Lesson 8 : Baryogenesis

Notes from Prof. Susskind video lectures publicly available on YouTube

Introduction

Baryogenesis means the creation of matter particles. Its study is more specifically concerned with explaining the excess of matter over antimatter in the universe.

The conditions which explain the imbalance of particles and antiparticles are called nowadays the Sakharov conditions [1].

In the previous chapter, we studied several landmarks in the temperature of the universe. Going backwads in time, we met the decoupling time when the universe become transparent. The temperature was about 4000°.

Further back at $T \approx 10^{10}$ degrees, photons could create electron-positron pairs. Protons had already been created earlier.

Even further back at $T \approx 10^{12}$ or 10^{13}, photons could create proton-antiproton pairs. More accurately they could create quark-antiquark pairs, because in truth it was too hot for protons to exist as such. But, at that temperature, let's say by convention that when we talk about a proton we really talk about three quarks. It does not make any difference in the reasonings.

1. Named after Andrei Sakharov (1921 - 1989), Russian nuclear physicist. Sakharov published his work on this in 1967 in Russian and it remained little known until the 1980s. In the meantime, the same conditions were rediscovered in 1980 by Savas Dimopoulos and Leonard Susskind who were trying to answer the question of why the number of photons, which is a measure of the entropy of the universe, exceeds by eight orders of magnitude the number of protons or electrons. Thus for a time the conditions were known in the West as the Dimopoulos-Susskind conditions.

So at that early time, when the temperature was, say, 10^{13} degrees, the universe was a hot soup of photons, protons, antiprotons, electrons and antielectrons, also named positrons. Their numbers [2] are denoted respectively

$$N_\gamma \quad N_p \quad N_{\bar{p}} \quad N_e \quad N_{e^+}$$

There was a thermal equilibrium and these five numbers were of the same order of magnitude.

But they were not exactly equal. There was a slight excess of protons over antiprotons, and a slight excess of electrons over positrons. Furthermore, the universe being electrically neutral, these two excesses exactly compensated each other. So we had

$$N_p - N_{\bar{p}} = N_e - N_{e^+} \tag{1}$$

As the universe expanded and progressively cooled, the average energy of photons decreased. First the creation of proton-antiproton pairs ceased. But since the expansion and cooling of the universe was relatively slow, the annihilation of protons and antiprotons could continue until there remained essentially only the excess protons. They are the protons we see in our universe today.

A little later, when the temperature went below 10^{10} degrees, the same process happened for electrons and positrons. The creation of electron-positron pairs ceased, but

2. Notice that, although we don't make it explicit with the notations, these numbers depend on time. When the temperature cooled N_p and $N_{\bar{p}}$ decreased, but their difference stayed fixed. The same comment applies to N_e and N_{e^+}.

the annihilation of electrons and positrons continued, eventually leaving only the excess electrons. They are the electrons we see today.

Then there was a period during which electrons and protons began to bind into atoms. But photons were energetic enough to maintain a part of the electrons and protons separated. In other words, the universe was ionized. In turn, the charged particles were still scattering photons and maintaining a thermal equilibrium.

Finally, when the temperature decreased to about 4000°, decoupling took place. Photons could no longer ionize atoms. The scattering of photons stopped. The universe became transparent. And the photons were frozen into a blackbody thermal spectrum, even though they were no longer being scattered. Moreover they continued to cool as $a(t)$ continued to increase, but in a separate process from matter particles.

Over the period from the early soup of photons, protons, antiprotons, electrons and positrons until today, the numbers that did not change are

$$N_\gamma \qquad N_p - N_{\bar{p}} \qquad N_e - N_{e^+}$$

More accurately, the first stayed about the same order of magnitude from then until now. And the last two are also simply N_p and N_e of today's universe.

Today we observe that

$$\frac{N_p}{N_\gamma} \approx 10^{-8} \tag{2}$$

Therefore when the temperature was high enough for photons to create protons and antiprotons, we had

$$\frac{N_p - N_{\bar{p}}}{N_\gamma} \approx 10^{-8} \tag{3}$$

or equivalently, since N_γ was of the same order of magnitude as $N_p + N_{\bar{p}}$,

$$\frac{N_p - N_{\bar{p}}}{N_p + N_{\bar{p}}} \approx 10^{-8} \tag{4}$$

A natural question is :

Why is there today so much more photons than protons or electrons ?

But the question is really upside down. Rather than why there is so much photons, it should be why there is so few protons and electrons.

There are basically three conditions, stated by Sakharov, such that if they are satisfied – and they are satisfied – then there will be an imbalance of matter over antimatter.

Then, once we explain the imbalance of matter over antimatter, the real question becomes the magnitude of it.

Let's pause to make a general comment on the questions we ask in cosmology, because it often comes up. Generally speaking we might wonder : is what we are asking a legitimate question ? Perhaps the answer is simply : well, the world started that way, end of story.

However, it focuses our attention a lot when we have a number – in this case 10^8 photons per proton. This number needs an explanation. The excess in itself was a fact about which we could have said : that is just the way it is. But once there is a number, and in particular if the number is some oddball figure like ten to the minus eighth, we start to think maybe this needs an explanation.

Sometimes the question is formulated as : why is the *entropy of the universe* today so large compared to the number of protons ? The reason is that roughly speaking, for a blackbody thermal spectrum, the entropy is simply the number of photons. The blackbody radiation carries entropy. It is thermal. And up to some proportionality factor, the entropy of blackbody radiation is just the average number of photons in a gas.

When people speak about the entropy of the universe, in many contexts what they are talking about is simply the number of cosmic microwave background photons. That is what they mean by it.

So, as explained in the last lesson, and recalled above, in the very early universe, when the temperature was a thousand billion degrees, there was a huge number of protons and a huge number of antiprotons. And they were basically equal to each other. Or, multiplying by three, the number quarks and the number of antiquarks were approximately the same.

The same is true for electrons and positrons, their numbers being approximately the same. And all these numbers were equal to the number of photons. In other words,

$$N_\gamma \quad N_p \quad N_{\bar{p}} \quad N_e \quad N_{e^+}$$

were all about equal.

The protons, antiprotons – or more accurately quarks and antiquarks –, electrons, positrons and photons, were in thermal equilibrium at some very high temperature.

Then the universe cooled, fixing first of all N_p, because the photons could no longer create quark-antiquark pairs, and then fixing N_e for similar reasons. So, when the universe cooled, the photons were leftover.

They eventually decoupled from matter because, when protons and electrons bound up to form atoms, the particles of matter became neutral and no longer scattered radiation. The universe became transparent. The photons just hung around, and it is those we see today in the CMB.

Let's emphasize : before the present situation, the protons and antiprotons annihilated each other. Somewhat later so did electrons and positrons – the usual name for antielectrons. The universe expanded fairly slowly so there was plenty of time for them to find each other and by and large annihilate each other.

All that was left over was a slight excess of protons, with positive charges, and correspondingly of electrons, with negative charges.

So from "why are there so many photons ?", the question became "why is there this tiny excess of 10^{-8} protons, when

measured in terms of the number of photons ?" Why isn't it just zero.

Already in the very early universe we had

$$\frac{N_p - N_{\bar{p}}}{N_\gamma} \approx 10^{-8} \tag{5}$$

The number of photons, N_γ, being approximately the same at that time, in order of magnitude, as the number of protons plus antiprotons, $N_p + N_{\bar{p}}$, equation (3) can be rewritten as

$$\frac{N_p - N_{\bar{p}}}{N_p + N_{\bar{p}}} \approx 10^{-8} \tag{6}$$

So there is a number to compute. Why does this small number appear ? Where does it come from ? Is it possible, within a theory, to compute it from more fundamental principles ?

We don't know the reason for the number. That is because we don't have a complete theory. But almost any theory that we write down, which explains it, gives a small number.

So we will talk about the kinds of conditions that are necessary – and, it turns out, sufficient – to make an imbalance of matter over antimatter.

Now, is it matter of antimatter ? How come it didn't come out antimatter over matter ? That is largely a definition. The thing we call matter is the thing we are made out of.

Baryonic excess

Let's begin with a hypothesis which is really believed to be wrong, but we will come to it. It concerns the baryon [3] number. What does the baryon number mean ?

If there are only protons, the *excess baryon number* in the world can be defined in terms of quarks and antiquarks. It is the number

$$B = \frac{1}{3}(N_q - N_{\bar{q}}) \tag{7}$$

where N_q is the number of quarks in the world, and $N_{\bar{q}}$ the number of antiquarks.

Why this factor $1/3$? Because originally the baryon number was defined in terms of protons and neutrons, and not in terms of quarks. One proton was given the baryon number 1, and so was one neutron. Then physicists established in the 1960s that a proton and a neutron were each made out of three quarks. So one quark received the baryon number $1/3$.

There are other kinds of objects that carry baryon numbers besides protons and neutrons. But they are all unstable. Even the neutron is unstable. Nevertheless there exist other kinds of objects. We can mostly focus by thinking about quarks themselves if we like.

The statement that there is a baryonic excess is the statement that there were in the early universe more quarks

3. Baryon means matter particle, or heavy particle. It is built from the Greek word *barýs* which means *heavy*.

than antiquarks. The numbers of quarks and antiquarks separately were about the same as the number of photons [4]. So the question is how it got that way ?

Let's suppose for the moment that baryon number is like electric charge. One of the things about electric charge is that it is conserved. It doesn't change with time.

Now baryon number is not really like electric charge. Electric charge is the source of Coulomb forces [5]. Long-range electric fields which create long-range electrostatic forces. Baryon number itself is not a source of any kind of Coulomb type force.

Of course the protons are electrically charged. Therefore they make conventional Coulomb forces between each other. They make electric fields. The neutrons are not electrically charged. They don't make electric fields. So what we would say is that it is the charge of the proton, not the baryon number of it, which is creating any kind of long-range field. And baryon number itself may truly be concerned. But it is not exactly like electric charge. It doesn't exhibit this tendency to make long range forces.

So, suppose it is conserved. Then if there was ever an excess, let's say in the beginning of the universe, whatever that means, then there will always be an excess. And that excess will be sort of frozen in. If you change the number of

4. We are talking about orders of magnitude. Therefore the number of protons and the number of quarks can be both about "equal" to the number of photons, even though there are three times as many quarks as protons. In orders of magnitude they are the same.

5. Named after Charles-Augustin Coulomb (1736 - 1806), French physicist best known for his study of electrostatic forces.

quarks, you must change the number of antiquarks by the same amount, if baryon number is conserved.

Moreover, experimentally baryon number appears to be highly conserved. Nobody has ever seen a proton disappear. We can talk more about experiments which search for the decay of protons and so forth. But in first approximation in our world protons are extremely stable.

Suppose the proton was to decay. What could it decay to ? It must decay into things which are lighter than itself. And it must decay to something which has a positive electric charge. So, if we want to assume that whatever it decays to is stable, there is really only one thing that it could decay to. It is a positron and something electrically neutral.

A proton could disappear and become a positron. That conserves electric charge. But it doesn't conserve energy. A positron is much much lighter than a proton. But it would compensate by giving off a neutral particle. What kind of neutral particles are around ? Photons. So a photon is a prime candidate.

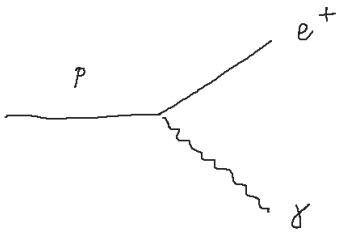

Figure 1 : Possible decay of a proton.

Thus a decay possibility for the proton would be a photon, γ, and an electron antiparticle, that is a positron, e^+. Let's draw it as a kind of Feynman diagram, figure 1. That is a possible thing that could happen. And we don't really know any deep fundamental reason why it can't happen.

Maybe it does happen. We are going to talk about whether it happens. In fact we think it happens, but for whatever reason, as mentioned, nobody has ever observed the decay of a proton yet.

Questions / answers session

Question : Couldn't the proton decay into a positron and a neutrino ?

Answer : No, it cannot decay into a positron and a neutrino. That is no good. The reason is that a neutrino is a fermion. So is a positron. And two fermions make a boson.

However the proton is a fermion. So whatever the proton decays to, it must decay into something with an odd number of fermions.

Q. : If the baryon number was conserved, it would correspond to a symmetry of some sort. Is there a natural candidate for that symmetry ?

A. : Yes. Given any conserved quantity, you can always make up a symmetry [6]. But then you can ask : well, alright, is it really a symmetry ? And the way to test it, is to ask whether the baryon number is conserved. So it is a little bit circular.

But yes, if baryon number was conserved it would correspond to a symmetry. And if it is not conserved, it doesn't correspond to a symmetry.

Q. : Is particule annihilation, like a proton and an antiproton annihilating each other into pure radiation energy, a decay ?

A. : No, we don't call that a decay. It is not the decay of a proton. It is just called *annihilation* of a proton.

But the important point here is that the baryon number doesn't change. Three quarks and three antiquarks came together. Before the annihilation the sum of the baryon numbers is zero [7]. And afterwards it is just a bunch of photons, whose baryon number is still zero.

6. A symmetry is a differentiable transformation of the degrees of freedom describing a physical system which leaves the Lagrangian invariant. See chapter 7 of volume 1 of the collection *The Theoretical Minimum* for Noether's theorem about the relation between symmetries and conservation laws. For example, the conservation of electric charge corresponds to global gauge invariance of the electric field.

7. If a particle has baryon number n, its corresponding antiparticle has baryon number $-n$.

C, P and T and Sakharov first condition

So for whatever reason – and we don't really know the reason – we haven't observed the decay of a proton as shown in figure 1.

Incidentally, the standard model does permit such a decay. In fact it not only allows it, it insists on it at some level.

Now, one thing we know about it immediately is that the decay rate of the proton cannot be very large. Our protons have been around for 13.8 billion years and they haven't disappeared on us.

If the decay time of the proton were a microsecond or even faster, whatever particle physics might permit, that proton just wouldn't be here anymore. So the lifetime of the proton is at least very very long.

It fact the lifetime of the proton is much longer than is necessary for them to still be here. The mean lifetime of the proton is more than 10^{32} years. So it is very long. And we don't completely understand why. In fact we don't understand why at all. But we take it as a fact, at least a temporary fact.

So let's assume for the moment that the proton doesn't decay. That implies that the total baryon number in the universe – counting antiquarks as negative –, or let's just say in box, doesn't change. Therefore, the only theory of the excess B, in equation (7), that we could have is that it was built in from the very beginning. And it still survives today.

That would be a consequence of baryon number conservation. We are not saying that it is true. We are saying that it is a consequence of baryon number conservation, that B is the same today as it was at the very beginning.

Now there are other transformations which *were thought* to be symmetries of nature. They are called respectively C, P and T, and stand for the following

- C = *Charge conjugation*
- P = *Parity, or mirror reflection*
- T = *Time reversal*

To say that C is a symmetry is the statement that if particles and antiparticles are interchanged, we cannot tell the difference. Of course we can tell the difference between a proton and an antiproton, because we are made of protons. But suppose I was made of protons, and you were made of antiprotons – don't get too close to me –, and somebody showed me a proton. I would say "yes, that is a proton". And if somebody showed you an antiproton, you would have exactly the same response, in the sense that you would say "yes, that is the particle I'm made of". You would see the same thing that I saw.

P has nothing to do with economics or fairness. It is mirror reflection. If there is a particle which exhibits an orientation, like a corkscrew spinning clockwise when it moves forward, or like a person with a long right arm and a short left arm, then it had been assumed that there would be a particle with the opposite orientation, i.e., with the corkscrew image, spinning counterclockwise when it also moves forward. The statement that P holds is also called left right

symmetry, or simply parity.

Finally T, the last putative symmetry that we are concerned with in this lesson, corresponds to running the film of nature backwards. Consider any process that can happen in nature. We can even restrict ourselves to laboratory processes, and not worry about the whole universe. If we take a movie of it and run it backwards, it should still be a process which is physically possible, which can happen.

These three transformations were thought to be, naturally, symmetries of nature, until Cronin and Fitch [8] discovered CP violation in 1964. That was only three years before Sakharov put out his theory of baryogenesis.

You could object that the time reversal T doesn't seem true in the real world of course. The second law of thermodynamics tells us that things get only worse. But that is true only in probabilities. It has to do with averaging over large numbers of possibilites. At the true microscopic level, we don't average things out. So any process that happens in nature was thought to be such that its time reversal was another possible process.

Now, there is a mathematical result in quantum field theory which states that the combination of C, P and T is a symmetry. It is not an axiom or a fact from observation, but a theorem – which we are not going to prove in this lesson – that was already known before the 1960s. It follows from the basic structure of quantum mechanics and relativistic field theory.

8. James Cronin (1931 - 2016) and Val Fitch (1923 - 2015), American particle physicists.

This theorem doesn't say that charge conjugation – the particle-antiparticle interchange – is a symmetry. It doesn't say that parity reflexion is a symmetry, nor that time reversal is a symmetry. But it says that the product of all three of them, CPT, is a symmetry. And this is for fundamental mathematical reasons. It would lead to mathematical inconsistencies if it was not a symmetry.

What does CPT symmetry mean ? It means that if we take any process, replace every particle by its antiparticle, reflect it in a mirror and run it backward [9], it is still a possible process in nature. We have to do all three in order to be sure that it is a symmetry.

Let's go back to the time when the three transformations were thought to be symmetries. And, since people focussed mostly on that, let's consider the product CP. It means changing every particle to its antiparticle and changing the orientation of space to its reflexion in a mirror. So CP was thought to be a good symmetry of nature.

If you believe in that, then you might ask : Well, gee whiz, if there is a complete symmetry between particles and antiparticles – after the change of orientation too, but that went along for the ride –, why should it be that at the very beginning there was an imbalance of one versus the other ?

Now, nobody can tell you that there isn't an imbalance

9. When we change the sign of t we also have to take the complex conjugate of the wave function representing the time evolution of the state-vector of the system. This can be seen for instance with the time-dependent Schrödinger equation. See chapter 4 of volume 2 in the collection *The Theoretical Minimum*, on quantum mechanics.

that just dates back to the very beginning. But you might also wonder : What is going on here ? The laws of physics seem to be completely symmetric between the two kinds of things, particles and antiparticles, and yet for some reason there was this small imbalance of size 10^{-8}. That doesn't sound right.

The modern theory of baryogenesis begins with the idea that there was a balance. It says that particles and antiparticles were balanced – again, not for any good reasons. It just assumes that, in the initial conditions we started with, there was no bias toward particles or antiparticles. It can actually be justified in some frameworks, but we are not concerned with that.

Then how is it possible that it got into an imbalance ? The only way it is possible for it to get into an imbalance is if the conservation of baryon number is not correct. In other words, if processes can happen in nature in which a proton becomes a positron, that is a violation of baryon conservation which allows the baryon number of the universe to change.

It would be the first requirement for a theory of baryogenesis, based on the assumption that the starting point was balanced between the two, explaining how we are going to wind up with an excess of quarks over antiquarks, or baryons over antibaryons – baryons meaning protons and neutrons. We must have a mechanism which violates the conservation. So that was Sakharov first condition :

1) *Baryon number violation.*

Violation means violation of the conservation law. And the

process of a proton becoming a positron and a photon is an example, if it happens in nature.

Let's comment on this first condition. If baryon number conservation is not a good conservation law of physics, then it must be of very very weakly broken one. As we said, protons are very old. They didn't disappear, so they must be very old. So whatever mechanism to change baryon number, such as the kind of decay shown in figure 1, it must have an extremely small probability per unit time.

In fact every current fundamental theory, unified or not, if it couples to gravity, violates baryon conservation. Therefore you can ask : In the known theories, why is the proton so stable ? The answer goes something like this. Any theory has a Feynman diagram explaining the decay of the proton as in figure 2.

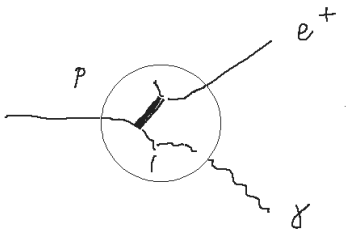

Figure 2 : Decay of a proton, with intermediate particles, and in particular a transient heavy mass M.

The proton comes in from the left, meaning from early. Out goes a positron and a photon. But somewhere in the guts of the Feynman diagram there are all kinds of particles. Let's not specify exactly which ones are in there.

But among them, it is assumed that there is one or more particles which are very very heavy.

In other words, the decay of the proton requires a particle type which is very very heavy in the process – particles which have not been discovered yet.

One of the reasons to believe this is that the standard model by itself, with its ordinary known particles, does not permit this decay to happen. So in order to make it happen you would have to have new additional particles that were not part of the standard model. And that certainly means that they are heavier, because they haven't been discovered yet.

But imagine making them very heavy. For instance 10^{16} GeV, that is 10^{16} heavier than a proton, which is not an unnatural number for heavy particles. Then what is true is that the kind of process in figure 2 is extremely unlikely. By extremely unlikely we mean that the quantum mechanical amplitude [10] for it is squashed by inverse powers of the heavy mass.

Let's call the heavy mass M. Then the Feynman diagram will contain a factor $1/M^2$, just because it is so hard to make a heavy particle.

The heavy particle won't last very long because it has to melt into the rest of the Feynman diagram. But, irrespec-

10. The reader is invited to go back to volume 2, in the collection *The Theoretical Minimum*, on quantum mechanics, to review what is the amplitude of a given observable state for a system, and its relation to probabilities.

tive of its lifetime it has to exist for a wink during the decay, and, if the particle is heavy enough, then the decay will have a very very small probability. That is how modern unified theories explain the stability of the proton.

We don't know if it is right. But it is a tentative explanation. The mechanism requires transient extra particles that are sufficiently heavy to squash the probability for the decay to happen.

Now let's take a second look at the factor $1/M^2$ which reduces the amplitude and the probability of decay. If for some reason the proton is given a lot of energy E somehow [11], then, without going into the quantum calculations about the Feynman diagram, the factor becomes

$$\frac{1}{(M-E)^2} \tag{8}$$

In other words, it may well be a much smaller denominator dividing the amplitude. We will see in a moment where that extra energy comes from, or can come from.

In the ordinary world, when we look at a proton sitting around, there is nothing to give it that enormously high kick E. It doesn't have a huge amount of energy. Protons, when they collide or anything else, don't have huge amounts of energy. That is the reason why we begin with $1/M^2$. And theories of this type can explain why the protons are stable.

Of course you could object that they don't explain anything. They just formulate the hypothesis that transient

11. That is an extra energy on top of its rest mass energy Mc^2. And in our calculations, we have chosen units such that $c = 1$.

particules, which are necessary for the proton decay, must be very heavy.

As we said, all known theories – and I will venture to say any theory that ever will exist – allow for the possibility of baryon violation.

So let's agree tentatively that baryon violation is not taboo, that processes such as shown in figure 2 can happen, they are not forbidden by any fundamental law of physics. It is just an accident of the particle spectrum that the proton is as stable as it is. Let's take that as an assumption, or a working hypothesis.

Baryon violation by itself is not sufficient to give us an average excess of protons over antiprotons. Why not ? Because for every process like in figure 1, which reduces the baryon number in the world by 1, since there is 1 baryon unit on the left and 0 on the right, there is the *charge conjugate process*, figure 3.

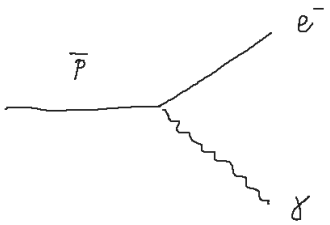

Figure 3 : Charge conjugate process of the proton decay.

An antiproton comes in. A true electron goes off. And the antiparticle of the photon is the photon itself. And that

charge conjugate decay cancels a -1 baryon, that is to say, eliminates an antiproton, of baryon number -1, to replace it by an electron and a photon, both of baryon number 0.

So, if we believe in particle-antiparticle symmetry, then, for every process like in figure 1, a process like in figure 3 will happen with the same probability. And, on the average, there will be as many proton decays as antiproton decays.

But as a consequence, if we started out with an equal population of quarks and antiquarks, of baryons and antibaryons, and we tried to rely only on the violation of baryon number, it would not be a very efficient way to create an excess.

Incidentally, in the very early universe, there were lots of electrons, positrons and photons around. And it was also perfectly possible for the opposite processes to go. A positron and a photon can come together and make a proton. An electron and a photon can come together and make an antiproton. So there is a statistical balance of things going on.

It is only statistical that there would be as many protons and antiprotons, because these processes are statistical processes.

But the imbalance of protons over antiprotons, that we measure, cannot be a statistical effect. The impossibility comes from an elementary argument in probability theory. Let's go through it.

The baryon excess is about 10^{-8} times the number of ba-

ryons in the very early universe. How many protons are there in the entire known observable universe today ? The answer is about 10^{80}.

We can view the excess of baryons as the result of N head or tail tosses where N is a number even larger than 10^{80}. On average that will produce $N/2$ baryons and $N/2$ antibaryons. But these won't be the exact numbers. There will be an excess of one over the other due to statistical variation. Probability theory tells us that this excess has an order of magnitude of square root of N [12].

For example, if you flip a coin 1000 times, you will get on average 500 heads and 500 tails, but not exactly. The standard deviation of the difference will be $\sqrt{1000}$, that is about 32. And the difference of heads and tails will almost surely be less than three times the square root of 1000, that is 100.

So the difference between baryons and antibaryons in absolute value, divided by N, if it was due to a statistical effect, should be less than three times the square root of 10^{80} divided by 10^{80}, that is three times 10^{-40} at the most.

We conclude that the excess of 10^{-8} that we measure is *way too big* to be the result of a statistical variation around zero.

The other observation that would be extremely hard to jus-

12. Consider a random variable X that can take the value $+1$ with probability p, and the value -1 with probability $(1-p)$. Then the expectation of X is $2p-1$. In the case $p = 1/2$, this is of course zero. The variance of X is $4p(1-p)$. If Y is the sum of N independent random variables like X, then the variance of Y is $4Np(1-p)$. And its standard deviation is $2\sqrt{Np(1-p)}$. If $p = 1/2$, this gives a standard deviation of $\sqrt{N}$.

tify, if it was only a statistical variation, is why it is the same everywhere we look in the universe? You would expect a patch over here with protons, a patch over there with antiprotons, etc.

What is the experimental evidence that no neighboring or even distant galaxies are antigalaxies? Astronomers point out that there is good evidence. If the population of galaxies were sort of symmetric between galaxies and antigalaxies, then we would expect cosmic rays, and especially very high-energy cosmic rays, which are thought to be cosmic in origin, to have as many nuclei as antinuclei. However, we do see for example helium nuclei in cosmic rays, but nobody has ever seen an antinucleus in cosmic rays.

We do see antiprotons, but that is fairly easy to explain. Even if there weren't any antiprotons in cosmic rays, when a high-energy particle hits the atmosphere it can make antiprotons. But it is extremely difficult to make an antinucleus. That would take an incredible piece of luck, a very high energy collision with the atmosphere creating an antinucleus. So all this is consistent with the hypothesis that there are no antigalaxies out there, and that the antiparticles we do observe are created in collisions with the atmosphere.

The complete absence of antinuclei strongly corroborates the idea that the universe is not equally populated with galaxies and antigalaxies.

So there is something to explain. The excess of baryons over antibaryons is not statistical, and we need to explain it.

Let's finally mention another thing we don't see. If our galaxy for example was made of particles and the Andromeda was made of antiparticles, then in the region in between we would expect to find both particles and antiparticles. For example there are plenty of electrons circulating in the region between galaxies. We would expect there would also be plenty of positrons. And then we would observe lots of positron-electron annihilations.

Positron-electron annihilations are very easy to detect. They produce pairs of photons of very definite energy. These pairs of photons could be observed. Now, it is not that we don't see any electron-positron annihilation, but we don't see nearly enough of it to account for the possibility that some neighboring galaxies would be antigalaxies.

So it is almost certainly ruled out that there are antigalaxies out there. And it is certainly ruled out that there is an equal population of them.

That closes our series of arguments for saying that the baryon excess that we obseve is not a statistical effect.

CP and Sakharov second condition

Particle-antiparticle symmetry tends to suggest rather strongly that the universe was created symmetrically, although it doesn't prove it.

So the next element of the argument, in order to account for the fact that the excess of protons over antiprotons that we

observe cannot simply be a statistical effect, is that charge conjugation symmetry – i.e. the idea that the particles and antiparticles are symmetric in the laws of physics – must fail.

If we have charge conjugation symmetry, C, or charge conjugation times parity, CP, or charge conjugation together with anything else, any symmetry that involves interchanging particles and antiparticles, then we have a very hard time explaining the imbalance that we observe.

Another way to say it is that if we only allow baryon violation – i.e. Sakharov first condition – that is not enough to explain the magnitude of the imbalance. That will just give us the statistical effect [13].

We need something in the laws of physics to bias it toward particles rather than antiparticles, in other words something, in the baryon violation, which makes one of the decays, in figures 1 and 3, more probable than the other.

The implications of all that are that you need violations of particle-antiparticle symmetry, in particular the so-called CP symmetry, but basically just particle-antiparticle symmetry.

Again, is there particle-antiparticle asymmetry in the world ? Yes, there is. We know with absolute certainty experimentally that particles and antiparticles don't behave the same

13. Violation of baryon number conservation doesn't imply violation of symmetry. We investigated the consequences of Sakharov first condition while keeping symmetry. And we showed that it lead to a wide discrepancy between the proton-antiproton imbalance permitted by statistical fluctuations and that observed.

way. The examples are hard to come by, but once you have one or two or three, you know that the laws of physics are not symmetric between particles and antiparticles.

The simplest example to explain is the so-called B meson. It is a particular kind of particle made of a quark and an antiquark [14]. There are actually four kinds of B mesons, made of an up quark and a bottom antiquark, denoted $u\bar{b}$, or a down quark and a bottom antiquark, $d\bar{b}$, or a strange quark and a bottom antiquark, $s\bar{b}$, or finally a charmed quark and a bottom antiquark, $c\bar{b}$.

To simplify the argument, let's focus on the B meson made of the u quark and the $\bar{b}$ antiquark, bound into $u\bar{b}$. This particle is simply denoted B^+.

B^+ has an antiparticle. You just interchange the quark and the antiquark. Instead of $u\bar{b}$, you consider the assemblage of $\bar{u}$ and b into $\bar{u}b$. This is the anti B meson denoted B^-.

Both B^+ and B^- are electrically neutral. But they are not their own antiparticles.

Each of these particles decays into two other particles.

$$\begin{aligned} B^+ &\longrightarrow \text{ particle 1 + particle 2} \\ B^- &\longrightarrow \text{ antiparticle 1 + antiparticle 2} \end{aligned}$$

To know the decay particles is not important for us. The only important thing to remember is that it is a process in

14. There are six types of quarks respectively named up, down, charmed, strange, top and bottom, and denoted by the letters u, d, c, s, t and b. And there are the six corresponding antiquarks, which are denoted with the same letters with a bar on top.

which a particle B^+ decays into two particles, versus the process in which the antiparticle B^- decays into the corresponding two antiparticles.

Now the rates for these decays to happen are measurable. They are measured. And they are different ! One of the decays is about two-thirds more important than the other. And so in this case it is a fairly gross violation of symmetry in the particle-antiparticle interchange.

It is definitely a real effect. There are lots of indirect measurements of it. It has been known for a long time that particle-antiparticle symmetry in reality is *not a symmetry.*

Once it is established that there exist fundamental processes in nature, buried deep inside Feynman diagrams somewhere, that are imbalanced between particles and antiparticles, it is also no longer the case that the decay of the proton and the decay of the antiproton, into respectively a positron and a photon, and an electron and a photon, as shown in figure 4 below have to have equal probability.

There is a rule however that the total half-life of the proton and of the antiproton have to be exactly the same. It is a theorem in relativistic field theory.

But the way that the proton and the antiproton decays can be different from each other, while their total half-lives are the same, is that there is more than one possible way for the proton to decay.

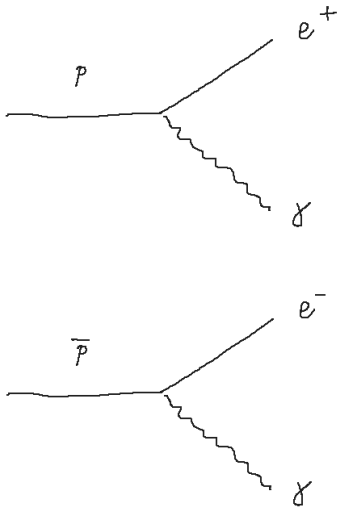

Figure 4 : Possible decays of a proton and an antiproton.

A proton can also decay for example into an antimuon μ^+ and a photon. And an antiproton can decay into a muon [15] μ^- and a photon.

What the theorem says is that if you consider all the possible ways that the proton can decay, and you calculate its total rate of decay, it must be exactly the same as the total rate of decay of the antiproton. But it doesn't say that decay rates of the processes shown in figure 4 must be the same.

15. The muon is an elementary particle similar to the electron, with the same negative electric charge, but 200 times more massive. The electron mass is 0.5 MeV, while that of the muon is approximately 100 MeV. Remember that the mass of the proton is ≈ 1 GeV.

The theorem doesn't say that any particular decay has to have a symmetry. So in particular, if at some fundamental level the charge conjugation symmetry is violated by something in the theory, then it will allow the two decays in figure 4 to be asymmetric, and the decay rates to be different.

It not only allows it, it basically insists that the two decays in figure 4 be not the same.

Once that is true, it says that there is a bias somewhere in the laws of physics – a bias toward either protons or antiprotons, or matter vs antimatter. And that is something which is absolutely necessary to add to baryon number violation to give it a directionality, that is to give it a push in one direction rather than the other.

So Sakharov second condition is CP *asymmetry*[16].

As said, Sakharov's paper was published only three years after the discovery of CP violation. That is remarkably short.

Let's list the two conditions we already have

1) *Baryon number violation.*
2) CP *asymmetry.*

Any theory that has been developped for particle physics always contains the first condition, including the standard model. It insists, as a theoretical statement, that baryon number violation must happen.

16. In this lesson, we don't distinguish parity separately from particle-antiparticle inversion, so for us C and CP are the same.

Experimentally, we know that CP is violated. In other words, particle-antiparticle interchange times parity has been discovered not to be a symmetry.

Now there is one remaining condition. But before going into it, let's do a questions / answers session.

Questions / answers session (2)

Q. : If the total decay rates of protons and antiprotons are the same, then why don't we have equal numbers of protons and antiprotons ?

A. : It comes from the fact that the μ particle is heavier than the electron.

To explain it, I need to explain first why the decay of the proton into a positron and a photon can happen rapidly, even though I said that the proton is very stable. It has to do with the environment.

How do we overcome the incredible stability of the proton ?

Remember that its mean lifetime is something like about 10^{33} or 10^{34} years or longer. The universe has only been around for about 10^{10} years. So on that scale the proton is very stable. So why don't we get to ignore the asymmetry between proton and antiproton decay ?

The reason has to do with the fact that the universe in its very early stages was very hot.

Because it was very hot, the protons or the quarks in particular – we can substitute quarks for protons in figure 2, it doesn't matter – were constantly engaged in very high energy collisions.

The high energy collisions meant that the protons moving in the plasma, that is the very hot gas or hot primordial soup, had lots of energy. How much energy ? It depends on the temperature.

But at some high enough temperature, where the average extra excitation energy E of the proton is high enough, the transient heavy particles shown in figure 2 no longer suppress the proton decay efficiently. We saw in equation (8) that the coefficient, which reduces the probability of decay, itself decreases from M^2 to $(M - E)^2$, thereby increasing the probability of the decay, or of the baryon violation.

In other words, to put it short, a proton in an environment where it isn't constantly being knocked around and having an enormously large excess energy E of some kind of another is very stable.

But when it's heated up to a high enough temperature, the transient heavy mass M shown in figure 2 is no longer an important factor, and the proton decay will happen quickly.

So if we go back to the very early universe when the temperature was very hot, then the proton and antiproton decays can happen.

Now the statement that the total decay rate is the same for

protons and antiprotons is a statement about zero temperature. It is a statement about a proton at rest in an environment where it has no excess energy. When it is being kicked around and it has extra energy then that is not necessarily true.

In fact it is generally not the case. In an environment which has some energy around, which is kicking the protons around, the decay rates don't have to be the same. So that is not a problem.

The problem has to do with the CPT symmetry.

Time-reversal and Sakharov third condition

We saw that there is no particle inversion symmetry. But there is a symmetry – at least in all known quantum field theories, string theories, any theory we know how to write down, which have relativity and quantum mechanics built into them : it is *charge conjugation* times *reflection* times *time reversal*, that is CPT.

Again what that means is

1. you exchange every particle for its antiparticle,
2. you reflect in a mirror,
3. and you run the film backward.

That is a symmetry of all theories.

What does that mean ? That means, among other things, that in thermal equilibrium, where the universe is just a pot static and hot, forward times is the same as backward time. In a universe in thermal equilibrium backward and forward in times are the same thing ; there is no asymmetry of time-reversal ; there is T symmetry.

But if the universe has T symmetry, and it also has CPT symmetry, then it must have CP symmetry. CP just means particle change to antiparticle for our purposes.

In other words, if particle goes to antiparticle combined with time goes to minus time is a good symmetry – and that, we know, is always true because it is a mathematical theorem of quantum field theory –, and the world has no bias toward one direction of time vs the other, then we are stuck again. We are back to having a particle-antiparticle symmetry. And we cannot have a proton antiproton excess in thermal equilibrium.

It is actually a theorem, which has been known for a long time, that says that in thermal equilibrium we cannot have an excess of protons over antiprotons. The thermal equilibrium will necessarily come to a configuration with equal number of protons and antiprotons.

But at early times the universe was not in thermal equilibrium. It was expanding fast – fast enough that it could not be considered in thermal equilibrium. If it is not in equilibrium that means that forward time and backward time are different. That is the reason : if the universe is expanding fast, forward time and backward time are not the same.

So in a rapidly expanding phase of the universe, we don't have to worry about time symmetry. It is definitely not symmetric. If time symmetry is broken, just by the rapid expansion of the universe, then we are in business. We have enough asymmetry of all possible kinds to explain the matter antimatter imbalance.

The baryon violation allows a change in the baryon number. The CP violation allows a directionality for it. And now the third condition is that the universe be *out of equilibrium.*

This third condition simply means the universe is expanding fast enough that running the thing backwards does not look like the original thing. It is not enough for the universe to be slowly expanding. It has to be expanding sufficiently rapidly so that all the microscopic processes don't have the time to adjust themselves to the equilibrium configuration. But whatever they are, it means that backward and forward times are different.

Movies of the universe must allow us to tell which way is forward in time and which way is backward in time [17], by the rapid expansion and cooling. The fact that at early times the universe has cooled rather suddenly is enough to completely ruin the time reversal symmetry. And if the time reversal symmetry is ruined, the CP symmetry is also necessarily ruined.

17. Notice that if we make a movie of the movements of particles in a hot static chamber, or even a chamber whose volume expands slowly while a piston moves, and we show it to spectators, they will not be able to tell whether we are showing them the movie with the time running forward or running backward.

We can finally list the three conditions of Sakharov :

1) *Baryon number violation.*

2) *CP asymmetry.*

3) *Out of equilibrium*

All three are believed to be really satisfied in the real world.

They are also believed to be sufficient. If of all three of these are true, it would be a complete accident if there was not some excess created.

The problem is that nobody knows enough about the physics of the early universe and the physics of very high energy collisions, the physics of very hot temperature, the nature of the particles that are in here, the details of what drives the CP violation and so forth, to be able to calculate the imbalance of protons to antiprotons.

So we know the three ingredients necessary and probably sufficient to explain an imbalance. It is Sakharov three conditions. But the ingredients needed to make a computation to show that

$$(N_q - N_{\bar{q}}) = 10^{-8}\ (N_q + N_{\bar{q}}) \tag{9}$$

that is out of reach. We don't know how to do that.

That is the status of this particular problem. It is the problem of baryogenesis. And it will await a much more detailed theory of both early cosmology and particle physics at very high energy.

We are now finished with baryogenesis.

Questions / answers session (3)

Q. : Doesn't the fact that we have a measurement of 10^{-8} put some constraints on what these theories can look like ?

A. : It does. But they are hard to use.

Nobody, to my knowledge, has ever used that number in a really effective way to constrain things. There are too many variables, hundreds of parameters. We just don't know enough.

Q. : What is the connexion between the positrons and electrons numbers that you mentioned previously, and the asymmetry in protons and antiprotons, or quarks and antiquarks ?

A. : Every time a proton disappears a positron appears [18].

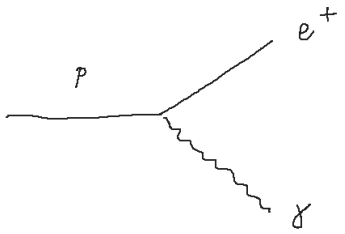

Figure 5 : Decay of a proton into a positron and an electron.

18. There can be other decays, for example into an antimuon μ^+ and a photon, but the reasoning is the same.

So the loss in baryon number is made up for by an increase in something with a positive electric charge. Conversely anything that creates a proton must either cancel a positive charge or create another negative charge too.

As long as all the processes we are looking at conserve electric charge, there is no alternative to the statement that if the baryon number shifts one way, the electron excess must shift the other way.

Q. : We hear claims that, in the standard model, the CP violation is not big enough to explain the baryonic excess. What to make of them ?

A. : I'm aware of those. And I have no reason not to believe them. I just am a little bit skeptical that anybody really knows how to do the calculation.

My involvement in baryogenesis ended with stating the three conditions, which I later discovered to have been stated thirteen years earlier by Sakharov.

Since then, dozens of people started to try to make computations based on standard models. Stephen Wolfram was one of them. He invented Mathematica, I think, to do the calculations.

In my opinion there is no way that they are going to be able to do this without, as I said, a much more detailed theory of both early cosmology and particle physics at very high energy, because there are just too many unknowns about

the early universe and so forth.

But it may be correct. And I don't know. I haven't followed this story for a long time. It may be that we know that CP violation in the standard model is too weak to drive the excess.

Once you admit CP violation into physics, though, there is no reason not to expect that at high energies the ratio could be tens or hundreds of times larger than 10^{-8}. Once you have opened the door to the imbalance, you don't have really much control over it.

Q. : Has the 10^{-8} number always been the same number ?

A. : The 10^{-8} is today's number for the ratio of baryon number or protons to photons.

In the early universe, there was a lot of annihilation between protons and antiprotons. But before the temperature fell low enough, it was compensated by the creation of new protons and antiprotons. There was an equilibrium.

As I said, the five numbers

$$N_\gamma \quad N_p \quad N_{\bar{p}} \quad N_e \quad N_{e^+}$$

were all about equal. But there appeared a slight excess of protons over antiprotons, and an exactly opposite excess of electrons over positrons. The excess is slight, but big enough not to be explainable by statistical fluctuations.

At very high temperature, there are very high-energy photons. Two photons come together and they create one way or another a proton and an antiproton. These photons must together have at least 2 GeV of energy.

As long as the temperature is high enough, these processes are going on backward and forward in equilibrium.

But then, when the temperature falls below a certain threshold, there simply aren't enough high energy photons around to create protons and antiprotons. Protons and antiprotons, however, continue to annihilate each other creating photons.

Those photons still have a lot of energy. But they go out into the soup and their energy gets lowered by coming to equilibrium with the lower temperature background stuff.

So as the temperature goes down the number of available high-energy photons decreases and you can't make the proton-antiproton pairs anymore. But the other way around still goes on.

Thus there is a point at which the annihilation is no longer compensated and eventually all the protons and antiprotons get eaten up, except for the excess of one kind over the other.

Once the excess reached a certain value in the early universe, it then never changed. $N_p - N_{\bar{p}}$ stayed the same. This difference is today's number of protons in the universe. So the ratio of 10^{-8} could be measured in the past, but as an

excess of protons over the total number of protons and antiprotons, or equivalently over the total number of photons, because *in orders of magnitude* the two denominators are the same.

Q. : When did the process of creation of protons and antiprotons start ?

A. : That was quite early, when the temperature was in the gigaelectron-volts for the energy of the photons.

It began within a second or so after the initial inflation which created space, and which is our next subject in this lesson and the following one. So it is at the left extremity of figure 6, right after inflation.

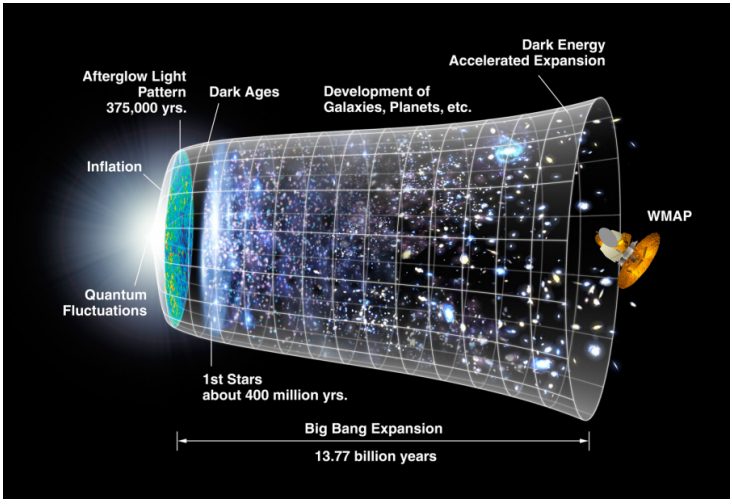

Figure 6 : Chronology of the universe.
Source : NASA/WMAP Team.

Q. : Why are the photons today not as energetic as when they were creating pairs of protons and antiprotons ?

A. : Today's photons – those of the CMB, which are the overwhelming majority of photons in the universe – are those that participated in the creation of proton-antiproton pairs and also electron-positron pairs. They remained as numerous in order of magnitude. But they are no longer as energetic.

Remember that radiation energy density decreases like $1/a^4$, where a is the scale factor, while the volume of the universe increases like a^3.

Q. : If we need these transient very massive particles in the decay of the proton, figure 2, does that mean the standard model is still incomplete ?

A. : Yes, it does.

That is what is known about the matter-antimatter imbalance in the universe. And, as I said, I don't think it is going to be too soon before we know a lot more and in particular are able to calculate the 10^{-8} ratio from theory.

The next topic we want to move onto is inflation, or inflationary universe. We shall do some groundworks in the remainder of this lesson. And next lesson will be entirely devoted to it.

Inflationary universe

The inflationary universe idea was put forward to try to explain another of those observations which were puzzling. The question was why is the universe so terribly homogeneous and isotropic. It became a rather critical issue when the cosmic microwave background was discovered.

The CMB quickly became a rather high precision affair. In short order the CMB wavelength distribution had been observed to correspond to the blackbody curve ; its temperature had been measured carefully ; it was known to be about 2.7 degrees Kelvin. Over the years, the fit between the CMB and the blackbody curve became incredibly precise. Today the error bars are microscopic.

So the temperature was very well defined. The CMB looked like a Planck distribution, that is like a blackbody distribution almost exactly – exactly as far as we could tell.

But, moreover, it was the same in every direction. Thus suddenly this rough idea that the universe is isotropic became a high precision idea.

In truth, it took years to establish this scenario. I'm condensing history. But, by now in any case, the idea that the universe is isotropic is a very high precision fact, and can be the starting point for further theoretical developments. So we can ask why is the universe so isotropic ?

If today it is isotropic, it must have been isotropic very early. There is no particular reason why an anisotropy, a lumpiness in the distribution, would decrease with time.

In fact, it is quite the opposite : lumpiness tends to increase with time, because of gravity. Gravity tends to take lumps and increase their magnitude.

Consequently, the universe must have started very early being extremely homogeneous and isotropic [19]. When I say very early, I mean at the time before there were galaxies, before there were stars and planets, at the time when the blackbody photons originated. In other words, at the decoupling time the universe was extremely isotropic.

It was of course known that it couldn't be completely isotropic and homogeneous. If it was exactly homogeneous, it would stay homogeneous. Anything which is exactly uniform and allowed to evolve will stay uniform.

But the universe is not uniform. It is full of galaxies and full of clusters of galaxies. It has a lumpiness.

That lumpiness, that we see now, clearly was much smaller in magnitude to begin with, for the simple reason that lumpiness tends to increase with time. As said, if you start with a world which is completely uniform of course it stays that way. But let's suppose there was a little bit of overdensity in some fairly big region. The density was just a tiny bit bigger than in the neighboring regions. What happens ?

In a gravitational theory what happens is the opposite of what happens in other kinds of theories.

19. Homogeneous and isotropic is a conventional wording. Logically isotropy does not imply homogeneity – humankind could happen to be at the center of the universe –, whereas homogeneity implies isotropy. So to say the universe is homogeneous would be sufficient. We even often use isotropic to mean homogeneous.

In other kinds of theories what tends to happen is that the overdensity will diffuse out and be eliminated. For example we have an overdensity of ink dropped into water. Over time the drop diffuses out and then disappears, the density of ink becoming homogeneous.

In gravity the opposite happens. And the argument is very simple. If we have an overdensity of matter in some region, because gravity is universally attractive it will tend to attract the stuff around it. It will pull into the region the stuff outside, decreasing the density outside and increasing it inside, figure 7.

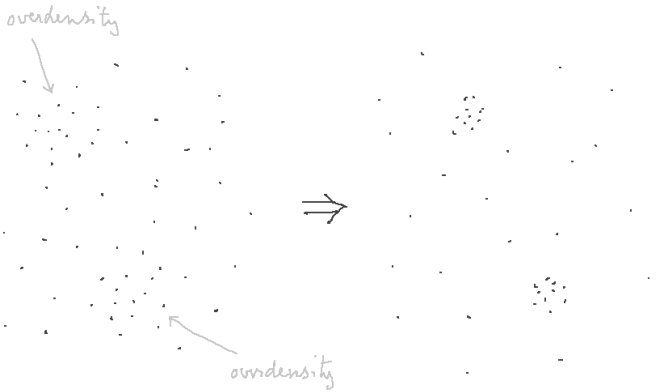

Figure 7 : Evolution under gravity of slightly overdense regions into aggregates of matter.

In other words, it is kind of a runaway situation where a little bit of inhomogeneity will tend to reinforce itself. So if we have, distributed in space, regions with overdensity, underdensity, overdensity, underdensity and so forth, in some pattern, the tendency of gravity will be to suck stuff out

of the underdense regions and put it in into the overdense regions, thereby magnifying the degree of inhomogeneity.

So the fact that we see inhomogeneity today does not mean that the universe very early was as inhomogeneous as it is today. It must have been much less inhomogeneous, just by running the argument backward.

In fact, rather early, cosmologists were able to estimate, by running the theory backward, just how much inhomogeneity there was at the time of decoupling in order that the galaxies could nucleate out of that inhomogeneity. In other words the picture is that the universe was very homogeneous, with just a little bit of inhomogeneity. There were little ripples, little bits of excess, depletion, excess, depletion of some kind.

The overdense regions eventually collapsed by the mechanism shown in figure 7. They formed galaxies and clusters of galaxies. And the underdense regions formed voids. And that is what we see today.

Jim Peebles [20] and other early cosmologists in the sixties estimated how much inhomogeneity was necessary at the time of decoupling.

Here is the way we quantify it. We characterize the lumpiness by the ratio

$$\frac{\delta\rho}{\rho} \tag{10}$$

20. Phillip James Edwin Peebles (born in 1935), Canadian-American physicist and theoretical cosmologist.

where ρ is the density, and $\delta\rho$ is roughly the mean excess density in a lump relative to the background. So the ratio is the fractional overdensity in a typical overdense region relative to the density itself.

That is the dimensionless measure of how much inhomogeneity there was. And it was pretty early recognized that this number had to be

$$\frac{\delta\rho}{\rho} \approx 10^{-5} \tag{11}$$

More precisely, so to speak, it had to be in a range between 10^{-4} and 10^{-5}.

Here we were again in the same kind of situation as with the excess of matter over antimatter. Did the question "Why was the universe almost perfectly homogeneous ?" really require an answer ? Perhaps it was enough to say : "Well, it just started that way." But now there was a number in addition to just saying the universe was almost homogeneous.

The universe was almost perfectly homogeneous, with a specific numerical magnitude to its inhomogeneity. And once you have a specific numerical magnitude, you want to know why that is the magnitude.

Now, it is not that we learned why this is the magnitude – we did not. We do not really understand why 10^{-5} is the right number. But the existence of this number focused our attention on two questions. The first is : Why is the universe almost perfectly homogeneous ? Later we will worry about the second : Why is it not exactly homogeneous ?

The first-order fact was that it was very very homogeneous. The explanation of that today – which seemed rather far-fetched when it was put forward by Alan Guth [21] in 1980 – was that the universe simply expanded by many orders of magnitude. And of course when something expands it stretches.

If you have an inhomogeneous universe with lumpiness in it, and you stretch it out enough, you will make it homogeneous, at least on the scales that are relevant. So the idea was the universe inflated, which means expanded. In particular it exponentially expanded for some period of time, and stretched itself out so much that it flattened itself.

It is literally like blowing up a balloon. When it is deflated, the balloon has a crinkled shape. We blow it up and it stretches out and flattens out, figure 8.

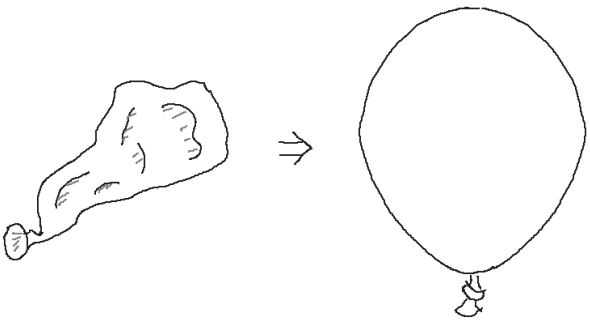

Figure 8 : Inflating a balloon flattens out its crinkles.

21. Alan Harvey Guth (born in 1947), American theoretical physicist and cosmologist.

Of course that is an analogy which we shouldn't take too far. But it may help feel how streching leads to an homogeneization of density.

Questions / answers session (4)

Q. : Can you explain how the universe can be homogeneous if at the same time it has lumpiness ?

A. : It is a matter of scale. If we look around this room, it does look homogeneous at all. But on an average over large distances, for all practical purposes it is homogeneous.

It is like the surface of the Earth. It is full of hills and mountains and valleys and so forth, and as a geologist you want to explain these geological phenomena. But on the whole the Earth is a very smooth ellipsoid.

The universe is the same. It may have lumpiness, galaxies and so forth, and yet be considered homogeneous when looked at on a sufficiently large scale.

Q. : If this inhomogeneity of the universe is a random process causing lumps to appear, wouldn't we expect much less overall homogeneity ?

A. : The initial inhomogeneity and the appearance of lumps is not a random process in the sense you think.

You are suggesting that if it were random we might expect that somewhere out there there would be large overdensities. There would be some statistical probability that we might find a large lumpiness.

But it is not random [22]. And we are going to talk about the pattern of $\delta\rho/\rho$.

Basically it is like this

Figure 9 : Inhomogeneities in one dimension.

Now squeeze this down until it is of size 10^{-5} relative to the background. On the whole it is very smooth. Nevertheless there are wiggles in it. The sizes of the wiggles are very small.

Suppose, furthermore, they are homogeneously, isotropically distributed through space. There are no tremendously large lumps out there. They are all very small in magnitude relative to the average. It is more or less like a very very smooth Earth with ripples on it.

We will see what the implications are.

22. Most paradoxes and popular riddles about randomness come from a deficient or mistaken specification of *what is the experiment to be replicated*, and what is the random variable to be measured.

Q. : What was the temperature like during inflation ? How did it decrease and what role did it play ?

A. : During the inflationary period temperature was not an important aspect of things. It was not a terribly relevant factor.

It was the exponential expansion which was the single most important thing.

We shall go through the basic equations of inflation in detail in the next lesson. To finish the present lesson, let's review some preliminaries about friction that will be useful.

Friction and viscosity

In the context which will occupy us, when we speak of *friction* we are speaking of *viscosity*. We are thinking about something like a stone falling through a viscous fluid.

What does friction or viscosity have to do with anything ? It has everything to do with everything. But to finish this lesson and prepare for the next one, we are just going to study elementary friction, the elementary equation of friction. We shall write it down. It is very simple. Why are we doing that ? Because it will come up.

Consider a stone falling through water or honey, figure 10. It is falling due to a force which is a gradient of potential energy, with a negative sign.

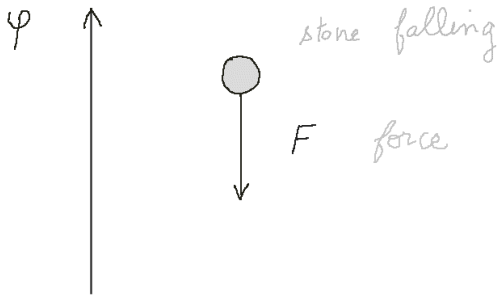

Figure 10 : Stone falling through a viscous fluid.

We are going to call its height ϕ, which is a stupid name for the height but it is not a stupid name for a field. And as you might expect what we are going to be talking about is fields.

Nevertheless, the analogy is ϕ is the height of the stone. And there is a force on the stone which is related to its potential energy. In this case it could just be the potential energy of gravitation. But let's call it $V(\phi)$ and not worry about where it comes from.

Let's assume the force is downward, which means $V(\phi)$ increases in the upward direction. The force exerted on the stone due to the potential energy is the derivative of V with respect to ϕ, with a minus sign.

$$F = -\frac{dV}{d\phi} \tag{12}$$

Now let's write Newton's equation for the stone, if there is only that force. For simplicity we take the stone to have unit

mass. The generic equation $F = ma$ in this case becomes

$$-\frac{dV}{d\phi} = \ddot{\phi} \tag{13}$$

The force on the left-hand side is minus $dV/d\phi$. And on the right-hand side, the mass is 1 and the acceleration is $\ddot{\phi}$, that is the second derivative of ϕ with respect to time.

This is just $F = ma$. There is nothing special there. In a uniform force field, which would mean the $dV/d\phi$ is a constant, the stone will fall with a uniform acceleration. So it will simply pick up speed, and continue to accelerate. Alright.

But that has ignored the viscosity of the fluid that the stone is moving through, which let's say is something like honey. When we take this into account, that means that there is another force on the left-hand side of equation (13).

The other force is zero if the object is at rest. The viscosity exerts no force on an object at rest. It is the motion through the fluid which creates the force. And so the extra force is going to depend on the velocity. The larger the velocity, the higher the force due to viscosity.

For simple fluids the force due to viscosity – which we sometimes simply call "the viscosity" – is actually proportional to the velocity itself. So, adding this force, and reversing sides in equation (13), we get

$$\ddot{\phi} = -\gamma\dot{\phi} - \frac{dV}{d\phi} \tag{14}$$

The term $\gamma\dot{\phi}$ has a minus sign in front of it because viscosity opposes motion. The coefficient γ is called the *drag*.

The potential energy increases with height, so, in equation (14), the term $-dV/d\phi$ is a force pulling downward. The stone is falling. Therefore $\dot{\phi}$ is negative, and $-\dot{\phi}$ is positive. Thus the force from viscosity is pointing up, opposing that from potential.

In the beginning, when the stone starts to fall, $\dot{\phi}$ is 0. So basically the stone starts out accelerating exactly as it would without the viscosity term.

But then $\dot{\phi}$ will increase. And, unless as we move down the force from potential gets bigger and bigger, because $\dot{\phi}$ is increasing the force from viscosity will eventually balance $dV/d\phi$. At that point forces cancel, the acceleration becomes nil, and the corresponding $\dot{\phi}$ is called the *terminal velocity*.

In particular, if the downward force is constant, for example like the force of gravity which is uniform near the surface of the Earth, then it is just F, and, if we take F to be negative, equation (14) simplifies into

$$\ddot{\phi} = -\gamma\dot{\phi} + F \tag{15}$$

Then the terminal velocity satisfies

$$0 = -\gamma\dot{\phi} + F \tag{16}$$

or equivalently

$$\dot{\phi} = \frac{F}{\gamma} \tag{17}$$

The terminal velocity has the same sign as F. Hence if F is negative, the terminal velocity is negative. And this is consistent, of course, with the fact that the stone is falling through the honey.

That is the basic theory of friction in a nutshell. And what does it do ? It slows things down.

What happens in particular if $V(\phi)$ has a shallow slope ? To view it comfortably let's plot $V(\phi)$ as in figure 11.

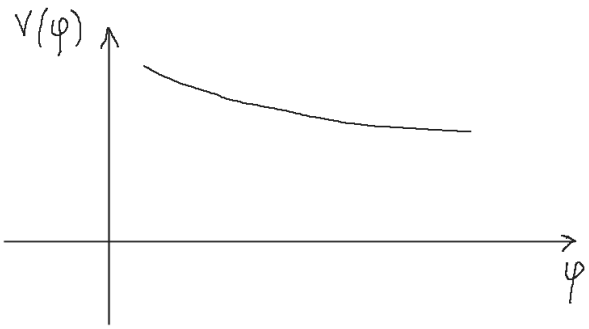

Figure 11 : Potential $V(\phi)$ as a function of ϕ.

Even though ϕ is the height, in this figure we represent it on the horizontal axis in order to view $V(\phi)$ on the vertical one. So we represented a potential which has a very small

gradient, or what we can call a shallow slope. Then the stone can be imagined, as time passes, rolling slowly along this curve [23].

If not only the potential has a shallow slope but the viscosity coefficient γ is large, then the stone falling along the potential energy curve of figure 11 will take a long long time to roll down the hill.

That is the situation we will want to be in. And ϕ will not be the position of a stone. It will be the value of a field. But we will want to be in a situation where a field evolves very slowly because of a lot of friction. And that will drive inflation.

Let's begin to use this model. And we will continue in the next lesson.

Classical field theory

What we are going to consider is classical field theory. The universe is filled with some field. The field is going to be a scalar field.

Now where does this come from ? What scalar field ? It was simply made up in order to be able to explain the isotropy and homogeneity of the universe.

23. Remember that Galileo, to study comfortably falling objects, let them rolling on wood planks with only a moderate slant. That way he created only a small horizontal gradient.

At the time that this was created, it was a long shot rather crazy idea, but it seems to be right. Here is the idea : in addition to the electromagnetic field, the gravitational field and all the usual fields, there is one more field in the universe. It is a scalar field. It is called the *inflaton*[24] and is denoted ϕ.

Why inflaton ? Because it has to do with the inflation of the universe when it began.

We are going to assume that ϕ is pretty much uniform in space. We could assume that it is not uniform in space. But, as the space expands, it will tend to stretch out the variations in ϕ. So let's just assume for simplicity that ϕ is uniform in space.

Let's think about the energy stored in a scalar field of this type. Does the reader remember what the energy density of a scalar field is ? We studied it in volume 3 of the collection *The Theoretical Minimum* on special relativity and classical field theory. It contains a kinetic term with a time derivative. There are other terms which have to do with gradients in space. So the energy density *begins with*

$$E = \frac{\dot{\phi}^2}{2} + \text{terms related to gradients in space} \qquad (18)$$

The term $\dot{\phi}^2/2$ is not the kinetic energy in the sense of movement in space but in the sense of due to time dependence.

As said, in general there are also differential terms in the energy which have to do with gradients in space. Gradients

24. Note the spelling and pronounciation : *inflaton*, not inflation.

in space also store energy.

But since we have assumed that the field is homogeneous, i.e. not varying in space, which we can justify – we will do that later –, there are no gradients in space. Therefore the only term involving derivatives is the one akin to a kinetic energy.

Then the other thing that can be on the right-hand side of equation (18) is a potential energy density $V(\phi)$. So the final form of the energy density is

$$E = \frac{\dot{\phi}^2}{2} + V(\phi) \tag{19}$$

$V(\phi)$ is just a thing that is made up : different values of the field have different energies. It doesn't have to do with derivatives, either in time or in space. Just in having a field of a given magnitude there is an energy density associated with it. It is called the field potential energy [25].

It is not by accident that we use the same notation as in the previous section. Here ϕ is a field. Before it was the coordinate of a stone.

Now let's follow the field in an expanding box. As usual we look at a box that is expanding with the universe. How big is the box at any given time ? The box has volume a^3, where a is the scale factor of expansion of the universe.

25. Pay attention to the fact that it is a *density*. Its units is energy per volume.

To get the energy of the field in the box, we use the energy density given by equation (19), and we multiply it by a^3.

$$E = a^3 \left[\frac{\dot{\phi}^2}{2} + V(\phi) \right] \tag{20a}$$

By a slight abuse of notation, in equation (19) E was an energy density, whereas now in equation (20) it is a plain energy.

But a is time dependent. So equation (20) is better expressed as

$$E = a(t)^3 \left[\frac{\dot{\phi}^2}{2} + V(\phi) \right] \tag{20b}$$

So we have a time-dependent expression for the energy in the "expanding unit box" [26].

You can think of equation (20), if you like, as formally, mathematically being the same as the as the energy of a particle : kinetic energy plus potential energy, except with a coefficient in front which depends on time – something you wouldn't ordinarily write down.

We might have a funny situation where the mass of a particle might depend on time. But ordinarily we wouldn't write down, for the energy of a particle, an equation like equation (20).

26. Remember the fictitious grid which we introduced in chapter 1. The grid-unit is by definition 1. But its real value in meters is a(t).

Nevertheless that is what we have. We have an energy, kinetic plus potential.

Questions / answers session (5)

Q. : What would happen if ϕ depended not only on time but also on space ?

A. : The general form of a scalar field is represented in figure 12.

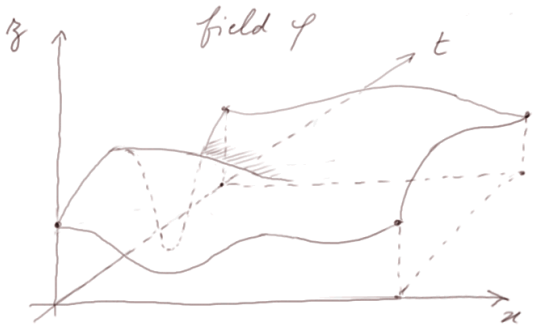

Figure 12 : General form of a scalar field, depending on time *and* space.

In general, a field depends on time *and* space. If there was a space-dependence, there would also be terms in the energy equation (20), in the brackets, which have to do with gradients in space. As I said, gradients in space also store energy.

But we assume for the moment that the field is uniform in space. If it is not, eventually it will be because the expansion of the universe will stretch it out. It will iron out, so to speak, the variations in space. Go back to figure 8, to see the comparison with an inflated balloon.

So that is reasonable to assume that, after a period of time, there are no gradients of the field in space.

Q. : So, just on a conceptual level, we are thinking of associating this field with a quantum space of states ?

A. : Yes. But here we are not doing quantum field theory, but classical field theory.

Furthermore we assume that the *inflaton field* is uniform in space. Therefore all we need to do to find the total energy is to multiply by the volume of the box, which is time dependent.

Q. : Is there anything physical we can think of which would correspond to this time-dependent and space-uniform field ϕ, and that would help us figuring it out and figuring out its evolution ?

A. : Well, I'm so used to it that I think of it as sort of obvious. But we are going to find out. We have to solve the equations of motion.

Your question is like asking before hand what will happen to a particle whose energy is

$$E = \frac{\dot{\phi}^2}{2} + V(\phi) \tag{21}$$

Is it moving up or moving down ?

That will depend on the initial conditions. It will depend on how long we wait. And it will especially depend on the sign of $dV/d\phi$, namely the sign of the force.

And whatever it will be doing, it will be doing it everywhere uniformly in space simultaneously.

Now we can back off that, and study what happens when it is not uniform in space. But this is the easy problem for us to study.

Now, how do we find the equation of motion when we know the kinetic energy and we know the potential energy ? There are various ways we could do it. But the most efficient way is through Lagrange's equations.

Let's write the lagrangian. As the reader remembers, in a simple situation like here, where the energy is made of two clearly separated terms, the lagrangian is the *difference* between the kinetic energy and the potential energy [27].

27. Of course, when talking of a field, we are talking first of all of the *density of energy* at a point in time and space. In due course, we shall integrate it to get a plain energy.

So we pick up our copy of volume 1 in the collection *The Theoretical Minimum* on classical mechanics, and we readily write

$$\mathcal{L} = a(t)^3 \left[\frac{\dot{\phi}^2}{2} - V(\phi) \right] \tag{22}$$

where ϕ is like the coordinate – what we called the *degree of freedom*. And equation (22) is the lagrangian.

The only new thing that wouldn't be there for an ordinary particle is this coefficient $a(t)$ to the cube in front.

From there, let's work out Lagrange's equations. When there is no space-dependence, Lagrange's equations for a field are simply

$$\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{\phi}} = \frac{\partial \mathcal{L}}{\partial \phi} \tag{23}$$

The calculation begins with

$$\frac{\partial \mathcal{L}}{\partial \dot{\phi}} = a(t)^3 \dot{\phi} \tag{24}$$

Then we have to take its time derivative. Equation (23) rewrites

$$\frac{d}{dt}\, a(t)^3 \dot{\phi} \;=\; -a(t)^3\, \frac{dV}{d\phi} \tag{25}$$

The derivative of V with respect to ϕ can be written as a plain derivative, as opposed to a partial one, because there

are no other independent variables driving V.

We could call the derivative $dV/d\phi$, the force, or more accurately minus the force. But notice that there is the extra explicit time-dependent coefficient $a(t)^3$ appearing on both sides of equation (25).

Therefore it doesn't quite look like Newton's equations. It would be Newton's equations if a was constant. But a is not constant. So let's work it out and see what it is.

For the left-hand side, differentiating with respect to time the product, we get

$$a^3\ \ddot{\phi}\ +\ \dot{\phi}\ 3a^2\ \dot{a} \tag{26}$$

If, on the right-hand side, we write $-dV/d\phi = F$, we get for equation (25)

$$a^3\ \ddot{\phi}\ +\ 3a^2\ \dot{a}\ \dot{\phi}\ =\ a^3\ F \tag{27}$$

It is very tempting, since it appears on both sides, to divide by a^3. Let's do it

$$\ddot{\phi}\ +\ 3\ \frac{\dot{a}}{a}\ \dot{\phi}\ =\ F \tag{28}$$

What is $\dot{a}$ over a called in cosmology ? The *Hubble constant.* Remember that it is a constant in space, but not necessarily in time. It can be and usually is time-dependent.

So equation (28) now takes the form

$$\ddot{\phi}\ +\ 3\ H\ \dot{\phi}\ =\ F \tag{29}$$

We recognize the same form as equation (14). Hence, we made the junction with our preliminary work on friction and viscosity :

Equation (29) is exactly the same equation as that of a stone falling through a fluid with a viscosity coefficient, i.e. a drag, $\gamma = 3H$, and there is a force which is minus the gradient of V.

Remember that the potential energy – or energy density to be precise – $V(\phi)$ is a made up thing which we added to the energy of the *inflaton field* ϕ, equation (19).

So our model for the way this field evolves can be envisioned by just supposing that ϕ was the position of a particle on a hill, where the height of the hill was V, and where the gradient of the altitude of the hill was the force $-dV/d\phi$, figure 13.

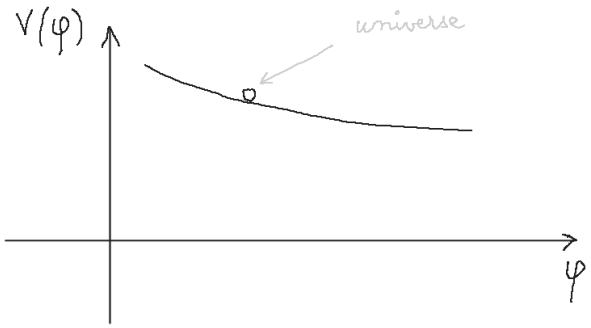

Figure 13 : Inflaton field ϕ and evolution of the universe.

The evolution of the universe can be compared to that of

a ball rolling down the hill, except that there is a viscous drag force proportional to the Hubble expansion rate. That is why we went through the exercise of studying elementary classical friction.

If the Hubble friction is strong, that is, if H is large, and if the hill in figure 13 is reasonably flat, not terribly flat, just moderately slanted, then our model is like a ball moving through motor oil on a cold day in Wisconsin.

If just left to its own devices without the friction term, the ball might roll down the hill in a few seconds. But with a large friction it could take years to roll down the hill, depending on the magnitude of the friction.

So the term $3H\dot{\phi}$ is called the *cosmic friction term* in the equation of motion of the scalar field ϕ, equation (29). It has the simple effect of slowing down the evolution of the system and keeping the ball from rolling down the hill.

In the next lesson, we are going to use this to study the cosmology of a universe which contains a field like ϕ. In other words, we are going to look at the Friedmann's equation with an energy density which is given by

$$E = a(t)^3 \left[\frac{\dot{\phi}^2}{2} + V(\phi) \right] \tag{30}$$

We shall study how the universe expands and evolves under the influence of an energy density resulting from a field which is slowly slowly slowly rolling down the hill.

That is the phenomenon of *inflation*, i.e. the way the universe responds to this very small slowly moving inflaton field ϕ.