change of coordinates as a *transform* that sends $X = 0$ into $Y = 0$. Then the change from g to g' could be for two reasons

1. It could be because the space itself is different at $X = 0$ and at $Y = 0$. We might discover that the curvature is larger at one point than the other. Or in some other way the shapes are different.

[1] The point Q we look at may be the same as point P, and we look at its new coordinates. Or we may look at the point Q whose Y-coordinates have the same values as those of P in the X-coordinates. Either way, we look at the effect of the coordinate change.

2. Or it could be not because the shapes are different at the two points but simply because we use screwy coordinates at the old origin and some other screwy coordinates at the new origin.

A space with metric g is homogeneous if, given any two points P and Q, we can find a coordinate transformation which will transform P into Q and such that the form of the metric is identical before and after the transformation.

For example let's just take two-dimensional flat space and a Euclidean coordinate system $(X_1,\ X_2)$. The metric is

$$dS^2 = dX_1^2 + dX_2^2 \tag{3}$$

Now suppose we simply make a translation of coordinates. We define a new system $(Y_1,\ Y_2)$ as follows

$$\begin{aligned} Y_1 &= X_1 - a_1 \\ Y_2 &= X_2 - a_2 \end{aligned} \tag{4}$$

This is a shift by a vector $(a_1,\ a_2)$ which sends the old origin into a new origin. The new origin, in the old system, has coordinates $(a_1,\ a_2)$. And of course in the new system it has coordinates $(0,\ 0)$.

What is the metric in terms of Y-coordinates? We can find it by simply expressing dX_1 in terms of dY_1, and dX_2 in terms of dY_2, in equation (1). But they are the same. When differentiating Y_1 in terms of X_1 in equation (4), the constant a_1 disappears. Of course we get $dY_1 = dX_1$. Likewise for dY_2 and dX_2. So for this kind of transformation the metric keeps the same expression in the new coordinates

$$dS^2 = dY_1^2 + dY_2^2 \tag{5}$$

The implication of that is that the neighborhood of the origin $X = 0$, for instance, has exactly the same geometric properties as the neighborhood of the new origin $Y = 0$.

Now since the shift $(a_1,\ a_2)$ could be anything, there is a coordinate transformation which takes the old origin to any other point whatever, and which preserves the form of the metric. That is what we call a homogeneous space, whose properties are everywhere the same.

So we proved that the flat plane is homogeneous in the rigorous mathematical sense we introduced : if there exists a coordinate transformation $X \to Y$, which takes the origin to any other point we choose, such that the form of the metric in terms of Y at the new origin is the same as its form in terms of X at the old origin, then the space is called homogeneous.

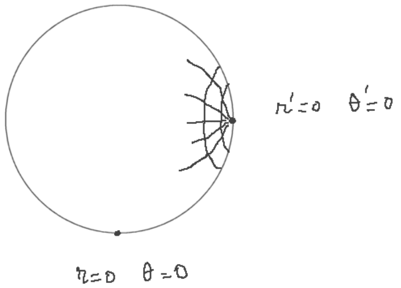

Figure 1 : Change of coordinates on a sphere.

Here is another example not flat, see figure 1. Let's consider the unit 2-sphere. And let's put the origin initially at the south pole. A first system of coordinates is $(r,\ \theta)$. We could perfectly call them $(X_1,\ X_2)$, but we prefer to use the standard notation for polar coordinates. r is the distance from the south pole[2]. And θ is the longitude. By now, we are familiar with the metric

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

Remember that $\sin^2 r$ is a lighter notation for $[\ \sin(r)\ ]^2$, that is the square of sine of r.

Then we consider a new coordinate system $(r',\ \theta')$ obtained by rotating the sphere. For instance let's put the new origin at the east pole, measure r' from there, and θ' from there too as shown on figure 1. It is a bit intricate to write the coordinate change – that is r' and θ' in terms of r and θ –, but we can show that the *same metric* of equation (6) takes on the new form

$$ds^2 = r'^2 + \sin^2 r' \ d\theta'^2 \tag{7}$$

In other words, we found a transformation such that the expressions of the metric and therefore the shapes of the space around the south pole and around the east pole are the same. We can also do it after having moved the south pole by rotation to any point. This establishes mathematically that the 2-sphere is homogeneous.

[2] We could say "r is the distance from the south pole measured along the surface of the sphere", but it is better to forget about embedding. Let's think of the 2-sphere as a space on its own, with two coordinates and a metric. And the metric is given by equation (6).

One test which works very well in two dimensions, to check for a uniform geometry, is that the curvature should be the same everywhere. That is not good enough in more dimensions. But the basic idea is : if the geometry is the same everywhere, then it is a homogeneous space. It can be then shown that it is also isotropic : at any point the space looks the same in every direction.

It is true in the flat plane. It is true of the sphere. It is also true of the hyperbolic plane. So the analog of the sphere where you replace sine by hyperbolic sine in equation (6) passes the same test. There are no others.

Question : Is the torus homogeneous ? Answer : No. The torus is translation invariant but is not isotropic. Therefore it is not homogeneous in the strong sense of everywhere and in any direction the same. Remember what a torus is. Mathematically it is assimilated to a rectangle, such that when we cross a border, we continue from the opposite point, figure 2.

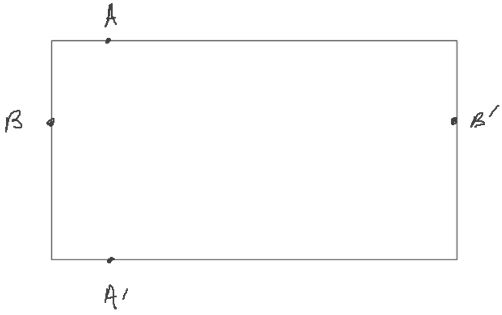

Figure 2 : Mathematical view of a torus.

The horizontal axis and the vertical axis are called the preferred coordinates axes. If we move the origin, keeping the axes parallel, the space looks the same. But if we tilt the axes, for example at 45°, the space won't look the same anymore. For instance even looking far enough you may not see you again.

Now the idea that our universe is isotropic and homogeneous has a status somewhere between being a postulate – therefore highly questionable –, and being an observational fact. On certain scales it really does look homogeneous, but on scales so big that we can see them we don't know.

There is a second remark I want to make before we go on. Since it comes up over and over again, and there appears to be an enormous amount of confusion when we say for example that our space may have a spherical shape, let's clarify what this means and a few related points.

When we talk for instance of a 2-sphere universe, people tend to get the idea in their head that it is really like a balloon, namely that it has an inside and an outside. And they ask questions like what happens if we move away from the balloon or into the balloon.

But we have to learn to view geometries thinking only of their intrinsic shape. The 2-sphere has no inside or outside. It is a two-dimensional space on its own – which happens to be finite and have no borders.

One of the most confusing kind of space is a one dimensional space. It can be infinite or finite, and in the latter case have borders, that is end points, or not. Let's talk of a closed one-dimensional space, also called a loop.

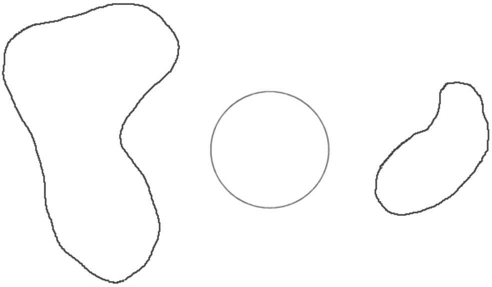

Figure 3 : Examples of closed one-dimensional spaces, all embedded in the page, that is in a 2D flat space. In their own intrinsic topology *they are not curved.* And the two on the right, which have the same length, are intrinsically identical.

From the point of view of the intrinsic geometry, all that counts is measuring distance along the space. In figure 3, the loop on the left is the longest, but the center loop and the right loop have the same length. They are absolutely identical intrinsically. If they appear different it is only because we chose to represent them embedded differently. The renowned little bug living in one of them and then in the other won't see any difference, because for it *there is no difference.*

The only parameter which distinguishes classes of loops is their total length : the distance the little bug has to walk into one direction before returning to its starting point. Had we drawn a square in figure 3, of the same length as the circle, it would also be intrinsically identical.

Question : Are the loops curved ? Answer : No. They are not curved. Their *embedded representations* on the page are curved, but the loops themselves, viewed in their one-dimensionality, viewed in their intrinsic topology, are not curved. If you live in one of these loops, and walk along you never have any sense of turning. This may seem strange. But to think of a turn, or of curvature of a trajectory, you have to think at least in 2D. You may also notice that any segment of a loop, embedded in 2D, can be straightened out, without stretching or changing it, and without changing whatsoever the experience of the little bug living and moving in it.

The right term about a curvy line drawn in 2D is :

The intrinsic curvature is zero, the extrinsic curvature is not zero.

One-dimensional spaces are all flat. They have no curvature. The fact that they appear curved in figure 3 only has to do with the way we drew them in two dimensions. And one dimensional spaces which are not loops, are either infinite or segments of string with end points. The latter have a topology similar to loops, except for the two end points. They are also flat, no matter how we represent them. And infinite one-dimensional spaces, whether we draw them in 2D as straight lines or otherwise, are flat.

So we have to learn to think about the intrinsic geometry of a space. It requires some effort. Then we encounter funny facts like : a billiard table is a donut, and spaghetti are flat !

Let's finish this clarification with a question about 3D : what is it after all that is special about three dimensions ?

Let's answer in steps. Do you think you can visualize a five-dimensional space ? Everybody answers no, and rightly so. What about four dimensions ? Same answer. Apart from some tricks – like for instance thinking about a collection of three-dimensional spaces, mimicking a fourth dimension [3] – we cannot visualize a four-dimensional space.

Now can we close our eyes and in our mind visualise a cube ? Yes. We can view it in our mind's eye. No problem.

Let's continue going to fewer dimensions. Can we visualize a two-dimensional surface (particularly a curved or bumpy one) ? Everybody will say : sure I can visualize the 2D surface. But in truth we only visualize it embedded in 3D, we don't perceive its intrinsic (flat or curved) shape without going into three dimensions.

Finally can we visualize a one-dimensional space ? Same answer : we really visualize it in three dimensions – neither intrinsically in one dimension, nor in two dimensions, since as just said we don't really perceive 2D without embedding them in 3D.

Even an abstract point cannot be imagined without seeing it suspended in three dimensions.

What is it that is special about three dimensions ? Is there something really mathematically special ? No. It is just that our brain architecture evolved for the purpose of navigating

[3] To see a nice natural representation of a four-dimensional space, where all the dimensions are treated on a equal footing, see this lesson in French on a game of tic tac toe in four dimensions https://www.lapasserelle.com/cours-en-ligne/3e_maths/tic_tac_toe/index.html

around in three dimensions. Therefore it's not surprising that our ability of visualizing is hung up at three dimensions. That doesn't mean that three dimensions are in any way special. Of course every dimensionality has its own special features. But three dimensions is not special.

Since we are mathematically sophisticated people, we can get around the 3D perception hardwired in our brain. If we want to discuss two dimensions we write an X and a Y. If we want three dimenstions, we write an X, a Y and a Z. If we need four, we add a W, and so forth. That is the reason why – the joke goes – the most elaborate string theory stopped at 26 dimensions.

One more example about 2D surfaces represented in 3D : Some of them even when appearing curved are intrinsically flat. That is the case when they can be put back flat on a table without stretching or deforming their intrinsic shape in any way. Such surfaces are called developable surfaces.

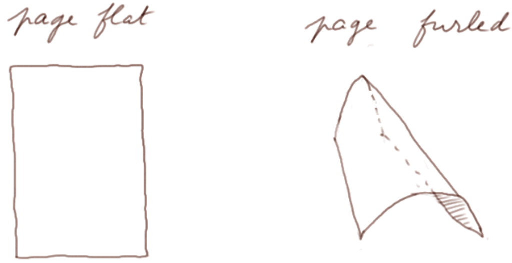

Figure 4 : A sheet of paper remains intrinsically flat, even when furled in 3D.

In figure 4, from the intrinsic point of view, the sheet of paper on the right is no less flat than that on the left. All the relationships within its surface are the same as they were before we furled it. On the other hand, spheres or sections of spheres are not developable. They are intrinsically not flat. The reader will find a more detailed treatment of these questions in volume 4 of the collection *The Theoretical Minimum* on general relativity, the first chapters of which are devoted to geometry.

To close these clarification points, in the coming lessons always remember that when we talk about the geometry of a space, we are talking about its intrinsic geometry and not the way it is embedded for purposes of visualization in some higher dimensional geometry.

Spacetime geometry

Let's accept that space is isotropic and homogeneous, and therefore is one of the three candidate type spaces : spherical, flat or hyperbolic. We give them colorful and imaginative names :

a) spherical : "$k = 1$"

b) flat : "$k = 0$"

c) hyperbolic : "$k = -1$"

The letter k stands for curvature. It is uniform over the whole space when space is homogeneous.

$k = 1$ is the positively curved type of space, that is a sphere. It is not the ordinary sphere. It is the analog of a sphere. It could be a three dimensional sphere, what we called a

3-sphere. And it can have a radius of any length, not only the unit radius.

Let's stress once more : *a 3-sphere is is not a ball with an outer surface.* It is a 3D space on its own right, homogeneous, finite, without any border, everywhere the same, such that if you go straight in any direction you eventually end up where you started. It cannot be embedded in the 3D Euclidean space, since it already has three dimensions. Anywhere locally, it looks like the ordinary 3D space. It could be embedded in 4D, if that helped, but we don't recommend trying. We should think of it intrinsically. Granted, it is not easy to do, but we will get used to it [4].

And again keep in mind when we say the spatial universe is either flat, or spherical, or hyperbolic, and is completely homogeneous, that it is not quite right. It could be true for the average properties of the space over scales big enough to average things out. It is a statement comparable to saying that the surface of the earth is a sphere.

Of course the real surface of the Earth has bumps on it. It has mountains and valleys. So it is certainly not a perfect sphere. However on big enough scale it is a satisfactory approximation. Mount Everest is five and a half miles high. So on a scale of, say, a hundred miles by a hundred miles, the world looks flat. On a scale of a thousand miles by a thousand miles it looks very flat, that is to say with just the smooth uniform small curvature of a huge sphere.

[4] We only have to remember the geometric properties of the surface of the Earth, homogeneous, finite, without any border, everywhere the same. If we go straight in any direction we come back around from the opposite direction. Anywhere locally it is like a plane, etc. And add one dimension while keeping all these properties.

So the idea of whether something is homogeneous or not is a scale dependent idea. Remember that, in the room, we consider that the air is homogeneous, but it is certainly not true at the molecular scale nor even at the millimeter scale.

Of course the Earth did not have to be homogeneous even on scales of a thousand miles or more. It could be cigar shaped. So there is some content in saying that the Earth on big enough scales is homogeneous and isotropic, that is, its surface is a sphere.

The same is true of cosmology. We consider that it is homogeneous, although it is certainly not true at the galaxy scale, nor even at the one hundred million light-year scale. It is homogeneous and it appears to be three-dimensional. It has one the three possible geometries : spherical, flat or hyperbolic, all of them being of the three-dimensional flavour.

We are now going to discuss *spacetime geometry*. In addition to the spatial dimensions and spatial geometry of the universe, we are going to consider a time dimension.

Once we start to talk about curved geometry of course we leave Newton and enter general relativity. That is true even without time, but all the more when looking at a curved spacetime.

Like any geometry, general relativity starts with a metric. And we are going to make an assumption about the spacetime metric : we assume that space and time aren't mixed with each other in the metric, in the sense that in the definition of ds^2 there are no elements like $dtdx$ or $dtdr$. In other words the metric has a form that looks like this

$$ds^2 = -dt^2 + a(t) \begin{cases} dr^2 + \sin^2 r \; d\Omega_2^2 \\ dr^2 + r^2 \; d\Omega_2^2 \\ dr^2 + \sinh^2 r \; d\Omega_2^2 \end{cases} \tag{8}$$

The part after the brace presents the three possible spatial metrics : the metric of the unit 3-sphere, that of the flat 3D space, or that of the unit hyperbolic 3D space. And they all carry the multiplicative scale factor $a(t)$ in front.

We use three dimensional polar coordinates because they are particularly convenient in cosmology. In each case $d\Omega_2^2$ is the metric of the unit 2-sphere. The top line, in equation (8), i.e. the metric of $k = 1$, is also denoted $d\Omega_3^2$. On the second line, $k = 0$, it is simply the metric of ordinary 3D Euclidean space which, of course, can also be expressed in Cartesian coordinates as $dx^2 + dy^2 + dz^2$. Finally on the bottom line, we spelled out what is also denoted $d\mathcal{H}_3^2$.

Remember in chapter 3 the description we gave of $\mathcal{H}_2$, using the Escher's image with devils and angels. We explained that $\mathcal{H}_2$ has a uniform negative curvature, and after stereographic projection we obtained a beautiful anamorphosis, different from that produced by stereographic projection of a sphere. Now we must think of the analog of all this in 3D. In particular the candidate hyperbolic space is no longer $\mathcal{H}_2$, but $\mathcal{H}_3$. Furthermore, in equation (8), there is a scale factor in the metric, and a time dimension in the form of an additional $-dt^2$.

The three metrics look similar. But qualitatively – and quantitatively – they are fairly different. They correspond to different 3D shapes. Only $k = 0$ is easy to imagine.

The scale factor $a(t)$ plays exactly the same role as it did in the expanding 3D Euclidean Newtonian universe we studied in chapter 1. Remember that we introduced a fictitious grid coordinate system, moving with the universe, so that galaxies kept the same grid-coordinates over time. We said that they were "frozen with the grid", except for possible local peculiar motions.

The same is true now with the metrics of equation (8). We can think of a 3D space, spherical, flat or hyperbolic, corresponding to the fictitious grid, with respect to which things are fixed or "frozen", and whose actual size is increasing with $a(t)$. Consider two points with fixed grid-coordinates. Let's call their grid-distance Δ, it can be in the unit 3-sphere, in the usual 3D space – as in chapter 1 –, or in the 3D hyperbolic space. Then their actual distance D is

$$D = a(t)\ \Delta \tag{9}$$

Likewise the velocity between those two points is

$$V = \dot{a}(t)\ \Delta \tag{10}$$

Again we have Hubble's law

$$V = \frac{\dot{a}}{a}\ D \tag{11}$$

$\dot{a}/a$ is the Hubble constant. As we stressed many times, it is a constant in the sense that it is doesn't depend on space. It doesn't depend on the position of the two points or on their distance Δ. But it can depends on time.

This is the basic setup. Now what we do want to do with it ? We want to know more about $a(t)$.

In volume 4 of the collection *The Theoretical Minimum*, on general relativity, we learned about Einstein's field equations, see chapter 9 of that volume. We learned that the calculations to derive them are rather complicated, that even simply writing them down in detail is daunting. So we are not going to go through the calculations. We will only sketch them and write down the general form of Einstein's equations.

How do we derive Einstein's equations ? We write down the metric. We calculate the Einstein tensor, then set it equal to whatever it is supposed to be set equal to. Well, remember that it is supposed to be equal to the energy-momentum tensor. And that yields the Einstein's field equations.

How do we calculate the Einstein tensor ? It requires the calculation of Christoffel symbols, lots of them. Each is computed from the metric, see chapter 3 of volume 4. They include derivatives of all kinds. It is a real nuisance.

In our case there aren't too many Christoffel symbols actually, but it is a drudgery anyway. We are not going to do the calculations here. We will just outline the basic ideas. Then we will use Einstein's field equations to find out not what $a(t)$ is, but what equation $a(t)$ satisfies in each of the three cases of equation (8).

The Einstein's field equation[5], like any equation, has a left-hand side and a right-hand side. The left-hand side, as said, is the Einstein tensor. It has to do with the geometry of the universe. If you don't remember it very well, it is not terribly important for our purpose. It is built out of the metric tensor. Its general form is

$$\mathcal{R}^{\mu\nu} - \frac{1}{2}\, g^{\mu\nu}\, \mathcal{R} \tag{12}$$

$\mathcal{R}^{\mu\nu}$ is called the Ricci tensor. It has two indices. We chose the indices to be upstairs indices. It is a contraction of the Riemann curvature tensor which has four indices. $g^{\mu\nu}$ is the metric tensor. And $\mathcal{R}$ is the curvature scalar. It is a further contraction of the Riemann tensor.

The right-hand side of Einstein's field equation is the energy-momentum tensor. It has to do with the distribution of masses and energy in the universe. And it has the constant $8\pi G/3$ in front. So it is

$$\frac{8\pi G}{3}\, T^{\mu\nu} \tag{13}$$

Equating expression (12) and expression (13) yields the Einstein's field equation, which is one of the most famous equation in physics. It shows how the energy and momentum of material in the universe affects its spacetime geometry, and conversely how the geometry of the universe affects the energy and momentum of its material.

$$\mathcal{R}^{\mu\nu} - \frac{1}{2}\, g^{\mu\nu}\, \mathcal{R} = \frac{8\pi G}{3}\, T^{\mu\nu} \tag{14}$$

[5] We use indifferently the singular or the plural because Einstein's field equations are *one tensor equation*, therefore they are also a bunch of equations, one for each component.

Both sides are tensors. Therefore if equation (14) is true in any frame it's true in every frame. It is a good tensor law of physics. From equation (14) our aim is to get an equation for $a(t)$, for the three types of spaces mentioned at the beginning of the section.

Let's start with the right-hand side. The energy-momentum tensor $T^{\mu\nu}$ contains a complex of things, which include the density of energy, the flux of energy, the density of momentum, and the flux of momentum. They are the different components of the tensor. In particular the time-time component T^{00} of the tensor is the one that we are going focus on. It is the energy density. As previously, let's denote it ρ.

In the tensor equation (14), the equality has to hold componentwise. Since we are focussing on the time-time components, the right-hand side becomes

$$\frac{8\pi G}{3}\ \rho \tag{15}$$

ρ stands for the ordinary energy in matter, whatever kind of matter or material the energy-momentum tensor is describing. $T^{\mu\nu}$ is completely sensitive to the kind of material that is in the universe. Is it particles ? Is it electromagnetic radiation ? Is it something else ? $T^{\mu\nu}$ knows about the material nature of the ingredients that are making up the universe.

The left-hand side, on the other hand, has nothing to do with the material making up the universe. The left-hand side is geometry.

So looking only at the time-time components, equation (14) becomes an equation whose right-hand side is the energy density and left-hand side is something that involves curvature.

We calculated the Riemann curvature tensor in chapter 3 of volume 4. Its forbidding expression was

$$\mathcal{R}_{srn}^{\ \ \ \ t} = \partial_r \Gamma^t_{sn} - \partial_s \Gamma^t_{rn} + \Gamma^p_{sn} \Gamma^t_{pr} - \Gamma^p_{rn} \Gamma^t_{ps} \tag{16}$$

There are two contributions : one which involves second derivatives with respect to the coordinates ; it is the first two terms of the right-hand side ; and the other which involves first derivatives squared, i.e. quadratic things in the first derivatives ; it is the last two terms.

The curvature tensor $\mathcal{R}_{srn}^{\ \ \ \ t}$ is used to compute by contraction the Ricci tensor and the curvature scalar appearing in the Einstein tensor $\mathcal{R}^{\mu\nu} - \frac{1}{2}\ g^{\mu\nu}\ \mathcal{R}$.

Fortunately for us, in the time-time component of the Einstein tensor, only the terms with first derivatives remain. The things which involve differentiating the metric twice actually cancel between $\mathcal{R}^{\mu\nu}$ and $\frac{1}{2}\ g^{\mu\nu}\ \mathcal{R}$. This cancellation is not true for the space-space or time-space components, but it is true for the time-time component. And that is sufficient for us.

So on the left-hand side of equation (14), when we look only at the time-time components, things are strictly proportional to squares of first derivatives, i.e. quadratic things of first derivatives. That is one fact.

The second fact on the left-hand side of equation (14) is that the Einstein tensor has two contributions : one comes from derivatives with respect to space, the other comes from derivatives with respect time.

The things which involve only derivatives with respect to space couldn't care less that there is time in the problem. They are only related to the curvature of space. And we know what it is. The spherical universe, called "$k = 1$", has positive curvature. The flat universe, called "$k = 0$", has zero curvature. And the hyperbolic universe, called "$k = -1$", has negative curvature.

So the curvature of space is one contribution on the left-hand side of equation (14). The other has to do with the way space is changing with time. But the only way in which space is changing with time is through the scale parameter $a(t)$. So there will be a factor involving the time derivative, $\dot{a}$, squared. It is sort of nuisance, but we can calculate it. When we work it out, we find that it is $(\dot{a}/a)^2$.

Let's turn to the term which has to do with the curvature of space itself, and let's think about it for a moment. How curved the space is as a function of its radius ? If the Earth were a thousand times bigger than it is, we would all agree that it would be less curved, at least locally. On the other hand, if it were a marble, it would be smaller, but its curvature would be larger.

The curvature of a space scales in a certain way according to its radius. In fact it is simply one over its radius squared. So it is proportional to $1/a^2$. The proportionality comes with a plus sign if space is a sphere, zero if space is flat, and a negative sign if space is a hyperboloid. In other words, it is the k of their names.

Finally the Einstein's field equation boils down to

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{8\pi G}{3}\,\rho \tag{17}$$

k is just a placeholder that tells us which of the three kinds of spaces we are talking about.

Let's switch k/a^2 to the other side. Einstein equation becomes

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\,\rho - \frac{k}{a^2} \tag{18}$$

This equation is absolutely identical to the Newtonian version. If we think of ρ as the mass density, it is the equation (19) we have already explored in chapter 2, and which bears the name of Friedmann equation.

The only thing new is that we now have a new interpretation of the term $-k/a^2$.

What did it stand for in the Newtonian example ? It stood for the energy : whether the total energy in the universe was positive, zero, or negative. It corresponded to whether we were above, at, or below escape velocity. Notice that above escape velocity corresponds to $k = -1$. At escape velocity is $k = 0$. And below escape velocity is $k = +1$.

So the term $-k/a^2$ is the same exact term as $+C/a^2$, of equation (19) in chapter 2, but it has a different interpretation, namely it now relates to the curvature of space. Same equation, somewhat different physical interpretation.

The reader might wonder how come the general theory of relativity, which contains among other things ingredients of special relativity, does lead to the same Newton equations ?

Basically it is explained as follows. Let's suppose the universe is curved but we only look on a small piece of it, figure 5.

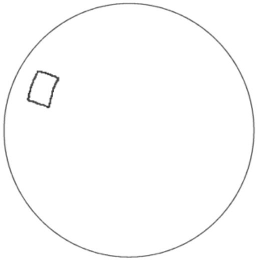

Figure 5 : Looking at the galaxies in a small region of a spherical expanding universe.

We cannot tell that the small piece is curved. And let's look at all the galaxies in that small region. The way they move must the same as in the flat expanding grid we studied in chapter 1. The curvature can't be important for a small enough piece.

Moreover the equations are exactly the same for a small piece or a big piece. Therefore we must somehow reproduce Newton's equations. Anyhow that is indeed what happens. The Newtonian version turns out to be correct even in the kind of spacetime under consideration.

But we have to remember that in Newton's physics ρ stood for the mass density. Setting the speed of light equal to one, energy $E = mc^2$ was just equal to mass. The parameter ρ was the density of ordinary mass – what is still sometimes called the rest mass.

In the derivation of Friedmann equation in the Newtonian framework, as we did in chapters 1 and 2, it was assumed that everything was moving much slower than the speed of light. And if we have a collection of particles, all moving much slower than the speed of light, the energy density is just the density of mass.

Now, on the other hand, Friedmann equation is more general. We sketched its derivation from Einstein's field equations in the framework of general relativity. For example it still holds in situations where particles may be moving fast relative to each other. In fact it even holds if the energy density is due to photons, i.e. massless radiation moving with the speed of light.

The Newtonian equations wouldn't know what to do with photons. The Einstein equations know what to do with photons or radiation. But otherwise the equations look very similar.

The next topic is the equation of state and how it determines information about ρ in Friedmann equation (18).

Equation of state

Friedmann equation involves the energy density ρ.

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\,\rho - \frac{k}{a^2} \tag{19}$$

To work with it we need more information about ρ. What do we mean by that ? We mean information on how ρ itself varies as a function of the scale factor a. There is not much we can do with this equation until we know something more about ρ.

Well, we do know more things about ρ. For example if ρ is just made of ordinary particles that are sitting there and the universe expands, it is quite clear that the density of energy decreases. And we even know how. We studied it in chapters 1 and 2, see for instance equations (19) and (20) of chapter 2. We will write it down again in a moment. If it is radiation, it decreases in another way.

So how ρ depends on a depends on the nature of the material that is making up the energy in the universe. But it is clear that in order to solve equation (19), to even think of it as having any content, we have to know something about how ρ depends on a. If we know how ρ depends on a then it just becomes an ordinary differential equation for a as a function of time.

The reader remembers the two examples we already studied. They have names. One of them is called *matter-dominated.* It simply corresponds to ordinary particles moving slowly relative to the mesh or grid. They are particles which are not moving so fast that we have to worry either about relativity or even kinetic energy very much. If we are standing next to such a particle, its energy relative to us is simply its mass.

It is the case where the energy density ρ is equal to some constant ρ_0 (read *rho naught*) divided by a cubed

$$\rho = \frac{\rho_0}{a^3} \tag{20}$$

We sometimes denoted ρ_0 as ν.

Incidentally, what is the meaning of a ? Let's take the spherical case, that is the case where the universe is a 3-sphere. In this spherical case, the meaning of a is extremely clear : it is the radius of the 3-sphere universe at any given instant.

Remember that it is difficult to imagine a 3-sphere, and therefore the universe as a 3-sphere. It is the analog in three dimensions of the ordinary 2-sphere in two dimensions. The ordinary 2-sphere is for instance the surface of the Earth, thought of intrinsically if we can, i.e. not thinking of it embedded in three dimensions.

Even thought of intrinsically, a 2-sphere has a radius. It is not related to any embedding in 3D. It is an intrinsic feature of the uniformly curved two-dimensional surface. The German mathematicien Carl Friedrich Gauss (1777 - 1855) is the first to have studied systematically the intrinsic curvature of a surface at any point[6]. Similarly a 3-sphere has

[6] Gauss was so pleased with his result, published in 1827, about the curvature of a surface, showing that in any point it is an intrinsic feature of the surface, that he called it the *Theorema Egregium.* Egregium if the latin word for remarkable. Gauss theorem is valid in

a radius, which is very difficult to visualize.

Of course we have to provide some units. Let's say we measure lengths in meters. Then, in meters, *a is simply the radius of the universe.* And what is ρ_0 ? It is just the density of mass in the universe at the time when a was equal to one.

Well, may be we don't want to talk about a universe of radius one meter because it is too small and the spherical model of the physics breaks down at that scale. So let's take one megaparsec for instance as the unit of length.

So ρ_0 is a sort of initial condition which we can change in the equations. ρ_0 is constant but every time we double the radius of the universe, or change the scale of the universe by rescaling a, the density ρ changes as the inverse of the cube power of a, equation (20).

In the flat case which is infinite, $a(t)$ itself has no invariant meaning, nothing to compare it with. It is less easy to think of an infinite 3D flat space inflating, than of a sphere – be it a 3-sphere – inflating. But we can select a time t_0 and a unit, and set $a(t_0) = 1$. This is what we did in chapter 1 when we considered the moving grid accompanying the expansion of the 3D Euclidean universe.

In the round case $a(t)$ is the radius of the whole universe. Also in the negatively curved space there is a natural definition of the radius of a hyperbolic geometry.

spaces of any dimensions and geometries. In the 1820's he carried out experiments with light rays emitted from different hills to check whether the sum of the angles of a triangle was exactly 2π, or any warp in the three-dimensional universe could be detected.

The other case that we talked about was *radiation-dominated.* We saw that then ρ is ρ_0 divided by a to the fourth

$$\rho = \frac{\rho_0}{a^4} \tag{21}$$

We talked at length about the difference between these two models of the universe. The fact is that, if we had a bit of both, like

$$\rho = \frac{c_1}{a^3} + \frac{c_2}{a^4} \tag{22}$$

then early on, in the history of the universe, the second term would likely be dominant because one over a to the fourth is much bigger than one over a cubed. At later times, the first term would take over on the right-hand side of equation (22). We discussed all that.

Switching from a Newtonian universe to an Einsteinian universe changes nothing to these considerations.

There is one new piece of information however. We remember that the sign of k in equation (19) determines whether the universe is going to continue to expand linearly (case $k = -1$), or is going to re-collapse (case $k = +1$).

Incidentally, this will change a little bit when we come to talk about dark energy. But up till now, with the forms of energy given by equation (19), if k is positive that is the spherical case, i.e. the case where the universe is a 3-sphere. In that situation the universe begins by expanding then re-collapses.

If it is flat, then it is as if every galaxy was exactly at the escape velocity. The universe continues to expand forever, though slowing down and slowing down asympotically to zero velocity, in other words asymptotically coming to rest. But it doesn't re-collapse. It is sort of a knife edge case.

The third case is k negative. It corresponds to being above the escape velocity. At late times the derivative of the scale factor stabilizes, and $a(t)$ just continues to increase linearly. It is similar to a stone thrown away from the Earth at a speed above escape velocity : eventually the attraction of the Earth becomes negligible and the stone then moves in a space with no force field, therefore at constant velocity.

So we did not waste our time doing the Newtonian case. The Newtonian case and the Einstein case are very close.

Now we want to understand equation (19) and its solutions even better. This will lead us to the equation of state.

If the only possibilities were matter-dominated and radiation-dominated, we might not care very much to build a more general understanding of equations (19) to (22). What is the point of a general understanding when there is only two cases ? There are however many things in between matter-dominated and radiation-dominated. And not only in between but also more extreme. There is a whole range of possible behaviors beyond those we described. And they are important.

So we want a deeper understanding of the link between the different kinds of material making up the universe and the evolution of the universe. For instance how the energy density would change as a function of the scale factor.

What is it that we need to know ? We are going to proceed in two steps. The important ingredient in determining how ρ varies as a function of a is called the *equation of state.* We will see what it is. Then first we will assume that we know the equation of state, and derive its consequences. The second part will be to derive for different kinds of materials the equation of state.

What do we mean by an equation of state ? It is basically a thermodynamic idea. It is a relationship between various thermodynamic variables describing a system. For our purposes temperature will not play any big role. The important variables will be pressure and energy density.

The equation of state will be a relationship between energy density and pressure. Now for an ordinary gas, made of moving molecules, in the room for example, elementary thermodynamics tells us that there is a connection between the energy density and the pressure : the higher the energy density, the higher the pressure. What energy density are we talking about in elementary thermodynamics ? Basically the kinetic energy of molecules. Of course there there is the mc^2 energy which is dominant by far, much bigger than anything else. But let's forget the mc^2 energy for a moment, just to get the idea.

The kinetic energy of motion of the molecules is proportional to the square of their velocity. The molecules bounce off the walls of the room and exert pressure. It is clear that the faster the molecules move the bigger the pressure on the wall is going to be[7]. So it is clear that there is a connection between energy density and pressure. And usually the way it goes – although there are some exceptions to it – is the higher the energy density, the higher the pressure.

The examples studied in elementary thermodynamics, but in fact which also cover pretty much the ground of interest of the cosmologists, can be described by very simple equations of state. The equation of state that cosmologists study the most is that saying that the pressure is a constant called w times the energy density ρ.

$$P = w\rho \tag{23}$$

In many cases of course, in elementary thermodynamics, it is not really true that the pressure is a strictly linear function of the energy density. Nonetheless, more or less by accident, the interesting cases in cosmology have the form of equation (23). We shall see presently what w is for the two cases of interest, matter-dominated and radiation-dominated universe. Later we will come back and derive it.

Now, pursuing the elementary thermodynamics simile, the matter-dominated case is the case where the particles are moving very slowly compared to the speed of light. Their total energy is mostly the mc^2 energy. Their kinetic energy is negligible. So in first approximation we can say the molecules are at rest in the room. And if the molecules are at rest in the room, then the pressure on the wall is zero. So for the matter-dominated world the pressure is equal to zero,

[7] We use the equation $Fdt = mdv$ applied to molecules hitting an area S of the wall to establish it.

or very very small compared with their total energy density.

Let's go over the analogous reasoning once more : in elementary thermodynamics we have the relationship $P = w\rho$, where ρ is the ordinary energy density of molecules. Then we say that the same relationship holds in cosmology with the total energy density ρ and a pressure P still resulting only from kinetic energy. In the matter-dominated universe the kinetic energy density is essentially zero compared to other energies making up ρ. So in the matter-dominated universe we have $w = 0$. And the equation of state is $P = 0$.

The harder case is the radiation-dominated universe. We won't prove it here, but in the radiation-dominated universe, $w = 1/3$. And the equation of state is $P = \rho/3$. It turns out that w is the inverse of the number of dimensions of the universe. That is why it is 1/3.

Now what does the equation of state have to do with anything ? What we are going to do is use the equation of state to derive how the energy density changes as a function of the scale factor. Again we will use the simplest kind of elementary thermodynamics identities.

Supposing we have a box containing material of some sort, see figure 6. It could be gas, it could be liquid, it could be molecules of a solid, it could be radiation. Whatever it happens to be doesn't matter.

Figure 6 : Box containing material.

The box has some energy E in it. We are talking about the total energy. We can write that this energy is equal to the energy density times the volume of the box

$$E = \rho V \tag{24}$$

Now let's imagine changing the volume of the box a little bit. It doesn't really matter whether we change the volume of the box sort of isotropically in all directions, or we just move, say, the right wall. Then how much does the energy change ? We have the simple mechanical relation

$$dE = -P\ dV \tag{24}$$

When the material inside the box pushes on the walls, the pressure does work, which is equal to a diminution of energy inside. Or, equivalently, the increase of energy is the opposite of that work. That is all we need to use. It is the basic identity we will use in working with the equation of state.

Actually $-P\ dV$ is only mechanical energy. In thermodynamics there is another term on the right-hand side of equation (24). It is TdS, the temperature times the change in entropy. The complete thermodynamic identity liking internal energy, pressure, volume, temperature and entropy is

$$dE = -P\ dV + T\ dS \tag{25}$$

However, when variations are slow in the gas – and the universe is expanding slowly by comparison with any other kind of time scale – the entropy doesn't change. It is called an adiabatic change. See volume 6 of the collection *The Theoretical Minimum*, on statistical mechanics, for a detailed treatment of the subject.

So for our purposes the change in entropy is zero. Therefore equation (24) holds for the universe and for boxes expanding like the universe.

Since $E = \rho V$ we can also calculate dE by applying the rule of differentiation of a product.

$$dE = \rho\ dV + V d\rho \tag{26}$$

Both ρ and V are changing. Moreover changing the volume will change the energy density in some way, as yet unspecified. Equations (24) and (26) enable us to write

$$\rho\ dV + V d\rho = -P\ dV \tag{27}$$

Let's regroup the terms with dV on the right-hand side.

$$V d\rho = -(P + \rho)\ dV \tag{28}$$

Now if we didn't know any connection between P and ρ, we would be stuck. We wouldn't know what to do with this equation. But let's assume that P is known in terms of ρ. In particular, let's take the very simple case where $P = w\rho$. Equation (28) becomes

$$V d\rho = -(1+w)\ \rho\ dV \tag{29}$$

Now we divide both sides by ρ and V.

$$\frac{d\rho}{\rho} = -(1+w)\ \frac{dV}{V} \tag{30}$$

For any positive function f, remember that

$$\frac{df}{f} = d(\log f)$$

So equation (30) can be rewritten

$$d \log \rho = -(1+w)\ d \log V \tag{31}$$

Integrating both sides we get

$$\log \rho = -(1+w)\ \log V +\ c \tag{32}$$

or equivalently

$$\rho = \frac{c}{V^{(1+w)}} \tag{33}$$

It is not the same constant c, but it doesn't matter.

Now, the volume of the box can be expressed in term of the scale factor. If we think of the box as expanding with the general expansion of the universe, its volume is proportional to a^3. So we reached the expression of ρ as a function of a which we were seeking.

$$\rho = \frac{c}{a^{3(1+w)}} \tag{34}$$

The constant c is still a generic constant which we will worry about another day. And again we can just say that, in equation (34), the constant c is equal to the value of the energy density when $a = 1$.

Equation (34) is what the thermodynamics of a nice homogeneous material would tell us. That is how ρ varies with the scale factor.

Let's see what equation (34) says for the various cases of the equation of state $P = w\rho$.

Suppose w is equal to 0. That is the case of matter dominance. Then equation (34) just says ρ is just equal to a constant over a^3. That is equation (20) already seen.

What if w is equal to one-third ? That is the case of radiation dominance. Then the exponent of a in equation (34) is $3(1 + 1/3) = 4$. And ρ is equal to a constant over a^4. That is equation (21) above.

So we see that there is a general framework in which $\rho = \rho_0/a^3$ or $\rho = \rho_0/a^4$ emerge. And really all we need to know is w, which describes a cosmology based on some sort of energy. What we need to know is the equation of state, in the form pressure is equal to w times energy density, and very little more.

This is a natural place to end the lesson. So we will just finish it with a questions / answers session.

Questions / answers session

Q. : When we went from the metric to Einstein's field equation, then to Friedmann's equation, we made the assumption that, in the metric, the space part and the time part are separate, in other words, there is no cross term like $dtdr$ for instance. Is this a physical restriction on what the universe can be ?

A. : No. It is just a consequence of isotropy and homogeneity. If the universe is isotropic and stays isotropic, and homogeneous and stays homogeneous, then there is really no alternative. It can be proved.

Q. : From the Einstein's field equation (14), we decided to work on the time-time component, and we eventually reached the Friedmann equation. What would choosing another component of the tensor equation produce ?

A. : Instead of the time-time component, we could have chosen the space-space component. What we would get would again be like Newton, except, instead of being the energy equation of Newton, it would be the $F = m\ddot{x}$ equation.

Remember that Friedmann equation, in the Newton framework, is really the energy equation, see equation (13) of chapter 2. So, like the Newton energy conservation, it involves only the first derivative $\dot{a}$ of the scale factor a. Whereas remember that $F = m\ddot{x}$ involves second time derivatives.

There are linear combinations of the space-space and time-time equations, in equation (14) above, which, instead of looking like the energy equation, look like $F = m\ddot{x}$. They also involve second time derivatives, and lead to $\ddot{a}$ appearing in the equation for $a(t)$. But the point is that there are all equivalent, just like $F = m\ddot{x}$ and energy conservation are equivalent. And they better be all equivalent.

Why is it that we don't need more than one of them ? Well, for the simple example there is only one function to calculate. It is $a(t)$. In a more general context, where geometry may be wavy and fluctuate and do other things, we may need all of the equations because there would just be a lot of functions to compute.

Notice though that we could not have picked the mixed space-time equations from equation (14), because this would just lead to zero equals zero.

Q. : In the radiation-dominated case, what is the radiation ? Is it the cosmic microwave background ?

A. : Yes, almost all of it is the microwave background radiation, by an overwhelming factor. There are other photons, of course : sunlight, starlight, etc. But they amount for a tiny fraction of all the radiation in the universe.

Furthermore, at the present time, the cosmic microwave background is very very small compared to the energy in ordinary atoms, which in turn is a somewhat small fraction of the dark matter.

In the standard model of the universe, today ordinary matter, protons, neutrons, atoms and so forth, is about 5% ; dark matter amounts to about 30% ; and the approximately two-thirds remaining is dark energy. Radiation is only a billionth of all this today.

If we run it backward into the past, at some point ρ_0/a^4 becomes bigger than ρ_0/a^3 and it becomes radiation-dominated. So the early universe was radiation-dominated.

Late universe, today, if it were not for dark energy, that is if it were not for the cosmological constant, which we shall study, would be matter-dominated.

And again, if we didn't worry about the cosmological constant, we would say that the parameter k, in Friedmann equation (18), is correlated with the future history of the universe, determining whether it collapses or continues to expand.

Now I emphasize that the model of expansion or re-collapsing, resting simply on $k = -1$, 0, or $+1$, in equation (18), is wrong because of one other ingredient missing. And that is *dark energy.*

We have gone far enough that I can say a little bit about dark energy, if we don't care where the equations come from. The cosmological constant CC, or dark energy DE, or vacuum energy VE are all different names for the same thing. And that thing is w.

A model universe with dark energy is a model universe where $w = -1$, that is whose equation of state is simply $P = -\rho$. It is a bit odd that the pressure and the energy density should have the opposite sign.

Usually pressure varies in the same direction as energy density. Yet we can think of ordinary physical systems where pressure and energy density have opposite sign. Think for instance of the box of figure 6, and for simplicity think of it in one dimension. Suppose the walls of the box are tied by a spring, as in figure 7.

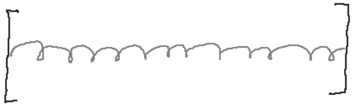

Figure 7 : System where pressure and energy vary in the opposite direction.

In figure 7, let's think of the energy density as the potential energy, and of the pressure as the opposite of the tension. Then pressure is negative. When pressure is negative, it is called tension. Since the potential energy is proportional to the elongation of the spring, therefore, by Hooke's law, to the tension, we see that the energy density varies like the pressure but with the opposite sign. In other words, increasing the potential energy makes the pressure more and more negative.

So it is possible for the energy density to go up when the pressure goes down. That is what $w = -1$ means.

Why such a tension should exist ? That is another question.

But let us just examine its consequences. In equation (34), reproduced below,

$$\rho = \frac{c}{a^{3(1+w)}}$$

if $w = -1$, what do we get ? We get ρ equals constant. That is the nature of dark energy : it does not change when you expand the size of the box. When ρ is constant, changing the size of the box changes the energy in the box, but it doesn't change the energy density.

How to explain that the dark energy density in the box doesn't depend on how big the box is ? It is because that dark energy density is a property of *empty space*. And empty space doesn't dilute when you stretch it.

In the next lesson, we will examine more critically the reason why w can have the various values we saw. And then we will study the behavior of the universe under the various kinds of conditions. In particular we will study what happens if we have dark energy. That raises new things that we haven't seen before.