LANC-TH 9319
astro-ph 9312022
November 1993

INTRODUCTION TO COSMOLOGY

Lectures given at the
Summer School in High Energy Physics and Cosmology
ICTP (Trieste) 1993

David H. Lyth

School of Physics and Materials,
Lancaster University,
Lancaster LA1 4YB. U. K.

Abstract

These notes form an introduction to cosmology with special emphasis on large scale structure, the cmb anisotropy and inflation. In some places a basic familiarity with particle physics is assumed, but otherwise no special knowledge is needed. Most of the material in the first two sections can be found in several texts, except that the discussion of dark matter and the cosmological constant is more up to date. Most of that in the remaining sections can be found in a review of structure formation and inflation done with Andrew Liddle, which describes original work by various authors including ourselves and Ewan Stewart. The reader is referred to these works for more detail, and a very complete list of references.

Contents

1 The Recent Universe

These notes form an introduction to cosmology with special emphasis on large scale structure, the cmb anisotropy and inflation. In some places a basic familiarity with particle physics is assumed, but otherwise no special knowledge is needed. Most of the material in the first two sections can be found in several texts [1], except that the discussion of dark matter and the cosmological constant is more up to date. Most of that in the remaining sections can be found in a review of structure formation and inflation done with Andrew Liddle [2], which describes original work by various authors including ourselves and Ewan Stewart. The reader is referred to these works for more detail, and a very complete list of references.

According to current thinking, the history of the observable universe broadly divides into three stages. First there is an *inflationary era*, when the energy density is dominated by the potential of a scalar field. Then there is a *radiation dominated era* when the energy density is dominated by relativistic particles, which are called 'radiation' by cosmologists. Finally, lasting till the present epoch, there is a *matter dominated era* when the energy density is dominated by the mass of non-relativistic particles, which are called 'matter'. This first lecture concerns the latter part of the matter dominated era, and in particular the present epoch.

Unless the contrary is implied by the specified units, I set $\hbar = c = k_B = 1$.

1.1 How one decides what is true

Before beginning let us address an important issue. More than most scientists, cosmologists tend to be asked 'how do you know?' The answer is that one proceeds in the same as in other areas of physics and related sciences. One starts with a *model*, which is a statement about the nature of physical reality, plus a theory formulated in terms of equations. Then one calculates ('predicts') the outcome of observations and experiments that have been performed, or better of ones that are about to be performed. If the agreement is impressive enough everyone agrees that the model is correct.

This sequence *cannot* be reversed, that is one cannot rigorously work back to deduce the model from observation. To put it the other way round, a model can strictly speaking only be falsified, not verified. Logically, two completely different models could both agree with observation but that doesn't seem to happen, though of course an old model can be seen to be a limiting cases of a better one.

So the answer to the question 'how do you know' begins with the admission that a guess has been made. At any one time, though, parts of the picture are universally accepted as correct, because they agree with such a huge body of observation. Other parts of the picture are just coming into focus, and agreement about the truth lies in the future. The scope of the agreed part is dictated not by logic, but by consensus.

Except for the mathematical theory, all of this is not really special to science, but is rather just the way that human beings go about things. For instance, at any given epoch during the centuries that the geography of the earth's surface was being established in the West, parts of it were universally accepted because the accounts of many people agreed (or sometimes because just one of two surveyors had done what looked like a good job). At the same time, parts of it were still in dispute. The portion of the earth's surface about which there was agreement was dictated not by logic, but by consensus.

Because our knowledge has been arrived at by comparing a large body of ideas with a large body of observation, it is not usually possible to point to a particular observation as the reason for believing a particular fact. For this reason, as well as because of the introductory nature of the lectures, I will often state a fact without giving specific justification. On the other hand, by the time we have finished it will hopefully be clear that the picture I put forward leads to agreement with an impressive body of observation, and that this picture could hardly be subjected to drastic alteration without spoiling this agreement.

1.2 The geography of the observable universe

Astronomers use the unit 1 pc= 3.26 light years= 3.09×10^{16} metres. For cosmology the megaparsec, $1\,\mathrm{Mpc} = 10^6\,\mathrm{pc}$ is appropriate. The unit of mass is the solar mass $M_\odot = 1.99 \times 10^{33}$ g.

Stars have mass in the range roughly 1 to $10 M_\odot$. They are found only in *galaxies* may be regarded as the basic building blocks of the universe, with masses ranging from maybe $10^6 M_\odot$ (dwarf galaxies) to $10^{12} M_\odot$ (large galaxies like our own). A galaxy typically has a luminous centre containing nearly

all of the stars, and a dark halo of unknown composition which extends of order 10 times as far and contains of order 10 times as much mass. In many (all?) cases there is a nucleus consisting of a black hole which is gobbling up stars and gas and emitting radiation. If the nucleus is the dominant feature it is called an AGN (active galactic nucleus), examples being Seyfert galaxies and quasars. (Astronomers tend to using different names for different examples of the same type object. The reason, of course, is that they don't know at first that they *are* examples of the same type of object.)

In round figures, large galaxies like our own have a size of .1 Mpc (including the dark halo) and are of order 1 Mpc apart. Many galaxies belong to gravitationally bound clusters containing from two to ~ 1000 galaxies. Small clusters are usually called groups. Big clusters, of order 10 Mpc in size, are the biggest gravitationally bound objects in the universe. There do exists, though, 'superclusters' with size of order 100 Mpc, which are conglomerations of galaxies and clusters representing regions of space which have higher than average density. Presumably they will become gravitationally bound at some time in the future. On the scale of 100 Mpc there also seem to be sheetlike and filamentary structures, as well as voids containing hardly any galaxies.

On scales bigger than 100 Mpc the distribution of matter in the universe is known to be very homogeneous, both from direct observation of the galaxies and from the isotropy of the microwave background. To be precise, if one throws down at random a sphere with radius R and measures its mass M, then the *rms* variation $\Delta M/M$ is a decreasing function of R, which is of order 1 at $R = 10$ Mpc and of order .1 at $R = 100$ Mpc

The biggest distance we can observe is of order 10^4 Mpc, the distance that light has travelled since the Big Bang. The sphere around us with this radius is thus the *observable universe*. As far as we can tell, distant parts of the observable universe are much like our own part, with the same kind of galaxies, clusters and so on. From the fact that the microwave background anisotropy $\Delta T/T$ is of order 10^{-5}, one can deduce that $\Delta M/M \lesssim 10^{-5}$ on scales R comparable with the size of the observable universe.

What's beyond the observable universe?

What is the universe like outside the patch we observe? Since the universe is very homogeneous and isotropic in the observed patch, it is reasonable to suppose that it remains so for at least a few orders of magnitude further. But what about happens after that?

Until the last decade or so, the prevailing assumption seems to have been that the entire universe is homogeneous and isotropic, like the bit that we observe. To avoid the embarrassing question of what would then be beyond the edge of the 'entire' universe, this view requires that one believes in a 'closed' universe, which is allowed in the context of general relativity. Such a universe is the three dimensional analogue of a sphere, and as I discuss later on this lecture its spatial 'curvature' could be detectable observationally.

More recently, the favoured view has been that as one examines larger and larger scales, the universe becomes steadily more inhomogeneous and anisotropic [3]. Going the other way, if one takes an overall view of a patch of the universe many orders of magnitude bigger than our own, it has a non-fractal nature; as one goes down to smaller and smaller scales, it looks more and more homogeneities, as opposed to the 'fractal' situation where new inhomogeneities would reveal themselves at every step. On this view, the observable universe is *extremely* homogeneous because it is a small part of a much large patch which is *roughly* homogeneous and isotropic. As I shall discuss in Lecture 6, this viewpoint is normally discussed within the context of inflation which indeed makes it useful and rather compelling, but one should understand that it is a separate hypothesis, not related *per se* to the inflationary hypothesis.

On this second view, there can be yet another hierarchy as one moves beyond the roughly homogeneous and isotropic patch. One could encounter regions where the universe is radically different, for instance contracting instead of expanding, or with different laws of physics, corresponding to different solutions of some super-theory. One would presumably still like the 'entire' universe to be closed, but on this second view it would be analogous to the surface of an extremely deformed 'sphere' so that no hint of the global behaviour would be evident from observations of our own region.

1.3 The isotropically expanding universe

On the large scales where it is homogeneous, the universe is also expanding isotropically. That is, the distance between any pair of galaxies separated by more than 100 Mpc is proportional to a universal

scale factor $a(t)$, the same for every pair. The *Hubble parameter* is defined by $H = \dot{a}/a$, where the dot is the time derivative.

A subscript 0 is generally used to denote quantities evaluated at the present epoch. It is convenient to set $a_0 = 1$, so that $a(t)$ is simply the size of any comoving region (one moving with the galaxies) relative to its present size. The present value of H, denoted by H_0 is called the Hubble constant. It is traditionally measured by observing the redshift $z \equiv \Delta\lambda/\lambda$ of galaxies receding from us with velocity $v \ll 1$. The velocity of such a galaxy is given by $v = Hr$, and its redshift is just the non-relativistic Doppler shift $z = v$, leading to Hubble's law

$$z(= v) = H_0 r_0 \tag{1}$$

Hubble's law is well established because *relative* distances are easy to establish. All one has to do is find a 'standard candle', that is a type of object (say as star of a given type) of which all examples have practically the same luminosity. Then its apparent luminosity will vary with (distance)$^{-3}$, and so measure relative distances.[1] On the other hand to fix H_0 which is the constant of proportionality one has to know the luminosity of some object, which is much harder to do. Different estimates give H_0 in the range 40 to 100 km sec^{-1} Mpc^{-1}, and it is usual to define a quantity h by

$$H_0 = 100h \,\mathrm{km\,sec^{-1}\,Mpc^{-1}} \tag{2}$$

Thus, the redshift determination gives $.4 < h < 1$. As one might expect, H_0 enters into many other equations besides Hubble's law, so the redshift determination is *not* the only method of determining H_0. I will suggest a 'best fit' value and error later on.

Since $H \equiv \dot{a}/a$, the time taken for the universe to expand by an appreciable amount is of order the *Hubble time*

$$H_0^{-1} = 9.76h^{-1} \times 10^9 \,\mathrm{yr} \tag{3}$$

During a Hubble time, light travels a distance of order the *Hubble distance*,

$$H_0^{-1} = 2998h^{-1} \,\mathrm{Mpc} \tag{4}$$

Hubble's law applies only to galaxies whose distance is much less than the Hubble distance, it is based on the non-relativistic Doppler shift which requires $v \ll 1$. As we will discuss later, much more distant galaxies are also observed, the radius of the observable universe being of order the Hubble distance. Thus we are at the moment discussing only the *nearby* part of the observable universe. The universe has expanded by a negligible amount since light from the nearby universe was emitted, but by a significant amount since light from the distant universe was emitted.

The Big Bang

At the present epoch the Hubble time is of order 10^{10} yr. In what follows we shall extrapolate back to an era when the universe is very hot and dense, and the Hubble time is only a tiny fraction of a second. This era is popularly known as the Hot Big Bang, but one should not be misled by the language. Since the early universe is even more homogeneous and isotropic than the present one the expansion cannot be said to originate from a central point. It is certainly not an explosion, which by definition is driven by a pressure gradient. Thus the undoubted fact of the Big Bang does not explain the expansion of the universe; it has to be laid down at the very beginning.

When is the beginning? Presumably it lies at the *Planck epoch*, when the Hubble time is of order the Planck time

$$t_{\mathrm{Pl}} = G^{1/2} = 5.39 \times 10^{-44} \,\mathrm{sec} \tag{5}$$

As we extrapolate back to this epoch, quantum gravity effects presumably invalidate the concept of time, which conveniently removes the need to discuss what is before the Big Bang!

[1] To establish that an object *is* a standard candle one should ideally understand it theoretically, but in practice it is fairly safe to assume that a sufficiently distinctive object is standard. As a check one can compare two different candles, observing pairs of them which are known for to be in the same region of space because, for instance, they are in the same galaxy or star cluster. If they both have luminosity proportional to (distance)$^{-3}$, either they are both standard or they conspire to be non-standard in the same way.

1.4 The matter content of the universe

If gravity were negligible, $a(t)$ would increase linearly with time and $\dot{a}$ would be constant. In fact, gravity slows down the expansion making $\dot{a}$ a decreasing function of time. To calculate the effect of gravity, consider a test particle of unit mass, on the surface of a comoving sphere (one expanding with the universe). If ρ is the mass density of the universe and r is the radius of the sphere, the mass inside it is $(4\pi/3)r^3\rho$ and the potential energy of the particle is $-(4\pi/3)r^3\rho G/r$. The kinetic energy of the particle is $\dot{r}^2/2$ and therefore

$$\dot{r}^2/2 - (4\pi/3)r^3\rho G/r = E \tag{6}$$

where E is the total energy of the particle. Writing $r(t) = a(t)x$ and dividing through by $a^2/2$ one arrives at the *Friedmann equation*[2]

$$H^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2} \tag{7}$$

where $K = -2E$. If $K \leq 0$ (energy $E \geq 0$) the expansion will continue indefinitely whereas if $K > 0$ it will eventually give way to contraction leading to a 'Big Crunch'. The critical value separating these possibilities is $K = 0$, which corresponds to Friedmann equation

$$H^2 = \frac{8\pi G}{3}\rho \tag{8}$$

The corresponding mass density is called the *critical density*, and its present value is

$$\rho_{c0} = 1.88 \times 10^{-29} h^2 \,\mathrm{g\,cm^{-3}} \tag{9}$$

$$= 10.5 h^2 \,\mathrm{GeV\,m^{-3}} \tag{10}$$

$$= (3.0 \times 10^{-3}\,\mathrm{eV})^4 h^2 \tag{11}$$

It is convenient to define the *density parameter* by

$$\Omega = \rho/\rho_c \tag{12}$$

Since mass is conserved, $\rho \propto a^{-3}$. Putting this into the Friedman equation and remembering that $H = \dot{a}/a$, we have an expression for dt/da which can be integrated to give $t(a)$ and hence $a(t)$. The result depends on K, or equivalently on the present density parameter Ω_0. As I shall discuss in Section 6, inflation strongly suggests the value $\Omega_0 = 1$, and almost everyone working in the field believes that this is the true value.[3] In that case Eq. (8) gives $a =\propto t^{2/3}$. Another case which is easy to solve is $\Omega_0 \ll 1$, which corresponds to negligible gravity and $a \propto t$.

Baryonic matter

What is the observed value of Ω_0? Consider first the contribution of ordinary matter *ie* nuclei and electrons. In the context of cosmology this is usually called *baryonic matter* since the baryons (nuclei) vastly outweigh the electrons. From the nucleosynthesis calculation we know that the baryon contribution to Ω_0 is [4]

$$\Omega_B = (.013 \pm .002)h^{-2} < .09 \tag{13}$$

where I have used $.4 < h$ to obtain the upper limit. Thus if $\Omega_0 = 1$ there exists *non-baryonic dark matter*, whose nature I discuss later.

The luminous matter in the universe, consisting of stars and radiation-emitting gas, accounts for only $\Omega_B \sim .01$. Unless h is close to 1 (which we shall see in a moment is impossible with the favoured value total value $\Omega_0 = 1$), it follows from Eq. (13) that there is a lot of *baryonic dark matter*. If $\Omega_0 = 1$ it constitutes on average a few percent of the total amount of dark matter. The fraction is presumably more or less the same throughout inter-galactic space. Within a given galaxy, one

[2] It is often stated that Newtonian gravity cannot be used to discuss the expansion of the universe, and in particular that one cannot justify the neglect of matter outside the sphere in the above argument. This seems to me to be quite wrong. Drawing a much bigger sphere around the one that we considered, the matter within the bigger sphere can certainly be neglected. There remains the matter outside the sphere but it accelerates the small sphere and the test particle equally so it too can be ignored. This is exactly the same argument that justifies the use of Newtonian gravity for a galaxy or for the solar system, and it seems to be just as good in the cosmological case.

[3] In this context Ω includes all forms of energy density, but in the matter dominated era other forms of energy are supposed to be negligible except conceivably for that due to a cosmological constant as discussed later.

might expect the baryons to be concentrated more in the central, luminous part than in the dark halo. The reason is that baryons (ordinary matter) can emit radiation whereas non-baryonic dark matter interacts too weakly to do so (or it would not be dark). In consequence baryons can lose more energy, allowing them to settle more deeply into the galaxy centre. The baryons in a galaxy might be in the form of a non-emitting gas, or they might be failed stars with[4] mass perhaps $\sim .001$ to $.1M_\odot$, or dead stars (old white dwarfs, non-emitting neutron stars and the occasionally black hole arising from the collapse of an exceptionally massive star) with mass $\sim M_\odot$. These objects are called MACHOS (massive compact halo objects), because they will occur only in galaxy halos. (As far as we know, bound objects form only within galaxies so that the intergalactic baryons have to be in the form of non-emitting gas.)

Detection of MACHOS in our galaxy has recently been claimed through their gravitational lensing of stars (microlensing), a variation in brightness of the star occuring on a timescale of a few days as a MACHO crosses the line of sight [5]. If the observed MACHO density reflected the universal average baryon dark matter density, it would correspond to $\Omega_B \gtrsim .1$, but since only the central part of the galaxy is probed this need not be the case, and even a much higher MACHO mass density would be quite consistent with the nucleosynthesis value $\Omega_B \simeq .05$.

Non-baryonic dark matter

One can try to estimate the total amount of matter through its gravitational effect. The gravitational field in a bound system such as a galaxy or galaxy cluster can be deduced from the velocities of its components, as evidenced by the Doppler effect (the 'components' can be gas molecules, stars or whole galaxies). One finds that each galaxy is surrounded by a dark halo accounting for most of its mass, the total galaxy contribution being $\Omega_0 \simeq .1$. Galaxy clusters contain in addition inter-galactic gas, whose contribution to Ω_0 is not yet known.

On larger scales, where the universe is almost homogeneous and isotropic, one can in a similar spirit observe the small departure from uniform expansion. This defines a 'peculiar velocity' field, which is usually called the *bulk flow*. If one knew the bulk flow $\mathbf{v}$ *and* the density perturbation $\delta\rho/\rho$, one could deduce Ω_0 through the relation ([1], cf. Eq. (110))

$$\frac{\boldsymbol{\nabla}.\mathbf{v}}{3H_0} = -\frac{1}{3}(\Omega_0)^{.6}\frac{\delta\rho}{\rho} \tag{14}$$

To estimate these quantities one has to rely on observations of the galaxies. Largely on the basis of numerical simulations of gravitational collapse, it is generally assumed that their motion accurately measures $\mathbf{v}$, but that their density contrast is equal to a *bias factor* b times the underlying density contrast, with $b \sim 1$ within a factor of 2 or so (its value depends among other things on whether one is looking at optical or infrared galaxies). A recent study using this method [6] gives $(\Omega_0)^{.6} = 1.28^{+.75}_{-.59} b_I$ where b_I is the bias factor for the infrared galaxies in the IRAS catalogue. Including non-linear effects strengthens this result slightly, so that for example if one assumes $b_I > .5$ one deduces $\Omega_0 > .3$ at 95% confidence level. A smaller bias factor would certainly lead to problems with interpreting other data so it seems fair to say that the bulk flow indicates that $\Omega_0 > .1$. In view of the nucleosynthesis limit $\Omega_B < .1$ this means that *non-baryonic dark matter seems to be needed*.

Further evidence about Ω_0 comes from what one might call the 'standard model' of structure formation. This model, which has been the almost universal paradigm for many decades and is by far the most thoroughly explored and successful one, supposes that the structure originates as a Gaussian adiabatic density perturbation, whose spectrum at horizon entry is approximately scale independent. It requires that Ω_0 is at the upper end of the range 0 to 1, with $\Omega_0 = .1$ definitely excluded.

Summary

From nucleosynthesis, baryonic matter contributes $\Omega_0 \simeq .01$ to .09. The observed total Ω_0, especially if one accepts the standard model of structure formation, is higher, and the theoretically favoured value is $\Omega_0 = 1$. Thus non-baryonic dark matter is indicated. In addition, baryonic dark matter is indicated because luminous matter accounts for only $\Omega_0 \sim .01$.

[4] A collapsing gas cloud has to have mass $\gtrsim M_\odot$ to achieve a high enough temperature for nuclear reactions to start, but on the other hand very light gas clouds are thought not to collapse at all.

Value of Ω at early times

Even if Ω is not equal to 1 at the present epoch, it quickly approaches 1 as we go back in time because the first term in Eq. (7) is proportional to $\rho \propto a^{-3}$, and it therefore dominates the second term which is proportional to a^{-2}. If $\Omega_0 \ll 1$ we can easily estimate the epoch before which Ω is practically equal to 1 as follows. As long as $\Omega \ll 1$, gravity is negligible and therefore $\dot{a}$ is constant, leading to $H \propto a$ and $\Omega \propto \rho/H^2 \propto a^{-1}$. If follows that $\Omega \simeq 1$ before the epoch

$$a \simeq \Omega_0 \tag{15}$$

Since $\Omega_0 \gtrsim .1$, we see that Ω is certainly close to 1 before $a \sim .1$.

The age of the universe

One would like to define the age t of the universe at a given epoch as the time since $a = 0$ (in other words, to set $t = 0$ at $a = 0$). As we noted earlier the Planck scale puts an absolute limit on the earliest meaningful time, and more seriously the matter dominated era that we are treating here goes back only to $a \sim 10^{-4}$. So a practical definition of the age of the universe at a given epoch is simply the time since a was much less than the current value.

The Friedmann equation allows one to calculate the present age of the universe in terms of Ω_0. For $\Omega_0 = 1$ the behaviour $a \propto t^{2/3}$ gives at any epoch

$$H = \frac{2}{3t} \tag{16}$$

and therefore

$$t_0 = \frac{2}{3} H_0^{-1} = 6.5 \times 10^9 h^{-1}\,\mathrm{yr} \tag{17}$$

The smallest conceivable value $\Omega_0 \simeq .1$ gives

$$t_0 = .9 H_0^{-1} = 8.8 \times 10^9 h^{-1}\,\mathrm{yr} \tag{18}$$

An upper limit on the age of the universe is provided by the age of the oldest stars (observed in globular clusters) which is bigger than 1.0×10^{10} years. With the favoured value $\Omega_0 = 1$ this requires $h < .65$, whereas with $\Omega_0 = .1$ the limit is $h < .88$.

The measured value of the Hubble parameter

We noted earlier that using Hubble's law, different astronomers have produced estimates of h in the range $.4 < h < 1.0$. These Hubble law estimates are generally agreed to be very difficult, whereas the age limits that we have discussed are generally agreed to be sound. If $\Omega_0 = 1$, and if there is no cosmological constant (see below), one should therefore discard those Hubble law estimates which conflict with the age bound. Since both of these conditions are widely agreed to be likely, one concludes that the current best guess for the true value of the Hubble parameter is

$$.40 < h < .65 \tag{19}$$

1.5 High redshift observations

So far we have discussed only the relatively nearby part of the universe, located at a distance small compared with the Hubble distance. The redshift Eq. (1) from galaxies in this region can be written

$$z = \mathrm{d}a/a \ll 1 \tag{20}$$

where da is the change in the scale factor since the light was emitted. Thus, we see the nearby universe as it was in the very recent past, when the scale factor had almost its present value $a = 1$. Now we consider the distant universe, which is seen as it was in the distant past.

In order to discuss the distant universe we need general relativity. The central idea of GR, and practically the only one that I shall use in these lectures, is that all measurements are to be made locally. Thus, one populates the universe with comoving observers. In the homogeneous, isotropic universe comoving observers are in freefall (their worldlines are geodesics) and as a result their clocks can be synchronized once and for all. In other words, they all agree on the age of the universe t. The hypersurfaces in spacetime with fixed t are orthogonal to the comoving worldlines, and will be called *comoving hypersurfaces*.

The redshift measures the scale factor

To calculate the redshift from a distant object, emitting its light when $a \ll 1$, we can add up the redshifts seen by a sequence of comoving observers (observers moving with the expansion of the universe), each separated by a distance $\ll H_0^{-1}$. In the small region around each one Eq. (20) still applies except that it is evaluated at the time t when the radiation passes by. Using $H = \dot{a}/a$ and integrating gives

$$\frac{\lambda + \Delta\lambda}{\lambda} \equiv 1 + z = a^{-1} \tag{21}$$

The redshift measures directly the scale factor of the universe at the time of emission of the radiation.

The size of the observable universe

We would like to know the present distance of an object observed with redshift z, defined as the sum of the present distances between the comoving observers along its line of sight. The distance dr from one comoving observer to the next is proportional to $a(t)$ and so can be written $dr = adx$ where dx is independent of time. Thus each observer has a time independent *comoving coordinate* x, related to his distance from us by $r = ax$. Let us work out the trajectory $r(t) = x(t)a(t)$ of a photon moving towards us. During time dt it passes between observers with coordinate distance $x + dx$ and x, who are separated by distance $a\mathrm{d}x = \mathrm{d}t$ (*ie* thes e observers measure the photon speed to be 1). The comoving distance of the source from us is therefore $x = \int_{t_e}^{t_0} \mathrm{d}t/a$, where the subscript 0 denotes the present and the subscript e the time of emission. Its present distance is therefore

$$r(t_0, t_e) = a_0 \int_{t_e}^{t_0} \mathrm{d}t/a \tag{22}$$

(It is convenient to display the present value $a_0 = 1$ of the scale factor in this context.) Given the time dependence of a, this relation gives the present distance of an object observed with redshift z.

In the limit $z \gg 1$ it gives the present distance that we can see in principle, corresponding to the limit where the object emits its light at $t = 0$. This distance is called the *particle horizon* and is given by

$$r_{\mathrm{p.h.}}(t_0) = a_0 \int_0^{t_0} \frac{\mathrm{d}t}{a} \tag{23}$$

Assuming that $\Omega_0 = 1$ one has $a \propto t^{2/3}$ and therefore $r_{\mathrm{p.h.}}(t_0) = 2H_0^{-1}$.

There is nothing special about the present, so we can define the particle horizon $R_{\mathrm{p.h.}}(t)$ at any epoch,

$$r_{\mathrm{p.h.}}(t) = a(t) \int_0^t \frac{\mathrm{d}t}{a} = a(t) \int_0^a \frac{1}{aH} \frac{\mathrm{d}a}{a} \tag{24}$$

It gives the size of the biggest causally connected region, across which a signal with the speed of light can travel since $t = 0$.

As we noted earlier, the mathematical epoch $a = t = 0$ has to be replaced in practice by an epoch when a and t are much less than their current values. Thus, $r_{\mathrm{p.h.}}(t)$ can really only be defined as $r(t, t_e)$ where t_e is some early epoch. As long as gravity is attractive, $aH = \dot{a}$ increases as we go back in time, and from Eq. (24) $r(t, t_e)$ converges as $t_e \to 0$, and is typically of order $H^{-1}(t)$ as we just found for our particular case ($\Omega_0 = 1$ and matter dominating). But if the universe begins with an inflationary era, where gravity is by definition repulsive, $r(t, t_e)$ diverges. If we push t_e back to the beginning of inflation, it is much bigger than the Hubble distance $H^{-1}(t)$, so a region much bigger than the Hubble distance is causally connected.[5] But to achieve this causal connection at a given epoch, one has to work with signals emitted at a much earlier epoch, and what usually matters physically is the size of a region that can be causally connected by signals emitted in the last Hubble time or so, namely the Hubble distance. In the context of the early universe, the Hubble distance is usually referred to as 'the horizon', and has largely displaced the 'particle horizon' as the sort of horizon that cosmologists are interested in.

[5] At all epochs t up to the present, on the usual assumption that the observable universe is inside the Hubble distance at the beginning of inflation.

Non-Euclidean geometry?

To go further in interpreting high redshift observations, one needs to know the spatial geometry, *ie* the geometry of the fixed t hypersurfaces. Einstein's field equation gives the spatial line element as

$$\mathrm{d}\ell^2 = \frac{\mathrm{d}r^2}{1 - Kr^2/a^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2) \tag{25}$$

where K is the quantity that appears in the Friedmann equation. The distance element dl is that measured by a local comoving observer.

A surface with fixed coordinate r clearly has the geometry of a sphere in Euclidean space with radius r, and in particular its area is $4\pi r^2$. However the radial distance between adjacent spheres is $(1 - Kr^2/a^2)^{-1/2}\mathrm{d}r$, which is equal to dr only if $K = 0$, corresponding to $\Omega = 1$. Thus *space is non-Euclidean ('curved') if* $\Omega \neq 1$. If $\Omega < 1$ space is infinite just as in the Euclidean case, but if $\Omega > 1$ it has the remarkable property of being finite but unbounded. In particular, a surface made out of geodesics, such as the 'equatorial plane' $\phi = \pi/2$, has the geometry of a sphere whose curvature is $R^2 = K/a^2$. The quantity R is the radius that the sphere would have if it lived in Euclidean space, or equivalently it is defined by saying that the area of the sphere is $4\pi R^2$. The total volume of space, obtained by integrating the volume $4\pi r^2(1-Kr^2/a^2)^{-1/2}dr$ between adjacent spheres, is $16\pi^2 R^3$.

One calls K/a^2 the 'curvature of space' whether it is positive ($\Omega > 1$) or negative ($\Omega < 1$). The departure from Euclidean geometry becomes significant when $r^2 \gtrsim |a^2/K|$, and one refers to the distance $|a^2/K|^{1/2}$ as the *curvature scale*. If $\Omega \simeq 1$ the curvature scale is much bigger than the Hubble distance H^{-1}, if $|1 - \Omega| \ll 1$ it is of order H^{-1}, and if $\Omega \gg 1$ it is much less than H^{-1}.

High redshift observations

Having displayed the basic tools I will briefly say what we learn from observation of the distant universe. The most important finding is that the universe is definitely evolving. The most dramatic case is that of quasars (active galactic nuclei), whose abundance per comoving volume peaks at $z \sim 3$ or so. Neither quasars nor any other objects are observed at $z \gtrsim 5$. Ordinary galaxies as well as clusters are observed out to a redshift of order 1 to 2, and they too show signs of evolution. (The mere fact that galaxies have size of order one tenth of their spacing and do not expand with the universe means, of course, that they cannot have existed before $z \sim 10$.)

Ideally, high redshift observations plus an understanding of galactic evolution would give us information the value of Ω_0. In practice the information obtained in this way does not add anything to our knowledge as yet, except for the issue of a cosmological constant that we look at next.

The Robertson-Walker metric

Before leaving the description of the large scale universe I should write down the famous *Robertson-Walker spacetime metric*,

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t)\left[\frac{\mathrm{d}x^2}{1 - Kx^2} + x^2(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2)\right] \tag{26}$$

It expresses the fact that the universe is homogeneous and isotropic, with spatial geometry given by Eq. (25). The spacetime metric is the traditional starting point for cosmology but in these lectures I will consistently avoid it, in favour of the physical notion of comoving worldlines.

1.6 A cosmological constant?

The discussion so far assumes that a) gravity is described by Einstein's field equation, and b) the only significant source of gravity is the matter. People have considered the possibility of alter assumption a) by adding a *cosmological constant* Λ to the lagrangian of Einstein gravity. This is equivalent to to keeping Einstein gravity, but modifying assumption b) by invoking an additional source of gravity in the form of a fluid with pressure p_Λ and energy density ρ_Λ related by $p = -\rho$. This second viewpoint is the most useful and I adopt it here. From Eq. (43) below it then follows that ρ_Λ is time independent, so we are considering the possibility that

$$\rho(t) = \rho_m(t) + \rho_\Lambda \tag{27}$$

The corresponding present density parameter is

$$\Omega_0 = \Omega_m + \Omega_\Lambda \tag{28}$$

If we are ever faced with a measured value $\Omega_m < 1$, we will probably want to invoke a cosmological constant to keep the total value $\Omega_0 = 1$ which inflation suggests. A cosmological constant is, however unattractive from a particle physics viewpoint because the corresponding tiny vacuum energy density $(.001eV)^4$ (Eq. (11)) is difficult to understand. What about observation? Assuming that the total $\Omega_0 = 1$, the present the situation is as follows [7, 8].

On scales much less than the Hubble distance the effect of a cosmological constant is to add a repulsive term to Newtonian gravity, but the extra term is insignificant.

Going on to observations of the distant universe, gravitational lensing statistics give a remarkably good limit $\Omega_\Lambda \lesssim .6$ [8], and structure formation gives a similar but less reliable limit. Thus $\Omega_m > .4$, which means that one cannot use a cosmological constant to avoid the need for non-baryonic dark matter.

Coming to the age of the universe, Eq. (45) shows that Ω_Λ corresponds to repulsive gravity, $\ddot{a} > 0$, so the age of the universe will be an increasing function of Ω_Λ. In fact, the upper limit on h increased from .65 to .88 as Ω_Λ increases from 0 to .6. Thus a high value of h measured by, say, Hubble's law would be evidence in favour of a cosmological constant, and conversely a confirmation of the expected low value would be evidence against it.

In summary, a cosmological constant is theoretically objectionable, and is not so far required by observation.

2 The Radiation Dominated Era

2.1 Overview

An epoch in the early universe is conveniently specified by giving the temperature T in eV, MeV or GeV (as usual in cosmology I set $k_B = 1$). The epoch from 1 MeV to 1 eV, which can be taken to mark the end of the early universe, is well understood, and is highly constrained because it includes the epoch of nucleosynthesis. At earlier epochs one encounters a lack of knowledge about the nature of the quark-hadron and electroweak phase transitions, and eventually a basic ignorance of the fundamental interactions and the initial conditions. Later one enters the matter dominated era, when structure formation occurs and again our knowledge becomes less secure. The various epochs are summarised in Table 1.

2.2 The cosmic microwave background (cmb)

The most direct evidence for a hot early universe comes from the cosmic microwave background (cmb). It is extremely isotropic, with a very accurate blackbody distribution corresponding to temperature $T = (2.736 \pm .017)^0\,\mathrm{K}$.

To understand what the cmb implies, recall that the blackbody distribution specifies the photon occupation number of each momentum state as

$$f(p) = 2[e^{(p/T)} - 1]^{-1} \tag{29}$$

The factor 2 counts the number of spin states, and p is the momentum. Since there are $(2\pi)^{-3}\mathrm{d}^3p\mathrm{d}^3x$ states in a given volume of phase space, the number density n and the energy density ρ of the photons are given by

$$n = \frac{2}{(2\pi)^3}\int_0^\infty f(p)4\pi p^2\mathrm{d}p \tag{30}$$

$$\rho = \frac{2}{(2\pi)^3}\int_0^\infty Ef(p)4\pi p^2\mathrm{d}p \tag{31}$$

with $E = p$. Evaluating the integrals gives

$$\rho_\gamma = 2(\pi^2/30)T^4 \tag{32}$$

$$n_\gamma = 2(\zeta(3)/\pi^2)T^3 \tag{33}$$

TEMPERATURE	EVENT	CONSEQUENCES
$m_{Pl} = 10^{19}\,\mathrm{GeV}$	Inflation begins?	Immediate collapse avoided
10^{16}–$10^{11}\,\mathrm{GeV}$?	Cosmological scales leave the horizon	Field fluctuations become classical
10^{16}–$10^{11}\,\mathrm{GeV}$?	Inflation ends	Particles created
$10^{16}\,\mathrm{GeV}$–1 TeV??	Particles thermalize	Radiation domination begins
1 TeV-100 GeV	Electroweak phase transition	Particles acquire mass Baryogenesis occurs?
100 MeV	Chiral and quark/hadron phase transitions	Pions acquire mass and quarks bind into hadrons
1 MeV	Neutrinos decouple	Neutrons start to decay
1 MeV	$e\bar{e}$ annihilation	Afterwards $(T_\gamma/T_\nu) = (11/4)^{1/3}$
.1 MeV	Nucleosynthesis	Most neutrons bind into ^{4}He
1 eV	Matter domination begins	The cold dark matter density contrast starts to grow
.3 eV	Atoms form	Photons decouple and the baryonic matter density contrast grows
$\sim 10^{-3}\,\mathrm{eV}$?	First galaxies form	
$2.4 \times 10^{-4}\,\mathrm{eV}$	The present epoch	

Table 1: The early universe. The left hand column defines the epoch. It gives the temperature, which is roughly equal to $\rho^{1/4}$ where ρ is the energy density (in the first three rows the universe is not actually thermalised so the number in the left hand column is *defined* as $\rho^{1/4}$). During radiation domination the time since the big bang is roughly $1\,\mathrm{MeV}/T$ seconds, and during both epochs the scale factor is roughly $a \simeq T_0/T$ where $T_0 \simeq 10^{-4}\,\mathrm{eV}$ is the present temperature of the cosmic microwave background.

where $\zeta(3) = 1.2021$. The photon energy density is negligible compared with the mass density of the matter, corresponding to present density parameter

$$\Omega_\gamma \equiv \rho_\gamma/\rho_{c0} = 2.51 \times 10^{-5} h^{-2} \tag{34}$$

Dividing n by the baryon number density $n_B = 11.2 h^2\,\mathrm{m}^{-3}\Omega_B$,

$$\eta \equiv \frac{n_B}{n_\gamma} = 2.68 \times 10^{-8} h^2 \Omega_B = (3.4 \pm .6) \times 10^{-10} \tag{35}$$

(The last equality is the nucleosynthesis result Eq. (13).) Thus there are many photons per baryon.

Now let us evolve the blackbody distribution back in time, assuming that the number of photons in a comoving volume is conserved (as we shall see this is correct back to the epoch of electron-positron annihilation). Because of the redshift, the momentum of each photon goes like a^{-1}, which is equivalent to saying that the blackbody distribution is preserved, with $T \propto a^{-1}$. Since $T \propto a^{-1}$, $\rho_\gamma \propto a^{-4}$, so $\Omega_\gamma/\Omega_m \propto a^{-1}$. We conclude that *before the epoch* $a^{-1} \sim 10^4$ *the energy density of the cmb photons dominates that of the matter.*

2.3 Physics in the radiation dominated era

Motivated by the observed cmb, one makes the hypothesis that the early universe is a hot ideal gas. The hypothesis is found to be consistent in the context of gauge theories like the Standard Model. It leads to the prediction that the *nucleosynthesis* of ^{4}He and a few other light nuclei occurs in the early universe, and the calculated primordial abundances are in striking agreement with observation.

Adiabatic expansion

The pressure of a gas is given by $p = v^2\rho/3$, where v^2 is the mean square particle velocity and ρ is the energy density. It follows that

$$p = \rho/3 \qquad \text{relativistic} \tag{36}$$

$$p \simeq 0 \qquad \text{nonrelativistic} \tag{37}$$

As the universe expands, the change in the energy $E = a^3\rho$ in a comoving volume a^3 is given by $dE = -pd(a^3)$ so that

$$a\frac{\mathrm{d}\rho}{\mathrm{d}a} = -3(\rho + p) \tag{38}$$

For a nonrelativistic gas this gives $\rho \propto a^{-3}$, but for a relativistic gas it gives

$$\rho \propto a^{-4} \tag{39}$$

The extra factor a^{-1} arises from the redshift of the particle energies between collisions, each collision conserving energy.

The homogeneity and isotropy of the early universe implies that its expansion is adiabatic (no heat flow), so that entropy is conserved. Keeping only relativistic particles in the gbb, the entropy density is

$$s = \frac{\rho + p}{T} = \frac{4}{3}\frac{\rho}{T} = \frac{4}{3}\frac{\pi^2}{30}g_* T^3 \tag{40}$$

Conservation of the entropy $a^3 s$ in a comoving volume therefore gives the *relation between the temperature T and the scale factor a*

$$aT \propto g_*^{-1/3} \tag{41}$$

For rough estimates one can ignore the time dependence of g_* to deduce that $T \propto a^{-1}$, and can also set $s \sim n$ to deduce that *particle number is roughly conserved*. In particular, Eq. (35) holds at least roughly in the early universe,

$$n_B/n_\gamma \sim 10^{-10} \tag{42}$$

For a detailed calculation one has to take into account that aT increases abruptly whenever the temperature falls below the mass of a particle species. Physically, the relative increase in temperature comes from radiation emitted when the species annihilates and/or decays.

The Friedmann equation

According to general relativity, the Friedmann equation Eq. (7) that we derived from Newtonian gravity remains valid, if ρ denotes the energy density instead of just the mass density. One can differentiate it using Eq. (38), which is equivalent to

$$\dot{\rho} = -3H(\rho + p) \tag{43}$$

The result is

$$\dot{H} = -H^2 - \frac{4\pi G}{3}(\rho + 3p) \tag{44}$$

or

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p) \tag{45}$$

This last expression reduces to the Newtonian result in the limit $p \ll \rho$, if ρ is identified with the mass density. But in general both energy *and pressure* are sources of gravity.

Returning to the Friedmann equation, the density parameter Ω is still defined by Eq. (12). Setting it equal to 1, corresponding to $K = 0$, and remembering that for radiation $\rho \propto a^{-4}$, one finds

$$a \propto t^{1/2} \tag{46}$$

corresponding to

$$H = \frac{1}{2t} \tag{47}$$

Thermal equilibrium

In the early universe collision and decay processes are continually creating and destroying particles. Let us suppose that it is in thermal equilibrium, which means that each process is taking place at the same rate as its inverse. Then the the number of particles of a given species, per momentum state, is given by

$$f(p) = g[e^{(E-\mu)/T} \pm 1]^{-1} \tag{48}$$

In this expression, g is the number of spin states, p is the momentum, and E is the energy, given in terms of the mass m by $E = \sqrt{p^2 + m^2}$. The sign is + for fermions (ensuring that there is

at most one particle per quantum state), and − for bosons. The quantity μ (which depends on temperature) is called the chemical potential of the species. Given the chemical potential, this expression determines the number density n and energy density ρ of the species through Eqs. (30) and (31).

The chemical potential is conserved in every collision, so that if a process AB → CD occurs then $\mu_A + \mu_B = \mu_C + \mu_D$. In the early universe all known particle species are freely created and destroyed (provided that their mass does not exceed the typical collision energy, which we shall see in a moment is of order T). The only significant restriction is that each collision must respect the conservation of any 'charges' that the particles carry. In the context of the Standard Model there are five such charges, namely electric charge, baryon number and the three lepton numbers.[6] Since the photon carries none of these 'charges', a single photon can be created through processes of the form A → Aγ, which implies that $\mu = 0$ for the photon, leading to the blackbody distribution. The same goes for any particle which is its own antiparticle, so it too has $\mu = 0$. If the antiparticle is distinct one can create a particle-antiparticle pair through some process of the form A → A+pair, which implies that the particle and its antiparticle have opposite μ. As a result, μ vanishes if the number densities n and $\bar{n}$ of particle and antiparticle are equal, and otherwise is determined by the imbalance $n - \bar{n}$.

More generally, one can show that until weak interaction equilibrium fails at $T \sim 1\,\mathrm{MeV}$, enough reactions occur to determine all of the chemical potentials in terms of the densities of the five conserved 'charges'. Furthermore, all of the chemical potentials are zero if the 'charges' are all zero. In that case one has what I shall call the *generalized blackbody distribution* or gbb

$$f(p) = g[e^{E/T} \pm 1]^{-1} \tag{49}$$

Putting this into Eqs. (30) and (31) gives n and ρ. If the temperature is well above the mass of the species in question, $m \ll T$, then one is in *relativistic regime* and it is a good approximation to set $E = p$. Using Eqs. (30) and (31) gives

$$\rho = \begin{cases} (\pi^2/30)gT^4 & \text{bosons} \\ (7/8)(\pi^2/30)gT^4 & \text{fermions} \end{cases} \tag{50}$$

$$n = \begin{cases} (\zeta(3)/\pi^2)gT^3 & \text{bosons} \\ (3/4)(\zeta(3)/\pi^2)gT^3 & \text{fermions} \end{cases} \tag{51}$$

Thus *according to the gbb, each relativistic species contributes* $\sim T^4$ *to* ρ *and* $\sim T^3$ *to* n. (Note that the typical energy $E \sim \rho/n$ is of order T, so this is indeed the relativistic regime.) As T falls below m one moves into the nonrelativistic regime, and according to the gbb ρ and n fall rapidly (like $e^{-m/T}$). The physical reason is that the typical energy E available in a collision is now insufficient to create the species.

Is the gbb valid in the early universe? The charge density of the universe is certainly zero to very high accuracy, or the expansion of the universe would be governed by electrical repulsion instead of gravity. The baryon number is not zero, but it is small in the sense that

$$\eta \equiv n_B/n_\gamma \ll 1 \tag{52}$$

The three lepton number densities cannot be measured directly, but let us suppose that they too are small in the same sense. It can be shown that if this is so, the gbb is valid to high accuracy for all relativistic species, the reason being that for each species one then has $|n - \bar{n}| \ll n_\gamma \sim T^3$. From now on I assume that the gbb is valid for all relativistic species in equilibrium, and the success of the nucleosynthesis calculation will support that assumption.

As the temperature falls below the mass m of a given species, particle-antiparticle pairs at first rapidly annihilate as dictated by the gbb. Then the small particle-antiparticle imbalance becomes significant, and annihilation soon stops because only particles (or only antiparticles) survive. Even if they do not decay, the surviving particles give a negligible contribution to ρ during radiation domination, and an even more negligible contribution to n.

As long as a non-relativistic particle species is in equilibrium, it follows from Eqs. (48) and (30) that its number density is

$$n = g\left(\frac{mT}{2\pi}\right)^{3/2} e^{-(m-\mu)/T} \tag{53}$$

[6] The non-perturbative violation of the last four is not relevant in this context.

We shall use this equation to determine the relative numbers of different species of nuclei in thermal equilibrium, and the degree of ionization of a gas in thermal equilibrium.

The upshot of this discussion is that unless one happens to be at an epoch when T is close to one of the particle masses, ρ and n come just from the relativistic species. In particular,

$$\rho = (\pi^2/30) g_*(T) T^4 \tag{54}$$

where the *effective number of degrees of freedom* g_* is given by

$$g_*(T) = \sum_{\text{bosons}} g + \frac{7}{8} \sum_{\text{fermions}} g \tag{55}$$

where the sums go over particle species with $m < T$. According to the Standard Model it is of order 10 for $1\,\text{MeV} \lesssim T \lesssim 10\,\text{MeV}$ and of order 100 for $T \gtrsim 1\,\text{GeV}$. Extensions of the Standard Model, notably supersymmetry, introduce new particles which can increase g_* by a factor of a few at $T \gtrsim 100\,\text{GeV}$, but usually leave it unchanged at lower energies.

Eq. (3) applies only if all relativistic species are in thermal equilibrium. This is true in the Standard Model (for $T \gtrsim 1\,\text{MeV}$), but not necessarily in extensions of it.

The timescale

Using Eqs. (54), (8) and (47), the timescale during radiation domination is

$$\frac{t}{1\,\text{sec}} = 2.42 g_*^{-1/2} \left(\frac{1\,\text{MeV}}{T} \right)^2 \tag{56}$$

2.4 The Standard Model of the early universe

Interaction rates and the ideal gas condition

Our basic assumptions are that the early universe is an ideal gas, and that it is in thermal equilibrium. The first requires that the interactions are not too strong, but the second requires that they are strong enough to make the relevant reaction rates (per particle) be bigger than the rate of expansion H of the universe. Which reactions are 'relevant' depends on what one supposes about the chemical potentials; each reaction that is in equilibrium implies a relation between the chemical potentials (if it creates or destroys particle species) and there have to be enough relations to ensure the postulated conditions on the chemical potential. In particular, if we want the gbb to hold for relativistic particle species, enough interactions have to be in thermal equilibrium to ensure that all of the chemical potentials vanish (on the assumption that the 'charge' densities all vanish). In the context of the standard model, one can show that this is indeed ensured if all processes of the form AB $\to$ CD going by the exchange of a single gauge boson are in equilibrium. Let us check that such reactions are indeed in thermal equilibrium, and that the interactions are yet weak enough to ensure the ideal gas condition.

Suppose first that the gauge boson is massless; this covers the case of the electromagnetic interaction before photon atoms form (photon decoupling) the strong interaction before hadrons form (the QCD phase transition) and the weak interaction before the W and Z acquire mass (the electroweak phase transition). For an isolated collision the cross section σ is infinite because the interaction has infinite range, but in a gas there is a cutoff at the particle spacing $n^{-1/3}$ (Debye shielding). Using $n \sim T^3$ one finds that $\sigma \sim \alpha^2/T^2$, where α is the electromagnetic, weak or strong coupling strength as appropriate. If at least one of the initial particles is relativistic, the rate per particle is $\Gamma \sim n\sigma$ (the formula is exact if one particle is at rest, and a factor 2 too small for a head-on relativistic collision.) Using Eqs. (47) and (56) this gives $\Gamma/H \sim .1\alpha^2 m_{Pl}/T$. Taking the high energy estimate $\alpha \sim .1$, this is indeed $\gg 1$ unless $T \gtrsim 10^{16}\,\text{GeV}$. Thus the interaction is strong enough to maintain thermal equilibrium, except at these very high temperatures. What about the ideal gas condition? We can take it to be the requirement that the mean free path r, determined for relativistic particles by $r\sigma n \sim 1$, is big compared with the particle spacing $n^{-1/3}$. Thus we require $n^{2/3}\sigma \ll 1$ which is equivalent to $\alpha \ll 1$. This is satisfied, except for the strong interaction just before the QCD transition.

Now consider the weak interaction after the electroweak phase transition. The cross section is $\sigma \sim \alpha^2 T^2/m_W^4$, which gives $\Gamma/H \sim m_{Pl}\alpha^2 T^3/m_W^4$. Taking the weak interaction coupling as $\alpha \sim .01$ this is equal to 1 when

$$T \sim \alpha^{-2/3} (m_W^4/m_{Pl})^{1/3} \sim 1 \text{ to } 10\,\text{MeV} \tag{57}$$

Weak interactions are in equilibrium only above this temperature. Since the cross section is a factor $(T/m_W)^4$ smaller than in the previous case, the ideal gas condition is amply satisfied for the weak interaction.

2.5 The history of the early universe

The electroweak phase transition

The electroweak symmetry of the standard model is restored above some critical temperature $T_c \sim 100\,\mathrm{GeV}$. The Higgs field is then zero, corresponding to a false vacuum with energy density $\sim (100\,\mathrm{GeV})^4$, and all particle masses vanish. As the temperature falls through T_c the Higgs field rapidly settles to its true vacuum value and particles acquire mass.

The consequences of a phase transition can be very different, depending on whether it is of first order or second order. In the first case the transition is relatively violent because bubbles of the new phase are produced. In the second case it is more gently, the new phase being produced smoothly everywhere (except near topological defects, which are not produced at the electroweak phase transition).

It is not clear if the electroweak transition is of first or second order, and the answer in any case should probably take into account an extension of the Standard Model. If it is of first order it will violate baryon- and lepton number conservation because of non-perturbative effects occurring in bubble walls. As Shaposhnikov explains in his lectures, such a mechanism might be responsible for generating the observed ratio $n_b/n_\gamma \sim 10^{-10}$, either within the standard model or within some extension of it.

The quark-hadron transition

As the temperature falls through $\Lambda_{\mathrm{QCD}} \sim 100\,\mathrm{MeV}$, chiral symmetry is spontaneously broken which, among other things, gives mass to the pion. More importantly, the *quark-hadron* phase transition occurs, which binds quarks and gluons into hadrons as their typical spacing $\sim T^{-1}$ becomes bigger than $\Lambda_{\mathrm{QCD}}^{-1} \sim 1\,\mathrm{fm}$.

The quark-hadron phase transition [9] is thought to be a second order one. If it turned out to be first order though, the bubbles could lead to dramatic effects, perhaps interfering with nucleosynthesis or even creating black holes accounting for the dark matter. The mass of such black holes might be of order $1M_\odot$ corresponding to Eq. (72) below, or much smaller.

Neutrino decoupling and e$\bar{\mathrm{e}}$ annihilation

At $1\,\mathrm{MeV} \lesssim T \lesssim 100\,\mathrm{MeV}$, the only particles present (according to the Standard Model) are photons, free protons and neutrons, electrons, positrons, and the three neutrino species plus their antiparticles. Other elementary particles as well as hadrons are absent because they are unstable, and are too heavy to be created in a typical collision ($m \gg T$). Nuclei, not to mention atoms and molecules, are absent because their binding energy is much less than T, so that collisions would destroy them.

The protons and neutrons turn into each other through reactions like $\mathrm{ep} \leftrightarrow \nu_e\mathrm{n}$, $\bar{\mathrm{e}}\mathrm{n} \leftrightarrow \bar{\nu}_e\mathrm{p}$ and $\mathrm{n} \leftrightarrow \mathrm{pe}\bar{\nu}_e$. As a result the ratio n/p of their number densities is determined by thermal equilibrium. The chemical potentials satisfy $\mu_p - \mu_n = \mu_{\nu_\mathrm{e}} - \mu_e$, with μ_{ν_e} and μ_e both $\ll T$ (on the assumption that the lepton number densities are sufficiently small that the e and ν_e satisfy the gbb to good accuracy). From Eq. (53) with $|\mu_p - \mu_n| \ll T$, it follows that the neutron to proton ratio is

$$n/p = e^{-\Delta m/T} \tag{58}$$

where $\Delta m \equiv m_n - m_p = 1.3\,\mathrm{MeV}$.

When the temperature falls through 1 MeV, three essentially unrelated things happen to occur at about the same time.

- Neutrinos decouple because of Eq. (57)
- Electron-positron annihilation occurs because $m_e \sim 1\,\mathrm{MeV}$
- Neutron creation stops because of Eq. (57).[7]

[7] It would stop anyway, because to create a neutron requires a collision energy bigger than $m_n - m_p = 1.3\,\mathrm{MeV}$.

A detailed calculation shows that the events actually occur in the indicated order; let us discuss them in turn.

After decoupling the neutrinos travel freely, maintaining their gbb with temperature $T_\nu \propto a^{-1}$. The photon temperature, though, is relatively raised by e$\bar{\text{e}}$ annihilation. Before annihilation there are electrons, positrons and photons in equilibrium giving $g_* = 11/2$, whereas after there are only photons giving $g_* = 2$. From Eq. (41) it follows that after annihilation, $T_\gamma = (11/4)^{1/3} T_\nu$. Remembering that there are two spin states for the photon and one for each massless neutrino or antineutrino species, Eq. (51) gives

$$n_\nu = (3/11) n_\gamma \tag{59}$$

for each neutrino species. If all three neutrino species are massless, Eqs. (33) and (51) give the neutrino contribution to Ω_0,

$$\Omega_\nu = (21/8)(4/11)^{4/3} \Omega_\gamma = .681 \Omega_\gamma \tag{60}$$

Using Eq. (34) the epoch of matter-radiation equality is therefore

$$z_{\text{eq}} = 2.37 \times 10^4 h^2 \tag{61}$$

After neutrons cease to be created, the np ratio is frozen in except for the slow decrease caused by neutron decay. Its initial value is $n/p \simeq 1/6$, and by the time that nucleosynthesis occurs at $T \simeq .1\,\text{MeV}$ it has fallen to

$$n/p \simeq 1/7 \tag{62}$$

Nucleosynthesis

A nucleus is defined by the number Z of protons, and the number $A - Z$ of neutrons. If strong nuclear reactions occur at a sufficiently high rate, while the weak interaction changing protons to neutrons is negligible, the chemical potential of a given nucleus satisfies $\mu = Z\mu_p + (A - Z)\mu_n$. By virtue of this relation, the μ's cancel in the ratio $n_p^Z n_n^{A-Z}/n$ given by Eq. (53), leading to the following number density for a given nucleus

$$n = g A^{3/2} 2^{-A} \left(\frac{2\pi}{m_N T} \right)^{3(A-1)/2} n_p^Z n_n^{A-Z} e^{B/T} \tag{63}$$

where B is its binding energy. Together with the np ratio and the total baryon density n_B this determines all of the nuclear densities. Using $n_B = \eta \text{n}_\gamma$ together with Eq. (33) for the photon number density n_γ, one can calculate the mass fraction in a given nucleus,

$$\begin{aligned} X &= g[\zeta(3)^{A-1} \pi^{(1-A)/2} 2^{(3A-5)/2}] A^{5/2} \eta^{A-1} \\ &\times X_p^Z X_n^{A-Z} \left(\frac{T}{m_N} \right)^{3(A-1)/2} e^{B/T} \end{aligned} \tag{64}$$

If the baryon to photon ratio η were of order 1, this expression would say that the nuclei form (X becomes of order 1) at roughly $T \sim B$, which is the epoch after which a *typical* photon has too little energy to destroy the nucleus. But because there are $\eta^{-1} \sim 10^{10}$ photons per baryon, the nuclei do not form until T is well below B. The reason is that until that happens there are still a lot of exceptionally energetic photons around with energy of order B.

To determine whether a given set of nuclei will actually be in thermal equilibrium at a given epoch in the early universe, one has to take the rates of the relevant nuclear reactions from accelerator experiments. (They cannot all be calculated accurately from first principles, and are certainly not amenable to the order of magnitude arguments that work for elementary particles.) One finds that at $T \gtrsim .3\,\text{MeV}$ the lightest few nuclei are indeed in thermal equilibrium, but that their number densities are negligible. In the range $.1\,\text{MeV} \lesssim T \lesssim .3\,\text{MeV}$, thermal equilibrium predicts that practically all of the neutrons are bound into ^{4}He, but when T first enters this range thermal equilibrium actually fails badly because the reactions forming ^{4}He do not occur fast enough to keep up with the 'demand' for ^{4}He (the deuterium bottleneck). As a result most of the neutrons bind into ^{4}He at $T \sim .1\,\text{MeV}$, rather than at $T \simeq .3\,\text{MeV}$ as nuclear equilibrium would predict. Assuming that all of the neutrons bind into ^{4}He, its abundance by weight is

$$X = \frac{2n}{n+p} \tag{65}$$

Figure 1: Nucleosynthesis predictions and observations, taken from [4]. The upper section gives the ^{4}He abundance by weight, and the lower section gives the other abundances by number. The horizontal lines indicate the observed values; from top down the upper and lower limits for ^{4}He, the upper limit for D+^{3}He, the lower limit for ^{3}He and the upper and lower limits for ^{7}Li. The horizontal axis is $10^{10}\eta$, and consistency with all observations requires that it be in the shaded region.

Taking $n/p \simeq 1/7$ gives $X \simeq 22\%$.

The assumption of thermal equilibrium gives a reasonable estimate of the ^{4}He abundance, but an accurate calculation of it and even an estimate of the other abundances requires an out of equilibrium calculation. This has to be done on a computer, taking the rates for the various nuclear reactions from laboratory experiments. The latest results from such calculations [4] are shown in Figure 1 as a function of η.

The primordial abundances of ^{4}He and other light nuclei can be measured by a variety of techniques, and the predicted abundances agree provided that η is in the interval indicated, which corresponds to the result already quoted in Eq. (35). The agreement is extremely impressive, and constitutes one of the strongest reasons for believing the hypothesis that went into the calculation. It also limits greatly ones freedom to alter them; in particular it severely constrains possible extensions of the Standard Model.

Photon decoupling

After nucleosynthesis there are photons, protons, ^{4}He nuclei and electrons in thermal equilibrium. When the temperature falls sufficiently, most of the electrons bind into hydrogen atoms, and one can use the thermal equilibrium fraction of hydrogen to determine when this happens. The atoms form through the process $\mathrm{p} + \mathrm{e} \to \mathrm{H} + \gamma$, so that $\mu_p + \mu_e = \mu_H$. The chemical potentials cancel in the ratio $n_p n_e / n_H$ given by Eq. (53), leading to

$$n_H = n_p n_e \left(\frac{m_e T}{2\pi} \right)^{-3/2} e^{B/T} \tag{66}$$

Here $B = m_p + m_e - m_H = 13.6\,\mathrm{eV}$ is the binding energy of hydrogen, and I have used $g_p = g_e = 2$ and $g_H = 4$. Ignoring the ^{4}He for simplicity, $n_e = n_p$ and $n_B = n_p + n_H$, which with Eq. (33)

determines the fractional ionization $X \equiv n_p/n_B$,

$$\frac{1-X}{X^2} = \frac{4\sqrt{2}\zeta(3)}{\sqrt{\pi}}\eta\left(\frac{T}{m_e}\right)^{3/2} e^{B/T} \tag{67}$$

This is called the *Saha equation*. As with nuclei, very few atoms form until T is well below B, because there are so many photons per baryon. In fact, one finds that $T \simeq .3\,\mathrm{eV}$ when X has fallen to 10%. The corresponding redshift is $z \simeq 1300$. Soon after this, at $z \simeq 1100$ to 1200, the photons decouple.

2.6 Beyond the Standard Model

There are several theoretical reasons for wanting to extend the Standard Model.

- One-loop contributions to the Higgs mass have to be cut off at the scale 100 GeV to 1 TeV, or they will be too big. To impose the cutoff one can either make the Higgs composite, as in technicolour type models, or invoke supersymmetry. Composite models are becoming increasingly unviable in the light of the ever more impressive agreement between the Standard Model and collider experiments, notably at CERN, and I will not consider them here. Supersymmetry requires that each known particle species has an as yet undiscovered partner. The lightest supersymmetric partner is typically weakly interacting and stable, and is therefore a dark matter candidate.

- Unless one of the quark masses vanishes, a global $U(1)$ symmetry called Peccei-Quinn symmetry is presumably needed to explain why the strong interaction is CP invariant. To avoid conflict with accelerator physics and astrophysics, the symmetry is broken at a very high energy scale $\gtrsim 10^{10}\,\mathrm{GeV}$.

- Going to a still higher scale, one would like to make contact with gravity, presumably involving superstring degrees of freedom around the Planck scale $\sim 10^{19}\,\mathrm{GeV}$.

- One may wish to invoke a GUT. Superstring theory can apparently relate couplings etc without one, but on the other hand the ratio of the strong, weak and electromagnetic interaction strengths predicted by supersymmetric GUTS agrees with experiment amazingly well, with a unification scale $\simeq 10^{16}\,\mathrm{GeV}$.

An observational reason for wanting to extend the Standard Model might be the desire to introduce neutrino masses to solve the solar neutrino problem as well as a problem with atmospheric neutrinos.

Because the cosmological predictions of the Standard Model agree so well with observation, many otherwise reasonable extensions of the Standard Model can be ruled out. Particularly sensitive are nucleosynthesis, and the absence of significant corrections to the shape of the cmb spectrum.

From the many things that could be discussed, I choose three of the most topical.

Neutrino masses

A massive neutrino species could be a dark matter candidate. From Eqs. (11), (33) and (59), its contribution to the density parameter is

$$\Omega_\nu = .109h^{-2}\left(\frac{m_\nu}{10\,\mathrm{eV}}\right) \tag{68}$$

A mass of order 10 eV is therefore required. Experiments observing β-decay practically rule out such a mass for the ν_e, but it is allowed for the ν_μ or ν_τ.

A firmer reason for wanting some neutrino species to have mass comes from the solar neutrino problem, whose favoured solution requires either the ν_μ or the ν_τ to have a mass of about $3\times10^{-3}\,\mathrm{eV}$ (with the ν_e mass much smaller).

A natural mechanism for generating neutrino masses is the see-saw mechanism, which invokes a right-handed neutrino with very large mass M, related to the lepton mass m and neutrino mass m_ν of each generation by

$$m_\nu = m^2/M \tag{69}$$

Taking, without very good justification, M to be the same for all three generations, the choice $M \sim 10^{12}\,\mathrm{GeV}$ gives $m_{\nu_\tau} \sim 10\,\mathrm{eV}$ and $m_{\nu_\mu} \sim 3\times10^{-3}\,\mathrm{eV}$, so that the first neutrino can be the dark matter and the second can solve the solar neutrino problem.

Neutrino dark matter is called hot, because it remains relativistic until the epoch when most cosmological scales of interest have entered the horizon (come into causal contact), and therefore cannot initially undergo gravitational collapse.

Cold and warm dark matter

Cold dark matter is by definition non-relativistic when all cosmological interesting scales enter the horizon. The lightest supersymmetric particle is a cold dark matter candidate. Its interaction strength is similar to that of the neutrino, but because it is much heavier it will be non-relativistic when it falls out of equilibrium. Its contribution to Ω_0 turns out to be

$$\Omega_{\rm lsp} \sim \left(\frac{m}{1\,{\rm TeV}}\right)^2 h^{-2} \tag{70}$$

Since one needs $m \sim 100\,{\rm GeV}$ to $1\,{\rm TeV}$ for supersymmetry to work, this is automatically of order 1!

Warm dark matter by definition remains relativistic until a cosmologically interesting epoch, which is however significantly earlier than the epoch for neutrino dark matter. Candidates include a right handed neutrino, a majoran, or a particle with interaction strength very much less than that of neutrinos with a mass of order $1\,{\rm keV}$.

The dark matter candidates discussed so far are known collectively as WIMPS (weakly interacting massive particles). One can hope to detect a WIMP in the laboratory by searching for its rare interactions, the detectors being of the same type as the ones that see solar neutrinos or do neutrino astronomy (so far only with supernova 1987A), and place limits on proton decay. One might also hope to detect WIMPS in the galaxy by looking for photons produced by their annihilation. Finally one might be able to create them at accelerators.

A quite different dark matter candidate is the axion, which is never in thermal equilibrium and would be cold dark matter. The axion is the Goldstone boson of Peccei-Quinn symmetry, which is the best bet for ensuring the CP invariance of the strong interaction (except perhaps for making one of the quark masses zero). Various possibilities exist for axion cosmology in different regimes of parameter space [10], but typically axions are emitted by Peccei-Quinn strings at $T \sim 1\,{\rm GeV}$ and quickly become non-relativistic. The axion contribution to Ω_0 is roughly

$$\Omega_a \sim (10^{-4}\,{\rm eV})/m_a \tag{71}$$

Accelerator experiments plus the requirement that axion emission should not have drained too much energy from supernova 1987A require that $m_a < 10^{-3}\,{\rm eV}$. Since $\Omega_a \lesssim 1$, we conclude that if the axion exists at all its mass is roughly of order 10^{-3} to $10^{-4}\,{\rm eV}$, and it gives Ω_a very roughly of order 1! One might detect solar axions through their tiny electromagnetic interaction, by exploiting the fact that their wavelength $(10^{-4}\,{\rm eV}/m_a) \times .197\,{\rm cm}$ is macroscopic.

In the context of supersymmetry the axion must have two supersymmetric partners, the axino (spin 1/2) and the saxino (spin 0), with rather well defined properties. The former might be a dark matter candidate and the latter might dilute the entropy by its decay, reducing predicted cosmological mass densities like Eqs. (75) and (71) by a factor of up to 10^{-4} [11].

These dark matter candidates are relatively uncontroversial, involving rather minimal extensions of the Standard Model. More ambitious extensions can produce dark matter with practically any desired properties; thus, since MDM (mixed, *ie* hot plus cold, dark matter) has proved so successful at fitting the observations, no fewer than three separate proposals have been made for naturally generating it in the early universe [12, 13, 14].

A completely different possibility is that the non-baryonic dark matter consists of black holes which form before nucleosynthesis. In that case one has no need of non-baryonic dark matter particles. A black hole will form if there is an overdense region with density contrast $\delta\rho/\rho \sim 1$, whose size is of order the Hubble distance. All of the energy in the overdense regions then ends up in the black hole, whose mass is therefore equal to ρ_r/ρ_m times the mass of the matter in a Hubble volume,

$$M \sim 10^{-1}(\rho_r/\rho_m)\left(\frac{1\,{\rm MeV}}{T}\right)^3 \sim 10^5\left(\frac{1\,{\rm MeV}}{T}\right)^2 \tag{72}$$

As we shall see later, the density contrast at horizon entry is only of order 10^{-5} on the scales $M \gtrsim 10^{10} M_\odot$ explored by large scale structure and the cmb anisotropy, and simple inflation models predict that this remains true down to very small scales. The simple models could be wrong however. One might also form smaller black holes on scales much less than the horizon size at a phase

transition. Light enough black holes could be evaporating during nucleosynthesis, or at the present, with dramatic effects.

Finally, I note that although WIMP or axionic dark matter is normally assumed to be gaseous, it might be bound into macroscopic objects. To achieve this during radiation domination one needs an isocurvature matter density perturbation (one not shared by the radiation) with most of its power on small scales. Such a perturbation might be generated during inflation, like the usually considered adiabatic density perturbation, but it could also be generated at a phase transition by, say, a collapsing bubble. In that case the maximum mass of the objects formed is that within a Hubble volume,

$$M \sim .1(1\,\mathrm{MeV}/T)^3 M_\odot \tag{73}$$

This mechanism has been considered in the context of axions, where the phase transition is at $T \sim 1\,\mathrm{GeV}$ giving objects with $M \sim 10^{-10} M_\odot$ [15].

Topological defects

In the early universe, the Higgs fields (and any other scalar fields present) can find themselves in a false vacuum, corresponding to a local minimum of their potential. The false vacuum is typically stabilized by finite temperature effects, but an interaction with the inflaton field can also to the job [16, 2, 17, 18, 19]. When the temperature, or the inflaton field, falls below some critical value, the false vacuum is destabilized and there is a phase transition to the true vacuum. Within the Standard Model the electroweak transition is the only example of this phenomenon, but further transitions could occur, at much higher energy scales, in extensions of it.

Such a phase transition can leave behind it topological defects, consisting of regions of space where the field is still trapped in the false vacuum. Whether this happens or not depends on the nature of the symmetry breaking. No topological defects form at the electroweak transition, but they might well form in transitions occurring in extensions of the Standard Model. A concrete example is provided by Peccei-Quinn symmetry breaking, which produces global strings which play an important role in producing axions before they disappear at $T \sim 1\,\mathrm{GeV}$. Another example is the breaking of a GUT which inevitably produces magnetic monopoles. They are so abundant as to be cosmologically forbidden unless one either finds a way of annihilating them, or invokes inflation.

3 The Evolution of the Density Perturbation

Cosmological perturbation theory develops linear equations for perturbations away from homogeneity and isotropy. Using it one can follow their growth on a given scale, until they become big enough for gravitational collapse. On scales $\gtrsim 100\,\mathrm{Mpc}$, where collapse has yet to occur, cosmological perturbation theory can be used right up to the present epoch. On smaller scales the only sure-fire way of performing calculations after perturbation theory fails is to perform numerical simulations, though analytic approximations can provide some insight. In these lectures I concentrate on the linear regime.

In the Newtonian regime cosmological perturbation theory has long been recognised to be a straightforward application of fluid flow equations, taking into account where necessary particle diffusion and free-streaming (collisionless particle movement). In the relativistic regime, cosmological perturbations was first discussed by Lifshitz [20] in 1946. His formalism, which has been the starting point for most subsequent work including the influential 'gauge invariant' formalism of Bardeen [21, 22], considers the perturbed Robertson-Walker metric. An alternative formalism, which makes no mention of the metric perturbation and works instead with relativistic fluid flow equations, was initiated by Hawking [23] in 1966. This approach, which treats the Newtonian and relativistic regimes in a unified way, is becoming increasingly popular [24, 25, 26, 27, 28] and is the one that I will use in these lectures.

3.1 Relativistic fluid flow

We populate the universe with comoving observers, who define physical quantities in their own region. By definition, the momentum density is zero with respect to a comoving observer; such an observer moves with the energy flow.

A crucial concept is the *velocity gradient* u_{ij}. It is defined by each comoving observer, using locally inertial coordinates in which he is instantaneously at rest, as the gradient of the velocity u^i

of nearby comoving worldlines,

$$u_{ij} \equiv \partial_j u^i \tag{74}$$

The velocity gradient can be uniquely decomposed into an antisymmetric vorticity ω_{ij}, a symmetric traceless shear σ_{ij}, and a locally defined Hubble parameter H,

$$u_{ij} = H\delta_{ij} + \sigma_{ij} + \omega_{ij} \ . \tag{75}$$

In the limit of homogeneity and isotropy, $\sigma_{ij} = \omega_{ij} = 0$. One can show from angular momentum conservation that ω_{ij} decays like $(\rho + p)a^{-5}$, so it is presumably negligible. We will see how to calculate σ_{ij} in Section 3.3.

Just as for a homogeneous isotropic universe, it is useful to consider 'comoving hypersurfaces', defined as those orthogonal the fluid flow worldlines.[8] On a given hypersurface, each quantity ρ, p and H can be split into an average plus a perturbation,

$$\rho(\mathbf{x}, t) = \bar{\rho}(t) + \delta\rho(\mathbf{x}, t) \tag{76}$$

$$p(\mathbf{x}, t) = \bar{p}(t) + \delta p(\mathbf{x}, t) \tag{77}$$

$$H(\mathbf{x}, t) = \bar{H}(t) + \delta H(\mathbf{x}, t) \tag{78}$$

Here t is the time coordinate labelling the hypersurfaces, and $\mathbf{x} = (x^1, x^2, x^3)$ are space coordinates. We would like to choose the space coordinates to be comoving coordinates, related to Cartesian coordinates by $r^i = ax^i$, with a the average scale factor given by $\dot{a}/a = \bar{H}$ (we shall not have occasion to define a perturbed scale factor). This cannot be done exactly, because the expansion is not isotropic and the comoving hypersurfaces are not flat. However the departure from flatness is of first order, and can therefore be ignored when describing perturbations which are themselves of first order. In other words *all perturbations 'live' in flat space.*

Independent scales

Each perturbation f can be written as a Fourier series, defined in a comoving box much bigger than the observable universe

$$f(\mathbf{x}, t) = \sum_{\mathbf{k}} f_{\mathbf{k}}(t) e^{i\mathbf{k}.\mathbf{x}} \tag{79}$$

The beauty of this expansion is that each Fourier mode propagates independently, as long as the cosmological perturbation theory that we are developing here is valid. The inverse wavenumber a/k is said to define a scale, which is specified by giving its present value k^{-1}.

Now consider a small density enhancement in the early universe, which is destined to become, say, a galaxy. If its size is of order $r = xa$, it is typically made out of Fourier components with $k \sim x^{-1}$. As long as it corresponds to a small density contrast, $\delta\rho/\rho \ll 1$, it will expand with the universe so that its comoving size x remains constant. When its density contrast becomes of order 1 it will collapse and then its *physical* size will remain more or less constant. In both cases, though, the *mass* of the enhancement remains fixed. It is therefore useful to associate with each scale k the mass of matter enclosed within a sphere of comoving radius $x = k^{-1}$ (taking the universe to be practically homogeneous, corresponding to the early universe). This mass is

$$M(x) = 1.16 \times 10^{12} h^2 (x/1\,\mathrm{Mpc})^3 M_\odot \tag{80}$$

One expects perturbations with comoving wavenumber k to be relevant for the formation of structures with mass $M(x)$, where $x = k^{-1}$.

Horizon entry

The ratio of a given comoving scale a/k to the Hubble distance H^{-1} is equal to $aH/k = \dot{a}/k$, which decreases with time. At the epoch when this ratio falls through 1, the scale is said to *enter the horizon*.

Well after horizon entry, the scale is small compared with the Hubble distance, which means that ordinary physical effects like diffusion, free-streaming (collisionless particle movement) and the propagation of sound waves can operate, with the expansion of the universe playing only a minor role. Well before horizon entry, the scale is much bigger than the Hubble distance, which means

[8] As noted later this definition has to be modified if ω_{ij} does not vanish.

that causal processes of this kind cannot operate (at least during a Hubble time, and in practice not at all). Instead, as we shall see, each part of the universe evolves independently.

The scale entering the horizon at a given epoch is given by

$$k^{-1} = (aH)^{-1} = \frac{a_0 H_0}{aH} H_0^{-1} \tag{81}$$

Except around matter-radiation equality at $z \sim 10^4$ one has

$$aH \quad \propto a^{-1} \qquad \text{radiation domination} \tag{82}$$

$$aH \quad \propto a^{-1/2} \qquad \text{matter domination} \tag{83}$$

Thus a crude estimate is that the scale entering the horizon at $z \lesssim 10^4$ is $k^{-1} \sim z^{-1/2} H_0^{-1}$, making the scale entering the horizon at matter-radiation equality $k_{\rm eq}^{-1} \sim 10^{-2} H_0^{-1}$, and and that the scale entering the horizon at $z \gtrsim 10^4$ is $k^{-1}(z) \sim 10^2 z^{-1} H_0^{-1}$. An accurate calculation shows that $k_{\rm eq}^{-1} = 40 h^{-1}\,{\rm Mpc}$, and that the scale entering the horizon at photon decoupling is $k_{\rm dec}^{-1} = 90 h^{-1}\,{\rm Mpc}$. We shall see that the first scale is crucial for structure formation, and the second for the cmb anisotropy. The smallest scale directly relevant for structure formation is presumably the one corresponding to a dwarf galaxy, which has mass $M \sim 10^6 M_\odot$ and from Eq. (80) corresponds to $k^{-1} \sim .01\,{\rm Mpc}$, which enters the horizon when $T \sim 10\,{\rm keV}$.

The evolution of the density perturbation

Now I derive differential equations for the perturbations. In doing so one has to remember that the comoving worldlines are not in general geodesics, because of the pressure gradient. As a result, the proper time interval $d\tau$ between a pair of comoving hypersurfaces is position dependent. Its average may be identified with the coordinate time interval dt, and one can show (using essentially the Lorentz transformation between nearby observers) that its variation with position is given by [26, 2]

$$\frac{d\tau}{dt} = \left(1 - \frac{\delta p}{\rho + p}\right) \tag{84}$$

Along each worldline the rate of change of ρ with respect to proper time τ is given by energy conservation and has the same form Eq. (43) as in the unperturbed case,

$$\frac{{\rm d}\rho}{{\rm d}\tau} = -3H(\rho + p) \tag{85}$$

The rate of change of H is given by the Einstein field equation, and to first order receives just one extra term in the presence of perturbations, coming from the pressure gradient [26],

$$\frac{{\rm d}H}{{\rm d}\tau} = -H^2 - \frac{4\pi G}{3}(\rho + 3p) - \frac{1}{3}\frac{\nabla^2 \delta p}{\rho + p} \tag{86}$$

This equation is called the *Raychaudhuri equation*. The operator ∇^2 is the Laplacian on a comoving hypersurface, given in terms of comoving coordinates by

$$\nabla^2 = a^{-2}\delta^{ij}\frac{\partial}{\partial x^i}\frac{\partial}{\partial x^j} \tag{87}$$

Perturbing H, ρ and p to first order and using Eq. (84) gives the following equations for the Fourier components

$$(\delta\rho_{\mathbf{k}})^{\cdot} = -3(\rho + p)\delta H_{\mathbf{k}} - 3H\delta\rho_{\mathbf{k}} \tag{88}$$

$$(\delta H_{\mathbf{k}})^{\cdot} = -2H\delta H_{\mathbf{k}} - \frac{4\pi G}{3}\delta\rho_{\mathbf{k}} + \frac{1}{3}\left(\frac{k}{a}\right)^2 \frac{\delta p_{\mathbf{k}}}{\rho + p} \tag{89}$$

Eliminating $\delta H_{\mathbf{k}}$ with Eq. (88) gives a second order differential equation for $\rho_{\mathbf{k}}$. It is convenient to use the *density contrast* $\delta \equiv \frac{\delta\rho}{\rho}$, in terms of which the equation is

$$H^{-2}\ddot{\delta}_{\mathbf{k}} + [2 - 3(2w - c_s^2)]H^{-1}\dot{\delta}_{\mathbf{k}} - \frac{3}{2}(1 - 6c_s^2 + 8w - 3w^2)\delta_{\mathbf{k}} = -\left(\frac{k}{aH}\right)^2 \frac{\delta p_{\mathbf{k}}}{\rho} \tag{90}$$

I have used the notation $w = p/\rho$ and $c_s^2 = \dot{p}/\dot{\rho}$.

In these equations for the perturbations the distinction between ρ and $\bar{\rho}$ is not meaningful (similarly for p and H), because we are working only to first order. For the same reason the distinction between τ and t is not meaningful. In order to have linear equations one will therefore obviously take ρ to mean $\bar{\rho}$ and and similarly for p and H, and will take the dot to mean differentiation with respect to t. Note that $c_s^2 = \dot{p}/\dot{\rho}$ is the speed of sound, because p and ρ vary adiabatically in a homogeneous isotropic universe (no heat flow).

The case of zero pressure gradient

The right hand side of Eq. (90), which involves the pressure gradient, is negligible after matter domination because the pressure is negligible.[9] It is also negligible well before horizon entry even during radiation domination, because the gradient is so small.[10] When it is negligible, Eq. (90) can be reduced to a first order equation, which has a very simple interpretation [29, 24]

The solution and its interpretation hinge on the introduction of a quantity K, defined locally through the Friedmann equation Eq. (7). Just as in the homogeneous, isotropic case, general relativity shows that K/a^2 is a measure of the curvature of the comoving hypersurfaces. In the homogeneous, isotropic case K is time independent by virtue of the energy conservation Eq. (43) and the gravitational deceleration equation Eq. (44). We have seen that in general the latter is modified to become the Raychaudhuri equation Eq. (86), and as a result K is not in general time independent. But when the pressure gradient is negligible they are both unmodified, so that K is time independent. We can say that *when the pressure gradient is negligible, each region of space evolves like a separate Friedmann universe.*

On a comoving hypersurface K can be split into and average $\bar{K}$ plus a perturbation δK, but the average can be set equal to zero because $\Omega \simeq 1$. Perturbing the Friedmann equation therefore gives, to first order,

$$2H\delta H_{\mathbf{k}} = \frac{8\pi G}{3}\delta\rho_{\mathbf{k}} - \frac{\delta K_{\mathbf{k}}}{a^2} \tag{91}$$

When $\delta K_{\mathbf{k}}$ is time independent, Eqs. (88) and (91) give a *first* order differential equation for the density contrast,

$$\frac{2H^{-1}}{5+3w}\frac{d}{dt}\left[\left(\frac{aH}{k}\right)^2\delta_{\mathbf{k}}\right] + \left(\frac{aH}{k}\right)^2\delta_{\mathbf{k}} = \frac{2+2w}{5+3w}\mathcal{R}_{\mathbf{k}} \tag{92}$$

where $w = p/\rho$ and I have introduced the useful quantity

$$\mathcal{R}_{\mathbf{k}} = \frac{3}{2}\frac{\delta K_{\mathbf{k}}}{k^2} \tag{93}$$

Remembering that $\delta K/a^2$ is the curvature perturbation and that it has units of $(\text{length})^{-2}$, we see that $\mathcal{R}_{\mathbf{k}} = (\ni/\in)(\delta\mathcal{K}/\dashv^{\in})(\dashv^{\in}/\|^{\in})$ essentially measures the curvature perturbation in units of the relevant scale a/k. (The factor $3/2$ is chosen so as to give the simple relation Eq. (199) with the inflaton field perturbation.) As we shall verify shortly, another interpretation of $\mathcal{R}$ is that it is essentially the Newtonian gravitational potential caused by $\delta\rho$.

During any era when w is constant, Eq. (92) has the solution (dropping a decaying mode)

$$\left(\frac{aH}{k}\right)^2\delta_{\mathbf{k}} = \frac{2+2w}{5+3w}\mathcal{R}_{\mathbf{k}} \tag{94}$$

In the radiation dominated era before horizon entry this becomes

$$\left(\frac{aH}{k}\right)^2\delta_{\mathbf{k}} = \frac{4}{9}\mathcal{R}_{\mathbf{k}}(\text{initial}) \tag{95}$$

and in the matter dominated era it becomes

$$\left(\frac{aH}{k}\right)^2\delta_{\mathbf{k}} = \frac{2}{5}\mathcal{R}_{\mathbf{k}}(\text{final}) \tag{96}$$

[9] Except for the baryons on scales below the Jeans scale, and we are presently assuming that the dark matter is cold, otherwise it is modified by free-streaming. Both of these points will be addressed shortly.

[10] Provided that p/ρ is not extremely large, which is ensured by the adiabatic initial condition defined shortly.

As the labels imply I am regarding the value of $\mathcal{R}_{\mathbf{k}}$ during the first era as an 'initial condition', which determines its value during the 'final' matter dominated era.

For future reference note that during matter domination, $H \propto t^{-1} \propto a^{-3/2}$ and

$$\delta_{\mathbf{k}} \propto a \qquad \text{(matter domination)} \tag{97}$$

3.2 The transfer function

For scales entering the horizon well after matter domination ($k^{-1} \gg k_{\rm eq}^{-1} = 40h^{-1}\,\mathrm{Mpc}$) the initial and final eras overlap so that $\mathcal{R}_{\mathbf{k}}(\text{initial}) = \mathcal{R}_{\mathbf{k}}(\text{final})$. Otherwise there is a *transfer function* $T(k)$, which may be defined by

$$\mathcal{R}_{\mathbf{k}}(\text{final}) = T(k)\mathcal{R}_{\mathbf{k}}(\text{initial}) \tag{98}$$

An equivalent, and more usual, definition is

$$a^{-1}\delta_{\mathbf{k}}(\text{final}) = AT(k)\delta_{\mathbf{k}}(\text{initial}) \tag{99}$$

where the (time dependent) right hand side is evaluated at an arbitrarily chosen time during the initial era, and the constant A is chosen so that T becomes equal to 1 on large scales.

The adiabatic initial condition

In order to calculate the transfer function one needs an initial condition, specifying the density contrast of each species of matter and radiation before horizon entry and before matter domination. The matter consists of baryons and one or more species of non-baryonic dark matter. The radiation consists of photons and massless neutrinos.

The most natural condition, is the *adiabatic* condition, which is that the density of each species depends only on the total energy density, or equivalently on the temperature. In other words, each density contrast vanishes on a hypersurface of constant energy density. Going along any comoving worldline, the density ρ_r of any species of radiation varies like a^{-4} and the density ρ_m of any species of matter varies like a^{-3}, so that

$$\delta\rho_m/\rho_m = \frac{3}{4}\delta\rho_r/\rho_r \tag{100}$$

Thus, the adiabatic initial condition implies that each species of radiation has a common density contrast and so has each species of radiation, and that they are related by a factor 3/4.

The most general initial condition is a superposition of the adiabatic one and some *isocurvature* initial condition, specifying a set of initial density perturbations $\delta\rho_i$ which add up to zero. An isocurvature initial condition is not very natural, especially in the context of inflation. Furthermore, it turns out that a purely isocurvature intial condition (with a reasonably flat, Gaussian spectrum) cannot lead to a viable structure formation scenario. The adiabatic condition is assumed from now on.

Cold and hot dark matter

The other ingredient needed to calculate the transfer function is the nature of the dark matter. It is practically always taken to have negligible interaction with anything else (to keep it dark). Neutrino dark matter is called hot dark matter (HDM) because it remains relativistic (in other words it is radiation rather than matter) until a very later epoch, which is after structure forming scales have entered the horizon for pure HDM. Cold dark matter (CDM) by contrast is nonrelativistic at all epochs of interest, and warm dark matter is in between. Only CDM and HDM have as yet been studied in the light of modern data. Pure HDM is ruled out (at least if we do not form structure with seeds such as cosmic strings, or with a density perturbation having so much small scale power that very small objects form first). Pure CDM performs rather impressively, going a long way towards explaining several different types of data, but it cannot actually fit them all simultaneously within their accepted uncertainties. For this reason people have considered *mixed dark matter* (MDM), roughly 70% cold and 30% hot, which at the time of writing seems to do rather well [30, 33, 40].

Figure 2: The transfer function $T(k)$, taken from Pogosyan and Starobinsky [30]. From top down, the curves correspond to $\Omega_\nu = 0, 0.05, ...0.95$. The transfer function is evaluated at the present epoch, but in the limit of pure CDM (top curve) it becomes time independent.

The CDM transfer function

The transfer functions for pure CDM and for MDM are shown in Figure 2. Although several different effects have to be taken into account to calculate them accurately, their general form is easy to describe. Let us deal first with the pure CDM transfer function.

On a given scale essentially nothing happens before horizon entry because there is no time for particles to move on that scale. As a result, the growth $\delta \propto (aH)^{-2}$ of the density contrast, that we established earlier as a consequence of the 'separate Friedmann universe' evolution of each comoving region of space, is shared by each species. After horizon entry, the density contrast of massless neutrinos rapidly *decays away*, because the neutrinos *free-stream* (that is, travel freely) out of any over dense region. The baryons and photons are thermally coupled until the epoch of photon decoupling, and their density contrast *oscillates* as a sound wave, which decays more or less rapidly (depending on the scale) as the photons diffuse away from overdense regions. The density of cold dark matter *grows*, but until matter domination the amount of growth is rather small because the matter does not have much self-gravity. For example, taking the radiation to be uniform and ignoring the baryons, one finds by following the steps leading to Eq. (90) that the matter density contrast satisfies

$$\ddot{\delta}_m + 2H\dot{\delta}_m - 4\pi G\rho_m\delta_m = 0 \tag{101}$$

Using $\rho_m/\rho_r = a/a_{\rm eq}$, one finds that the solution of this equation is proportional to $2+3a/a_{\rm eq}$ (plus a decaying mode), which indeed hardly grows before matter domination at $a = a_{\rm eq}$. Since it would have grown like $(aH)^{-2}$ had horizon entry not occurred, the total amount of growth missed by a given scale is roughly

$$\frac{(aH)^2_{\rm hor}}{(aH)^2_{\rm eq}} = \left(\frac{k}{k_{\rm eq}}\right)^2 \tag{102}$$

The CDM transfer function is therefore roughly

$$T(k) \simeq 1 \qquad (k^{-1} > k_{\rm eq}^{-1} = 40h^{-1}\,{\rm Mpc}) \tag{103}$$

$$T(k) \simeq (k_{\rm eq}/k)^2 \qquad (k^{-1} < k_{\rm eq}^{-1}). \tag{104}$$

This indeed gives an extremely crude representation of the correct CDM transfer function. To calculate it accurately one needs to follow in detail the oscillation and/or decay of each component.[11]

The above discussion applies only to the CDM density perturbation. We need to know also what happens to the baryon density perturbation, since baryons after all are what we observe. Unlike the CDM density contrast, the baryon density contrast is small at the photon decoupling epoch because it has been decaying since horizon entry. Afterwards, the baryons are unaffected by the photons, but two competing forces act on them. First there is gravity, which tends to make them fall into the potential wells caused by the CDM density contrast, and second there is their own pressure gradient which tends to keep them out of the wells.

To rigorously see which effect wins one should generalize Eq. (90) to treat the CDM and the baryons as a pair of uncoupled fluids [26, 28]. In practice the following order of magnitude estimate is enough. Ignoring the pressure, the time taken for the baryons to fall into a well is[12] of order $(G\rho)^{-1/2}$. The time taken for the pressure to adjust itself to prevent the collapse is of order λ/c_s where $\lambda = 2\pi/k$ is the wavelength and c_s is the speed of sound. Collapse occurs if $\lambda/c_s \lesssim (G\rho)^{-1/2}$ because the pressure cannot act quickly enough. Inserting by convention a factor 2, one concludes that collapse occurs on scales in excess of $k_{\rm J} = (4\pi G\rho/v_s^2)^{1/2}$. This is called the *Jeans scale* and the corresponding mass, given by Eq. (80), is called the Jeans mass.

Just after decoupling the speed of sound is given by $c_s^2 = 5T/3m_N$ where $T \simeq 1100 \times 2.74\,{\rm K}$. The corresponding Jeans mass is of order $10^6 M_\odot$. As yet structure formation on such small scales is not well understood because the necessary numerical simulations of the collapse are too difficult, but it may be significant that there are no galaxies with mass less than the Jeans mass.

The MDM transfer function

The hot (neutrino) component becomes non-relativistic only at the epoch

$$1 + z_{\rm nr} = (1.7 \times 10^5) h^2 \Omega_\nu \tag{105}$$

On scales entering the horizon before this epoch, the neutrino dark matter will free-stream out of any density enhancement so that its density contrast will decay. Afterwards the HDM density contrast is free to grow to match that of the CDM, on scales in excess of an effective Jeans length

$$k_J^{-1} = .11(1+z)^{1/2}\Omega_\nu\,{\rm Mpc} \tag{106}$$

By the present time, the CDM and HDM have a common density contrast on scales $k^{-1} \gtrsim .1\,{\rm Mpc}$, which are the only ones of interest. Note, though, that the MDM continues to evolve right up to the present.

3.3 The peculiar velocity field

After matter domination it is very useful to introduce the concept of a peculiar velocity field. Traditionally this concept is formulated in the context of the Newtonian treatment of perturbations, but it has been recognised recently that the same concept is valid, and useful, also in the general relativistic treatment [2, 31].

At a given epoch the Newtonian treatment is valid on scales much less than the Hubble distance, to the extent that gravity waves can be ignored. There is a well defined fluid velocity $\mathbf{u}$, and choosing the reference frame so that $\mathbf{u}$ vanishes at the origin the PV field $\mathbf{v}$ is defined[13] by

$$\mathbf{u}(\mathbf{r}) - \mathbf{u}(0) = \bar{H}\mathbf{r} + \mathbf{v}(\mathbf{r}) - \mathbf{v}(0)\ , \tag{107}$$

An equivalent statement in terms of the velocity gradient Eq. (75) is

$$\delta u_{ij} = \partial_i v_j\ . \tag{108}$$

[11] A surprisingly good approximation [32] is to ignore the decay, treating the radiation and matter as uncoupled perfect fluids [26, 28], but I do not know of any simple reason why this is so.

[12] If a particle falls from rest towards a point mass M, its velocity at distance r is given by $mv^2 = 2GM/r$ so it falls a significant distance in a time $t \sim r/v \sim (GM/r^3)^{-1/2}$. We are replacing the point mass by a perturbation with size r and density $\rho \sim M/r^3$.

[13] Up to a constant, which can be chosen so that the average of $\mathbf{v}$ vanishes.

where $\partial_i = a^{-1}\partial/\partial x^i$. Like any vector field, $\mathbf{v}$ can be written $\mathbf{v} = \mathbf{v}^{\rm L} + \mathbf{v}^{\rm T}$, where the transverse part $\mathbf{v}^{\rm T}$ satisfies $\partial_i v_i^{\rm T} = 0$ and the longitudinal part is of the form $\mathbf{v}^{\rm L} = \boldsymbol{\nabla}\psi_v$. Eq. (14) defines the local Hubble parameter H, shear σ_{ij} and vorticity ω_{ij}, this last being given by

$$\omega_{ij} = \frac{1}{2}(\partial_i v_j^{\rm T} - \partial_j v_i^{\rm T}) \tag{109}$$

Angular momentum conservation gives $\omega_{ij} \to a^{-2}$, so $\mathbf{v}^{\rm T}$ decays like a^{-1} and may be dropped (remember that $\partial_i \equiv a^{-1}\partial/\partial x^i$).

Taking $\mathbf{v}$ to be purely longitudinal, it is determined by the density perturbation in the following way. First, take the trace of Eq. (14) to learn that $\boldsymbol{\nabla}.\mathbf{v} = 3\delta H$. From Eqs. (16), (91), (93) and (94), it follows that

$$\boldsymbol{\nabla}.\mathbf{v} = -(4\pi G\delta\rho)t \tag{110}$$

The solution of this equation is

$$\mathbf{v} = -t\boldsymbol{\nabla}\psi \tag{111}$$

or

$$v_i(\mathbf{x}, t) = -(t/a)\frac{\partial\psi(\mathbf{x})}{\partial x^i} \tag{112}$$

where

$$\psi(\mathbf{x}) = -Ga^{-2}\int \frac{\delta\rho(\mathbf{x}', t)}{|\mathbf{x}' - \mathbf{x}|}d^3x' \tag{113}$$

The factor a^{-2} converts coordinate distances into physical distances. Since it is related to the density perturbation by the Newtonian expression, ψ is called the peculiar gravitational potential. It is independent of t because, from Eq. (97), $\delta\rho \propto a^2$.

From Eq. (96) we see that the peculiar gravitational potential is related to the spatial curvature perturbation by

$$\psi = -\frac{3}{5}\mathcal{R}(\text{final}) \tag{114}$$

From Eqs. (111) and (113) the Fourier components of $\mathbf{v}$, ψ and δ are related by

$$\mathbf{v}_\mathbf{k} = i\frac{\mathbf{k}}{k}\left(\frac{aH}{k}\right)\delta_\mathbf{k} \tag{115}$$

$$\psi_\mathbf{k} = -\frac{3}{2}\left(\frac{aH}{k}\right)^2\delta_\mathbf{k} \tag{116}$$

The extension of these results to the relativistic regime turns out to be very simple [2, 31]. The vorticity decays $(\rho + p)a^{-5}$ and is again presumably negligible. In that case one can show that the velocity gradient perturbation is of the form

$$\delta u_{ij} = \partial_i v_j + \frac{1}{2}\dot{h}_{ij} \tag{117}$$

The extra term, which is absent in Newtonian physics, represents the effect of gravitational waves. The success of Newtonian theory on scales $\lesssim 100\,$Mpc implies that it is negligible there, but as we shall see it could be significant on larger scales, and contribute to the cmb anisotropy. Even if present though, it does not spoil Eqs. (110)–(116) because it is transverse, $\partial_i h_{ij} = 0$, and traceless, $\delta^{ij}h_{ij} = 0$.

One can avoid dropping the vorticity by proceeding as follows [31]. First, a transverse velocity $\mathbf{v}^T$ may be *defined* by Eq. (109). The worldlines with velocity $-\mathbf{v}^T$ relative to the comoving worldlines have no vorticity, and comoving hypersurfaces are defined to be orthogonal to them (there are no hypersurfaces orthogonal to worldlines with nonzero vorticity). The velocity gradient δu_{ij} receives an extra contribution $\frac{1}{2}(\partial_i v_j^{\rm T} - \partial_j v_i^{\rm T}) + \frac{1}{2}(\partial_i w_j^{\rm T} + \partial_j w_i^{\rm T})$ where $\mathbf{w}^{\rm T} = \left[1 + 6\left(1 + \frac{p}{\rho}\right)\left(\frac{aH}{k}\right)^2\right]\mathbf{v}^{\rm T}$ For scales well inside the horizon $\mathbf{w}$ is negligible and the Newtonian result is recovered.

3.4 The spectrum of the density perturbation

In order to discuss the perturbations in a given region of the universe around us, one has to perform the Fourier expansion Eq. (79) in a box much bigger than this region. If the box is a cube with sides of length L, the possible values of $\mathbf{k}$ form a cubic lattice in k space with spacing $2\pi/L$.

When discussing an isolated system, which is the usual case in physics, one can take the limit $L \to \infty$ in a straightforward way, the coefficients $f_{\mathbf{k}}$ tending to a constant limit which is a smooth function of $\mathbf{k}$. But cosmological perturbations do not fall off at large distances, and their Fourier coefficients are not smooth functions of $\mathbf{k}$. They are the spatial analogue of a signal extending over an indefinite period of time, as opposed to an isolated pulse.

Although the coefficients $f_{\mathbf{k}}$ are not smooth functions, it is reasonable to suppose that $|f_{\mathbf{k}}|^2$ is smoothly varying when smeared over a region d^3k of k space, which is large enough to contain many lattice points. I shall denote this average by $\langle |f_{\mathbf{k}}|^2 \rangle$. It depends only on $k = |\mathbf{k}|$, and up to a k dependent factor it is called the *spectrum* of f, because of the analogy with a signal. A convenient choice of the factor is to define the spectrum as

$$\mathcal{P}_f \equiv \left(\frac{Lk}{2\pi}\right)^3 4\pi \langle |f_{\mathbf{k}}|^2 \rangle \tag{118}$$

The normalisation is chosen to give a simple formula for the dispersion (root mean square) of f, which I shall denote by σ_f. From the Fourier expansion one has $\sigma_f^2 = \sum |f_{\mathbf{k}}^2|$, and since the possible values of $\mathbf{k}$ form a cubic lattice with spacing $2\pi/L$ the transition from sum to integral is

$$\left(\frac{2\pi}{L}\right)^3 \sum_{\mathbf{k}} \longrightarrow 4\pi \int k^2 dk \tag{119}$$

The dispersion σ_f is therefore given by

$$\sigma_f^2 \equiv \langle f^2(\mathbf{x}) \rangle = \int_0^\infty \mathcal{P}_f(k) \frac{dk}{k} \tag{120}$$

with the brackets now denoting the average over position $\mathbf{x}$.

For the density perturbation $f = \delta$ it is useful to define the correlation function $\xi(r)$ by

$$\xi(r) = \langle f(\mathbf{r} + \mathbf{x}) f(\mathbf{r}) \rangle = \int_0^\infty \mathcal{P}_f(k) \frac{\sin(kr)}{kr} \frac{dk}{k} \tag{121}$$

The analogous quantity is useful for other perturbations like the peculiar velocity components, though it is not then called the correlation function. For $r = 0$ it clearly reduces to σ_f^2.

If the phases of the Fourier coefficients are random, f is said to be Gaussian, and then all of its stochastic properties are determined by its spectrum. In particular the probability distribution of f, evaluated at randomly chosen points, has a Gaussian profile. As long as the perturbations are evolving linearly this Gaussian property is true of perturbations originating as inflationary fluctuations, and I take it for granted from now on.

From Eqs. (96) and (98), the spectrum of the density contrast after matter domination may be written

$$\mathcal{P}_\delta(k) = \left(\frac{k}{aH}\right)^4 T^2(k) \delta_H^2(k) \tag{122}$$

The quantity δ_H specifies the initial spectrum. In fact, from Eq. (95) it is related to the spectrum of the initial curvature perturbation $\mathcal{R}$ by

$$\delta_H^2(k) = \frac{4}{25} \mathcal{P}_\mathcal{R}(k) \tag{123}$$

The subscript H has been used because δ_H^2 is exactly equal to the value of $\mathcal{P}_\delta$ on horizon entry on scales $k^{-1} \gg k_{\mathrm{eq}}^{-1}$, and approximately equal to it on smaller scales. As we shall see later, this means that $\delta_H(k)$ is *roughly* equal to the *rms* value of δ at horizon entry, for a density enhancement with comoving size of order k^{-1}.

The standard assumption is that δ_H^2 is independent of k. A more general possibility is to consider a spectrum

$$\delta_H^2 \propto k^{n-1} \tag{124}$$

where the exponent n is called the *spectral index*. (The definition of the index as $n-1$ instead of n is a historical accident.) The standard choice of $n=1$ was first advocated by Harrison (1970) and Zel'dovich (1970) on the ground that it is the only one making the perturbation small on all scales, at the epoch of horizon entry. Although this is a powerful argument in favour of a value of n *of order* 1, it does not require that n is actually close to 1 unless one assumes, without justification, that the power law dependence holds over a huge number of decades of scale[14]

In this context, inflation makes a striking prediction [33]. It gives the spectral index is given in terms of two parameters ϵ_1 and η_1,

$$n = 1 + 2\eta_1 - 6\epsilon_1 \tag{125}$$

The parameters depend on the model of inflation, but both are much less than 1 in magnitude and ϵ_1 is positive. As a result most inflationary models predict that n is rather close to 1, and typically smaller rather than bigger. The prediction that n is close to 1 seems to be confirmed by observation, which is a striking success for the general idea of inflation. In the very near future, one will be able to use the observed value of n as a powerful discriminator between different inflationary models.

4 The Cosmic Microwave Background Anisotropy

The first detection last year of the intrinsic anisotropy of the cmb was surely the most important advance in observational cosmology for a long time. Exploring scales two orders of magnitude bigger than the biggest that can be probed by galaxy surveys, the fact that it agreed within a factor of two with extrapolation from these surveys using the pure CDM model with a flat spectrum was extremely impressive. Before exploring the significance of this fact in the next section, let us study the cmb anisotropy is some detail.

4.1 The spectrum of the cmb anisotropy

At a given wavelength, the cmb is anisotropic *ie* its intensity depends on the direction of observation. The anisotropy is usually specified by giving the equivalent variation in the temperature of the blackbody distribution. Denoting the direction of observation by a unit vector $\mathbf{e}$, the anisotropy may be expanded into multipoles

$$\frac{\Delta T(\mathbf{e})}{T} = \mathbf{w}.\mathbf{e} + \sum_{l=2}^{\infty} \sum_{m=-l}^{+l} a_l^m Y_l^m(\mathbf{e}) \tag{126}$$

The dipole term $\mathbf{w}.\mathbf{e}$ is well measured, and is the Doppler shift caused by our velocity $\mathbf{w}$ relative to the rest frame of the cmb (defined as the frame where it has zero momentum density). Unless otherwise stated, ΔT will denote only the non-dipole part from now on. Its mean square over the whole sky is

$$\left\langle \left(\frac{\Delta T}{T} \right)^2 \right\rangle = \frac{1}{4\pi} \sum_{l=2}^{\infty} \sum_{m=-l}^{l} |a_m^l|^2 \tag{127}$$

The multipoles, and therefore the mean square anisotropy, depend on the position of the observer. Averaging over position gives the result for a randomly placed observer,

$$\left\langle \left\langle \left(\frac{\Delta T}{T} \right)^2 \right\rangle \right\rangle_{\text{position}} = \frac{1}{4\pi} \sum_{l=2}^{\infty} (2l+1)\Sigma_l^2 \tag{128}$$

where

$$\Sigma_l^2 = \left\langle |a_l^m|^2 \right\rangle_{\text{position}} \tag{129}$$

In contrast with $|a_l^m|^2$, the quantities Σ_l^2 are expected to be smooth functions of l. For large l one can therefore replace the sum by an integral,

$$\left\langle \left\langle \left(\frac{\Delta T}{T} \right)^2 \right\rangle \right\rangle_{\text{position}} \simeq \frac{1}{4\pi} \int_2^{\infty} l(2l+1)\Sigma_l^2 \frac{dl}{l} \tag{130}$$

[14] On cosmologically interesting scales the observed magnitude of δ_H is of order 10^{-5}, so if $n = 1 - \delta n \simeq 1$ one has to extrapolate over $5/\delta n$ orders of magnitude before δ_H becomes of order 1.

Figure 3: This figure is reproduced from [34]. The top Figure shows the spectrum of the cmb anisotropy for a slightly tilted standard CDM model with spectral index $n = 0.85$ and $\Omega_B = 0.05$. As discussed in the text, the case $n = 1$ is recovered for $l \gg 1$ by multiplying the curves by a factor $l^{(1-n)/2} = l^{.075}$. The contribution of an adiabatic density perturbation is the middle line, labelled 'S', and the contribution of gravitational waves is the bottom line, labelled 'T'. The light dashed line is the density contribution with $\Omega_B = 0.01$. For each curve, the quantity plotted is $l(l+1)\Sigma_l^2$, normalised to 1 for the quadrupole, $l = 2$. If the gravitational wave contribution to the quadrupole is equal to that of the density perturbation, as is roughly the case for power-law inflation with $n = 0.85$, the top curve indicates the total. On the other hand, it could well be that the gravitational contribution is negligible. The bottom Figure shows the filters F_l for various experiments, which are cited in [34].

Another nice feature of large l is the relation $\theta \sim 1/l$ between the angular size of a feature in the sky (in radians) and the order l of the multipoles that dominate it. (This is analogous to the relation $r \sim 1/k$ between the linear size of a feature and the wavenumbers that dominate its Fourier expansion.) Translating to degrees we have the following relation between l and the angular scale

$$\frac{\theta}{1^0} \simeq \frac{60}{l} \tag{131}$$

By analogy with the Eq. (120) for the density contrast, one may call $(2l+1)l\Sigma_l^2/4\pi$ the *spectrum* of the cmb anisotropy.

4.2 Predicting the cmb anisotropy

The predicted spectrum of the cmb anisotropy is shown in Figure 3. I will discuss the prediction qualitatively, then give a detailed calculation of the large scale (small l) part which is the most interesting.

Discounting the possibility of early re-ionisation the cmb last scattered at the epoch of photon decoupling. Its surface of last scattering therefore lies practically at the particle horizon whose comoving distance is $x = 2H_0^{-1} = 6000h^{-1}\,\mathrm{Mpc}$. At this surface an angle θ degrees subtends a comoving distance

$$\frac{x}{100h^{-1}\,\mathrm{Mpc}} \simeq \frac{\theta}{1^0} \tag{132}$$

This gives, at least roughly, the linear scale explored by observing the cmb on a given angular scale.

It is convenient (at least pedagogically) to separate the anisotropy into an initial anisotropy, specified on a comoving hypersurface shortly after last scattering, and the additional anisotropy acquired on the journey towards us,

$$\frac{\Delta T(\mathbf{e})}{T} = \left(\frac{\Delta T(\mathbf{e})}{T}\right)_{\rm em} + \left(\frac{\Delta T(\mathbf{e})}{T}\right)_{\rm jour} \tag{133}$$

The initial anisotropy represents the variation of the intensity of the cmb, as measured by different comoving observers, each with their detector pointing in the relevant direction $\mathbf{e}$.

Let us discuss these contributions, from the largest scales downwards. Scales in excess of a degree or so explore linear scales in excess of 100 Mpc, neatly complementing galaxy surveys which explore scales $\lesssim 100\,$Mpc. Such scales have not entered the horizon at last scattering, so the adiabatic initial condition Eq. (100) still holds at that epoch. The cmb is practically isotropic for each comoving observer, but its intensity varies from place to place because $\rho_r \propto T^4$. Since last scattering occurs during matter domination, the density contrast is practically that of the matter, so the adiabatic condition gives

$$\left(\frac{\Delta T(\mathbf{e})}{T}\right)_{\rm em} \simeq \frac{4}{3}\delta(\mathbf{x}_{\rm em}, t_{\rm em}) \tag{134}$$

However we shall see that this effect is dominated by the additional anisotropy acquired on the journey towards us which is called (on these large angular scales) the Sachs-Wolfe effect.

Scales in the *arcminute* regime explore distance scales of order 10 to 100 Mpc, the same as the one explored by large scale galaxy surveys. As the scale decreases through this regime, the anisotropy at first exhibits a peak, which can be interpreted as of the Doppler shift caused by the peculiar velocity of the initial comoving hypersurface. Thereafter the anisotropy acquired on the journey towards us becomes negligible.[15] The initial anisotropy also falls, but it remains significant until the scale falls to a few arcseconds.

On scales less than a few arcminutes, the distance Eq. (132) subtended at the last scattering surface is less than the thickness of this surface. As a result the cmb is isotropised by its passage through the last scattering surface, and it turns out that no significant anisotropy is acquired by the cmb on its journey towards us either. The anisotropy is therefore expected to be very small.

The cmb transfer function and the cosmic variance

By evaluating these physical effects, one can calculate the multipoles a_l^m seen by an observer at any given position, in terms of the initial density perturbation $\delta_{\mathbf{k}}$(initial).[16] Of course we do not know $\delta_{\mathbf{k}}$(initial), nor are we interested in it. Instead we know, or rather are prepared to hypothesise, its stochastic properties, which are that it is Gaussian with a spectrum specified by $\delta_H^2(k) \propto k^{n-1}$. The Gaussianity implies that the real and imaginary part of each multipole a_l^m has a Gaussian probability distribution as a function of the observer's position, with no correlation between them (except for the trivial one $a_l^{m*} = a_l^{-m}$). The spectrum Σ_l^2 of the cmb anisotropy—specifying the widths of the Gaussian distributions—is related to the spectrum $\delta_H(k)$ of the initial density perturbation by a transfer function $T_l(k)$,

$$\Sigma_l^2 = \int_0^\infty T_l^2(k)\delta_H^2(k)\frac{{\rm d}k}{k} \tag{135}$$

In accordance with Eq. (132), $T_l(k)$ is peaked at

$$\frac{l}{60} \simeq \frac{100h^{-1}\,{\rm Mpc}}{k^{-1}} \tag{136}$$

The transfer function is independent of the nature of the dark matter on large scales as we discuss in a moment. On smaller scales the effect of adding in some hot dark matter has not yet been investigated in detail, but going from pure CDM to MDM (30% hot dark matter) is not expected to have a big effect. The prediction in Figure 3 is for pure CDM.

This prediction gives the expectation value Σ_l^2 of the multipole $|a_l^m|^2$ evaluated at a random position, but we can only measure $|a_l^m|^2$ at our position. The best guess is that $|a_l^m|^2 = \Sigma_l^2$, but one

[15] I here discount the Sunyaev-Zel'dovich effect by which the cmb acquires anisotropy through photon up-scattering off dust on its way through a galaxy or cluster. It occurs on scales of order one arcminute and is easily identified for nearby clusters. The cumulative effect of distant clusters is expected to be negligible.

[16] Discounting gravitational waves, which I consider in a moment, and isocurvature density perturbations, cosmic strings etc which I do not consider at all.

can also calculate the expected error, which is the square root of the variance of the guess (the mean value of $(|a_l^m|^2 - \Sigma_l^2)^2$). This variance is called the *cosmic variance*. For the real or imaginary part of a single multipole, the Gaussian distribution implies that the variance is equal to 2 times the mean square prediction. The lth multipole has $2l+1$ independent components, so summing over m makes the cosmic variance only $2/(2l+1)$ times the mean square prediction. A typical observation typically involves a (weighted) sum over several multipoles, which further reduces the cosmic variance so that in practice it is not usually a dominant factor. One must always bear it in mind though.

4.3 The Sachs-Wolfe effect

Now we calculate the anisotropy on scales in excess of a few degrees [2, 31].

Consider a photon passing a succession of comoving observers. Its trajectory is $a\mathrm{d}\mathbf{x}/\mathrm{d}t = -\mathbf{e}$ and between nearby observers its Doppler shift is

$$-\frac{d\lambda}{\lambda} = e_i e_j u_{ij} dr = -\frac{d\bar{a}}{\bar{a}} + e_i e_j \delta u_{ij} dr \,, \tag{137}$$

where the first term is due to the average expansion, and the second is due to the relative PV of the observers. Integrating this expression gives the redshift of radiation received by us, which was emitted from a distant comoving source. The unperturbed result is $\lambda/\lambda_{\rm em} = 1/a_{\rm em}$, and the first order perturbation gives

$$\frac{\Delta T(\mathbf{e})}{T} = \left(\frac{\Delta T(\mathbf{e})}{T}\right)_{\rm em} + \int_0^{x_{\rm em}} e_i e_j \delta u_{ij}(\mathbf{x}, t) a(t) dx \,, \tag{138}$$

where $x_{\rm em} \simeq 2H_0^{-1}$ is the coordinate distance of the last scattering surface, and the integration is along the photon trajectory

$$x(t) = \int_t^{t_0} \frac{dt}{a} = 3\left(\frac{t_0}{a_0} - \frac{t}{a}\right) \,. \tag{139}$$

Using Eqs. (108), (139) and (111) and integrating by parts one finds

$$\begin{aligned} \frac{\Delta T(\mathbf{e})}{T} &= \left(\frac{\Delta T(\mathbf{e})}{T}\right)_{\rm em} + \frac{1}{3}[\psi(\mathbf{x}_{\rm em}) - \psi(0)] + \\ &\quad \mathbf{e}\cdot[\mathbf{v}(0,t_0) - \mathbf{v}(\mathbf{x}_{\rm em}, t_{\rm em})] \,. \end{aligned} \tag{140}$$

Here $\mathbf{x}_{\rm em} = x_{\rm em}\mathbf{e}$ is the point of origin of the cmb coming from direction $\mathbf{e}$.

A better expression follows [31] if one uses the divergence theorem and Eq. (139) to project out the dipole part of $\psi(\mathbf{x}_{\rm em})/3$. One finds that it is equal to $\langle \mathbf{v}(t_{\rm em}) - \mathbf{v}(t_0)\rangle$ where $\langle\rangle$ denotes the average within the last scattering surface $x = x_{\rm em}$. Defining $\tilde{\mathbf{v}} = \mathbf{v} - \langle \mathbf{v}\rangle$ this gives

$$\begin{aligned} \frac{\Delta T(\mathbf{e})}{T} &= \left(\frac{\Delta T(\mathbf{e})}{T}\right)_{\rm em} + \frac{1}{3}\left[\psi(\mathbf{x}_{\rm em})\right]_{l>1} + \\ &\quad \mathbf{e}\cdot[\tilde{\mathbf{v}}(0,t_0) - \tilde{\mathbf{v}}(\mathbf{x}_{\rm em}, t_{\rm em})] \,. \end{aligned} \tag{141}$$

On angular scales $\gg 1^0$, corresponding to linear scales at last scattering which are outside the horizon, Eqs. (115) and (116) show that

$$|\psi_{\mathbf{k}}| \gg |\delta_{\mathbf{k}}| \gg |\mathbf{v}_{\mathbf{k}}| \tag{142}$$

On these scales we can therefore drop the term $\tilde{\mathbf{v}}(\mathbf{x}_{\rm em}, t_{\rm em})$, as well as the the initial anisotropy Eq. (134). This leaves

$$\frac{\Delta T(\mathbf{e})}{T} = \frac{1}{3}\left[\psi(\mathbf{x}_{\rm em})\right]_{l>1} + \mathbf{e}\cdot\tilde{\mathbf{v}}_0 \,. \tag{143}$$

The last term in this expression is the dipole and it defines the rest frame of the cmb by giving our velocity $\tilde{\mathbf{v}}_0 = \mathbf{v}_0 - \langle \mathbf{v}_0\rangle$ relative to that frame. Since $\mathbf{v}_0$ is our peculiar velocity, the peculiar velocity of the cmb rest frame is $\langle \mathbf{v}_0\rangle$, the average peculiar velocity (bulk flow) of the region inside the surface of last scattering (at roughly the Hubble distance from us). As we shall discuss in Section 5 the bulk flow decreases as the scale increases, and is quite negligible on the Hubble scale, so in practice we can just say that the rest frame of the cmb has zero peculiar velocity. This indeed is what is usually assumed.

Inserting the Fourier expansion of ψ and projecting out a multipole leads, for an observer at the origin of coordinates, to

$$a_l^m = -2\pi i^l \sum_{\mathbf{k}} \left(\frac{aH}{k}\right)^2 j_l(kx_{\rm em}) Y_l^m(\Omega_{\mathbf{k}}) \delta_{\mathbf{k}} \tag{144}$$

where $\Omega_{\mathbf{k}}$ is the solid angle in $\mathbf{k}$ space. For a randomly placed observer the coefficients $\delta_{\mathbf{k}}$ have random phases (except for the trivial correlation implied by the reality condition $\delta_{\mathbf{k}}^* = \delta_{-\mathbf{k}}$), and this implies that the multipoles a_l^m have a Gaussian distribution. The variance of the distribution is (Peebles 1982b)

$$\Sigma_l^2 = \pi \int_0^\infty \frac{{\rm d}k}{k}\, j_l^2\,(2k/aH)\;\, \delta_H^2(k) \tag{145}$$

where j_l is the spherical Bessel function. With $\delta_H^2(k) \propto k^{n-1}$ this becomes

$$\Sigma_l^2 = \frac{\pi}{2}\left[\frac{\sqrt{\pi}}{2} l(l+1) \frac{\Gamma((3-n)/2)}{\Gamma((4-n)/2)} \frac{\Gamma(l+(n-1)/2)}{\Gamma(l+(5-n)/2)}\right] \frac{\delta_H^2(H_0/2)}{l(l+1)} \tag{146}$$

The square bracket is equal to 1 for $n = 1$. For $l \gg 1$ and $l \gg |n|$ it can be replaced by 1, if $\delta_H(k)$ is evaluated on the scale Eq. (136) which dominates the integral.

4.4 The contribution of gravitational waves

We have so far ignored gravitational waves. They correspond to a metric perturbation h_{ij} which is traceless, $\delta^{ij}h_{ij} = 0$, and transverse, $\partial_i h_{ij} = 0$. This means that each Fourier component is of the form $h_{ij} = h_+ e_{ij}^+ + h_\times e_{ij}^\times$, where in a coordinate system where $\mathbf{k}$ points along the z-axis the nonzero components of the polarisation tensors are defined by $e_{xx}^+ = -e_{yy}^+ = 1$ and $e_{xy}^\times = e_{yx}^\times = 1$. The spectrum $\mathcal{P}_g$ of the gravitational wave amplitude may be defined by summing Eq. (23) over all four components,

$$\mathcal{P}_g = 2\left(\frac{Lk}{2\pi}\right)^3 \left(\langle |h_+(\mathbf{k})|^2\rangle + \langle |h_\times(\mathbf{k})|^2\rangle\right) \tag{147}$$

Each Fourier component satisfies the massless wave equation, which in comoving coordinates is

$$\ddot{h}_{ij} + 3H\dot{h}_{ij} + (k/a)^2 h_{ij} = 0 \tag{148}$$

Well before horizon entry it has constant initial value. For scales entering the horizon after matter domination its subsequent evolution is

$$h_{ij}(t) = \left[3\sqrt{\frac{\pi}{2}}\frac{J_{3/2}(x)}{x^{3/2}}\right] h_{ij}(\text{initial}) \tag{149}$$

where $x = 2k/(aH)$. Well after horizon entry one has redshifting radiation.

As noted in Eq. (117), the contribution of gravitational waves to the velocity gradient is $\delta u_{ij} = \dot{h}_{ij}/2$. By substituting this expression into Eq. (138) one can calculate the cmb multipoles in terms of the initial amplitude, and hence calculate cmb spectrum Σ_l^2 in terms of the initial gravitational wave spectrum $\mathcal{P}_g(k)$. Each gravitational wave gives its dominant contribution as it enters the horizon, since its amplitude is practically constant before that and redshifts away afterwards. As a result the gravitational wave contribution to the cmb anisotropy cuts off below the angular scale $\simeq 1^0$, corresponding to the smallest linear scale which enters the horizon after decoupling $k_{\rm dec}^{-1} = 90h^{-1}\,$Mpc. The corresponding multipole cutoff is at $l \simeq k_{\rm dec}^{-1} H_0/2 \simeq 70$.

On scales well in excess of a degree ($l \lesssim 10$), the gravitational wave contribution is given in the case of a flat spectrum ($\mathcal{P}_g(k)$ independent of k) by Starobinsky [35],

$$l(l+1)\Sigma_l^2 = \frac{\pi}{36}\left(1 + \frac{48\pi^2}{385}\right)\mathcal{P}_g C_l \tag{150}$$

If one ignored the cutoff due to the redshift, the coefficient C_l would become 1 in the limit $l \gg 1$. Starobinsky gives the values $C_2 = 1.118$, $C_3 = 0.878$ and $C_4 = 0.819$. Numerical calculation, including the effect of the cutoff, shows that a value C_l close to 1 is indeed achieved for $l \sim 10$, before the cutoff takes effect (Figure 3).

For $l \gg 1$ the above result is good also if $\mathcal{P}_g(k)$ has moderate scale-dependence, provided that it is evaluated at the scale Eq. (136) which dominates the lth multipole.

Within a given inflationary model one can calculate $\mathcal{P}_g(k)$. Defining the spectrum n_g of the gravitational waves by $\mathcal{P}_g \propto k^{n_g}$, one finds [33]

$$n_g = -2\epsilon_1 \tag{151}$$

where ϵ_1 is the same small positive parameter appearing in Eq. (125) for n. Setting the coefficient C_l in Eq. (150) equal to 1, the ratio of the gravitational and density contributions is also given in terms of this parameter, by [33]

$$R \equiv \frac{\Sigma_l^2(\text{grav})}{\Sigma_l^2(\text{density})} \simeq 12\epsilon_1 \tag{152}$$

In some models of inflation, ϵ_1 is very small, corresponding to a very small gravitational wave contribution with a very flat spectrum. If ϵ_1 is significant, the potentially is typically of the form ϕ^α with α at least equal to 2 and often much bigger (I include the behaviour $e^{A\phi}$ in this class since it corresponds to the limit $\alpha \to \infty$). In that case one has $-n_g = 1 - n > 0$ and

$$R \simeq 6(1-n) \tag{153}$$

Thus one's expectation from inflation is that *either* gravitational waves are negligible *or* that their relative magnitude is related to the spectral index by Eq. (153) (on scales in excess of a few degrees). It should be noted that this expectation comes from particle physics, being a consequence of the kind of potentials that typically arise. It is *not* generic to the idea of inflation *per se*, which in fact provides no relation whatever between R and n since the two parameters ϵ_1 and η_1 can be chosen independently.

4.5 Observing the cmb anisotropy

By analogy with Eq. (121), one can define a temperature correlation function,

$$C(\alpha) = \left\langle \frac{\Delta T(\mathbf{e})}{T} \frac{\Delta T(\mathbf{e}')}{T} \right\rangle \tag{154}$$

Here $\mathbf{e}$ and $\mathbf{e}'$ specify the directions in which the anisotropy is observed, and the average goes over directions separated by an angle α. It is given in terms of the multipoles by

$$C(\alpha) = \sum_{l=2}^{\infty} Q_l^2 P_l(\cos\alpha) \tag{155}$$

where

$$Q_l^2 \equiv \frac{1}{4\pi} \sum_{m=-l}^{+l} |a_l^m|^2 \tag{156}$$

Since one is not interested in very small scale fluctuations, it is convenient to smear $\Delta T(\mathbf{e})/T$ over a patch of sky with angular size θ_f, to obtain a 'filtered' anisotropy $\Delta T(\theta_f, \mathbf{e})/T$. This cuts off the multipole expansion above $l \sim \theta_f^{-1}$; to be precise, with a Gaussian profile for the smearing function, its effect is

$$a_l^m \to \exp\left[-(2l+1)\theta_f/2\right] a_l^m \tag{157}$$

Associated with the smeared quantity is a correlation function

$$C(\theta_f, \alpha) = \sum_{l=2}^{\infty} \exp\left[-((2l+1)\theta_f/2)^2\right] Q_l^2 P_l(\cos\alpha) \tag{158}$$

A given experimental setup typically measures something which can be directly related to a suitably smeared C. The simplest one is a single beam whose resolution can be represented by a Gaussian profile. Averaged over the sky the anisotropy measured by such a beam is

$$\sigma_T^2(\theta_f) \equiv \left\langle \left[\frac{\Delta T(\theta_f, \mathbf{e})}{T}\right]^2 \right\rangle = C(\theta_f, 0) \tag{159}$$

For more complicated setups involving two or three beam switching, still with Gaussian profiles, the measured anisotropy is

$$\left\langle \left[\frac{\Delta T(\theta_f, \mathbf{e})}{T} - \frac{\Delta T(\theta_f, \mathbf{e}')}{T} \right]^2 \right\rangle = 2\left[C(\theta_f, 0) - C(\theta_f, \alpha) \right] \tag{160}$$

and

$$\begin{aligned} & \left\langle \left[\frac{\Delta T(\theta_f, \mathbf{e})}{T} - \frac{1}{2}\frac{\Delta T(\theta_f, \mathbf{e}')}{T} - \frac{1}{2}\frac{\Delta T(\theta_f, \mathbf{e}'')}{T} \right]^2 \right\rangle \\ = \; & \frac{3}{2} C(\theta_f, 0) - 2C(\theta_f, \alpha) + \frac{1}{2} C(\theta_f, 2\alpha) \end{aligned} \tag{161}$$

In the second expression, $\mathbf{e}'$ and $\mathbf{e}''$ lie on opposite sides of $\mathbf{e}$, aligned on a great circle and each at an angular distance α. In a typical setup the 'beam throw' α is of the same order of magnitude as the 'antenna resolution' θ_f.

If we denote the left hand side of Eq. (159), (160) or (161) generically by $(\Delta T/T)^2$, the prediction in terms of multipoles may be written

$$\left(\frac{\Delta T}{T} \right)^2 = \sum_{l=2}^{\infty} F_l Q_l^2 \tag{162}$$

where F_l is a filter function. For the single beam expression Eq. (159) the filter keeps all multipoles with $l \lesssim \theta_f^{-1}$, but for the two- and three beam expressions it keeps only multipoles with $l \sim \theta_f^{-1}$.

For a realistic experimental setup these Gaussian expressions require modification. Accurate filter functions for some currently mounted observations are shown in Figure 3, which is reproduced from [34].

Present observations

Until recently the only published observation of the non-dipole cmb anisotropy was that of the COBE group [36]. They give the quadrupole and some information about higher multipoles, but because of cosmic variance their most useful statistic is the rms anisotropy on the scale 10^0, the smallest that they explore.

$$\sigma_T(\theta_f) = (1.10 \pm .18) \times 10^{-5} \tag{163}$$

Assuming a Gaussian beam profile one can use this expression to normalize the spectrum $\delta_H(k)$ of the density perturbation, given the spectral index n and the relative contribution R of gravitational waves. With $n = 1$ and $R = 0$ it gives $\delta_H = (1.7 \pm .3) \times 10^{-5}$. Subsequent analysis [37] using the correct beam profile and allowing for incomplete sky coverage raises this to by about one standard deviation to

$$\delta_H = (2.0 \pm .3) \times 10^{-5} \tag{164}$$

Since the ten degree measurement probes scales of order the Hubble distance $k^{-1} \sim 10^4\,\mathrm{Mpc}$, this provides a reasonable estimate of δ_H on that scale even if $n \neq 1$ is somewhat different from 1, *provided* that gravitational waves are negligible. Most inflationary models predict either that they will be negligible, or that they will be given by Eq. (153) (with $n < 1$). In the latter case the normalisation Eq. (164) becomes

$$\delta_H = (2.0 \pm .3) \times 10^{-5} [1 + 6(1 - n)]^{-1/2} \tag{165}$$

At the time of writing additional measurements on smaller scales are being reported. They have bigger uncertainties than the COBE measurement, and do not add substantially to what we learn by combining this measurement with the results of galaxy surveys in the manner discussed in the next section.

5 The CDM and MDM Models of Structure Formation

A model of large scale structure should simultaneously account for at least half a dozen completely different types of data, exploring a wide range of scales. Pure CDM with a flat spectrum ($n = 1$) performs rather impressively, but cannot actually fit all data within the accepted uncertainties, and the same remains true if one allows n to vary. Adding about 30% of hot dark matter does work, as far as theory and observation have been compared [30, 33, 40].

5.1 The filtered density contrast

At the present epoch the universe is highly inhomogeneous on small scales. In order to use linear cosmological perturbation theory one must therefore filter out the small scales, by smearing each perturbation over a region of size $\gtrsim 100\,\mathrm{Mpc}$. The same is true at epochs in the relatively recent past, except that the comoving filtering scale goes down. Only in the truly early universe is the universe (presumably) homogeneous on all relevant scales.

The filtering is done by means of a 'window function' $W(R_f, r)$, which is equal to 1 at $r = 0$ and which falls off rapidly beyond some radius R_f [1]. Taking for definiteness the density contrast, the filtered quantity is

$$\delta(R_f, \mathbf{x}) = \int W(R_f, |\mathbf{x}' - \mathbf{x}|)\delta(\mathbf{x}')d^3x' \tag{166}$$

and its spectrum is

$$\mathcal{P}_\delta(R_f, k) = \left[\widetilde{W}(R_f, k)/V_f\right]^2 \mathcal{P}_\delta(k) \tag{167}$$

where

$$\widetilde{W}(R_f, k) = \int e^{-i\mathbf{k}\cdot\mathbf{x}} W(R_f, r)d^3x \tag{168}$$

and

$$V_f = \int W(R_f, r)d^3x \tag{169}$$

The filtered dispersion is

$$\sigma^2(R_f) = \int_0^\infty \left[\widetilde{W}(R_f, k)/V_f\right]^2 \mathcal{P}_\delta(k)\frac{\mathrm{d}k}{k} \tag{170}$$

The quantity V_f is the volume 'enclosed' by the filter. It is convenient to define the associated mass $M = \rho_0 V_f$, where ρ_0 is the present mass density. One normally uses M instead of R_f to specify the scale, writing $\delta(M, \mathbf{x})$ and $\sigma(M)$.

The two popular choices are the Gaussian filter

$$\begin{aligned} W(R_f, r) &= \exp(-r^2/2R_f^2) && (171)\\ V_f &= (2\pi)^{3/2}R_f^3 && (172)\\ \widetilde{W}(R_f, k)/V_f &= \exp(-k^2R_f^2/2) && (173)\\ M &= 4.36\times 10^{12}h^2(R_f/1\,\mathrm{Mpc})^3 M_\odot && (174) \end{aligned}$$

and the top hat filter which smears uniformly over a sphere of radius R_f

$$\begin{aligned} W(R_f, r) &= \theta(r - R_f) && (175)\\ V_f &= 4\pi R_f^3/3 && (176)\\ \widetilde{W}(R_f, k)/V_f &= 3\left(\frac{\sin(kR_f)}{(kR_f)^3} - \frac{\cos(kR_f)}{(kR_f)^2}\right) && (177)\\ M &= 1.16\times 10^{12}h^2(R_f/1\,\mathrm{Mpc})^3 M_\odot && (178) \end{aligned}$$

The Gaussian filter is the most convenient for theoretical calculations, but the top hat filter is widely used to as a means of presenting data.

It is useful to write Eq. (170) in terms of the spectrum δ_H^2 of the primeval curvature. Using a Gaussian filter it becomes

$$\sigma^2(R_f) = \int_0^\infty e^{-k^2R_f^2}\left[T^2(k)\delta_H^2(k)\left(\frac{k}{aH}\right)^4\right]\frac{\mathrm{d}k}{k} \tag{179}$$

On the scales $k^{-1} \gtrsim 1\,\mathrm{Mpc}$ of most interest, the factor in square brackets increases quite strongly with k, so that the entire integrand peaks on the scale $k^{-1} \sim R_f$ and

$$\sigma^2(R_f) \sim \mathcal{P}_\delta(k) = T^2(k)\left(\frac{k}{aH}\right)^4 \delta_H^2(k) \tag{180}$$

with $k^{-1} = R_f$. Thus, $\sigma(R_f)$ falls off as the smearing scale R_f increases.

Figure 4: The dispersion $\sigma_0(M)$, of the linearly evolved filtered density contrast evaluated at the present epoch, is shown for pure CDM and MDM, with COBE normalization. This figure is taken from [2], which documents the precise input used.

One can filter the other perturbations as well. The filtered peculiar velocity is called the *bulk flow*. From Eq. (115), the dispersion of a given component of the bulk flow is

$$\sigma_v^2(R_f) = \int_0^\infty e^{-k^2 R_f^2} \left[T^2(k) \delta_H^2(k) \left(\frac{k}{aH} \right)^2 \right] \frac{\mathrm{d}k}{k} \sim \left(\frac{aH}{k} \right)^2 \mathcal{P}_\delta \tag{181}$$

It too falls off with increasing scale, but not as fast as the density contrast.

The bottom-up picture of structure formation

In Figure 4 is shown the prediction for the linearly evolved $\sigma(R_f)$, evaluated at the present epoch, for both pure CDM and for MDM. From Eq. (97), the pure CDM prediction scales at earlier epochs like $(1+z)^{-1}$.

The striking thing about the shape of $\sigma(R_f)$ is that it decreases monotonically as R_f increases. This feature implies a 'bottom-up' picture, with structure forming on successively larger scales. The epoch $z_{\rm nl}(M)$, when a significant fraction of the mass in the universe collapses into objects with mass M, is roughly the epoch when $\sigma(M) = 1$. The linear evolution $\sigma \propto a = (1+z)^{-1}$ given by Eq. (97) then ceases to be valid, but if it *had* remained valid the present value of σ would have been $(1 + z_{\rm nl})$. Thus

$$1 + z_{\rm nl}(M) = \sigma_0(M) \tag{182}$$

where the 0 denotes the present value of the *linearly evolved* quantity.

Figure 4 leads therefore to the following general picture. First, a large fraction of the matter in the universe collapses into objects with the mass of a small galaxy, or even smaller. Then successively large objects form, presumably gobbling up some or most of the smaller objects. At the present epoch, collapse is only just beginning on the scale $M \sim 10^{15} M_\odot$ which corresponds to large clusters.

5.2 Constraining the spectrum of the density contrast

A full comparison of theory with observation requires the use of numerical simulations, to follow the process of gravitational collapse which occurs on each scale after cosmological perturbation theory

OBSERVATION	NORMALISATION	REQUIRED σ_8
COBE	$\sigma_T(10^0) = (1.1 \pm .2) \times 10^{-5}$	$1.15 \pm .15$
bulk flow	$\sigma_{19} = .37 \pm .07$	$.85 \pm .15$
cluster abundance	$\sigma_8 = .46 \pm .16$	$.46 \pm .16$
galaxy correlation	$E = 1.21 \pm .07$	$E = .95$
quasar abundance	$\sigma_1 > .24$ at $z = 4$	$> .4$

Table 2: Key data and their implied normalisation, taken from [38]. The right hand column gives the value of σ_8 needed to fit each observation with pure CDM and a flat spectrum, except in the fourth row which gives E defined by Eq. (186)

breaks down. It turns out, though, that the linear theory can be applied on a wide variety of scales, so that one can obtain powerful constraints on the parameters by considering it alone. Working from the top down, some of the most important linear constraints are explained below.

- *The large scale cmb anisotropy* The COBE data explores scales of order the size of the observable universe, say 10^3 to 10^4 Mpc.

- *The bulk flow* Smearing the galaxy peculiar velocities over a sphere of radius tens of Mpc to get what is called the *bulk flow*, one should (just) be in the linear regime. In principle [6] one can observe the radial component of the bulk flow, construct the corresponding potential by integrating Eq. (112) radially, reconstruct $\mathbf{v}$ and finally determine the density perturbation $\delta\rho(\mathbf{x})$ from Eq. (110). In practice one has to use in addition the number density contrast $\delta n/n$ of infrared galaxies (for which the most complete data are available), assume that it is equal to $b_I\delta\rho/\rho$ with a position independent bias factor b_I, and estimate ψ from Eq. (113).

- *Galaxy cluster number density* The average number density $n(> M)$ of clusters with mass bigger than $M \sim 10^{15}M_\odot$ gives information on a scale of order $10h^{-1}$ Mpc. Within linear theory one can estimate $n(> M)$ by assuming that the matter in regions of space where $\delta(M, \mathbf{x})$ exceeds some critical value δ_c of order 1 is bound into objects with mass $> M$. The fraction of space occupied by such regions is

$$f(> M) = \mathrm{erfc}\left(\frac{\delta_c}{\sqrt{2}\sigma(M)}\right) \tag{183}$$

  From this assumption Press and Schechter derived the formula [2]

$$m\frac{\mathrm{d}n(> M)}{\mathrm{d}M} = \frac{\langle k^2\rangle}{12\pi^2 R_f}\nu e^{-\nu^2/2} \tag{184}$$

  where $\nu = \delta_c/\sigma(M)$ is the number of standard deviations that δ_c represents, and

$$\langle k^2(M)\rangle = \sigma^{-2}(M)\int_0^\infty k^2 e^{-k^2R_f^2}\mathcal{P}_\delta(k)\frac{\mathrm{d}k}{k} \tag{185}$$

  (This formula includes a more or less unmotivated factor 2). An alternative prescription is to identify each peak in $\delta(M, \mathbf{x})$ higher than δ_c with an object of mass $> M$, which gives a roughly similar result. Yet another method, which in principle is superior, is to run a numerical simulation, which again gives roughly similar results.

- *The shape of the galaxy correlation function* The galaxy correlation function Eq. (121) can be used to probe the shape of $\sigma(M)$ on scales between those explored by the last two items, if the bias factor is taken to be scale independent. The result is usefully quantified by giving the 'excess power'

$$E = 3.4\frac{\sigma(25h^{-1}\mathrm{Mpc})}{\sigma(8h^{-1}\mathrm{Mpc}}) \tag{186}$$

  where the prefactor is chosen so that $E = 1$ with pure CDM and $n = 1$. The observed value is somewhat higher.

Figure 5: The ratio of the MDM and pure CDM predictions shown in Figure 4 is shown, as a function of the filtering scale r. Also shown is the result of replacing MDM by pure CDM with a tilted spectrum. The figure is taken from [2], where the input is fully documented.

- *Quasar number density* Given some astrophysics assumptions, the observed quasar abundance can provide a lower limit on the density contrast at high redshift. For instance, one estimate [39] suggests that at $z = 4$, the fraction of mass bound into objects with $M > 10^{13} M_\odot$ is at least 1×10^{-7}. Using the Press-Schechter estimate with $\delta_c = 1.33$ [2, 33], this fraction is equal to 2 times Eq. (183), which gives $\sigma(M) > .24$ at $z = 4$.

A subjective view [38] of the normalisation of $\sigma(R_f)$ on various scales by these observations is given in the central column of Table 2, with the notation $\sigma_r \equiv \sigma(rh^{-1}\,\mathrm{Mpc})$. All values refer to the present except the last one.

First let us see how pure CDM with a flat spectrum fares. The value of σ_8 required to fits each observation is given in the right hand column. The striking thing is that all of the observations are consistent with this model to within a factor of 2. On the other hand the observations are not actually fitted; normalising to COBE the data fall below the prediction quite sharply as one moves down through the tens of Mpc regime.

In Figure 5 is plotted the ratio of the MDM to the CDM prediction, and one sees that it has a sharp drop of about the desired magnitude. Also shown is the effect of tilting the pure CDM spectrum, by making $n < 1$. This too reduces the small scale power, but the drop is not sharp enough to accomodate the galaxy correlation data quantified by E.

One can ask what constraints on n and the fraction Ω_ν of HDM are allowed by the data, if both are allowed to vary. Two investigations have been performed [38, 40]. The second analysis concludes that $.20 \lesssim \Omega_\nu \lesssim .35$, and that $n > .70$ (no gravitational waves) or $n > .85$ (gravitational waves as in Eq. (165)), and the first suggests even tighter constraints. Similar bounds on n also hold if one tries to rescue pure CDM by throwing away the troublesome galaxy correlation data [33, 2] (justified perhaps on the ground that it assumes a scale independent bias parameter).

These bounds on n are extremely significant in the context of inflation as we discuss in Section 6.

5.3 Alternatives to the MDM model

We have seen that pure CDM roughly accounts for observation, but that to get good agreement one needs a rather sharp reduction in power as one goes down in scale from $\sim 30h^{-1}\,$Mpc to $\sim 10h^{-1}\,$Mpc. The MDM proposal is one way of giving such a reduction. What about others?

Two have been tried so far. One is to invoke a cosmological constant with $\Omega_\Lambda \simeq .6$, leaving only $\Omega_m \sim .4$ in cold dark matter. While it goes in the right direction, this fix does not seem to be as successful as MDM, and is far less pleasing from a particle physics viewpoint. The other is to leave the transfer function alone, and require that the primeval power spectrum already has the sharp drop in power. Using two inflationary fields one can put this (or practically any other feature) into the spectrum, but such 'designer' models are quite unmotivated from a particle physics viewpoint. Alternatives which have not been tried yet are to use some kind of 'warm' dark matter.

Instead of fixing the CDM model one can contemplate throwing it away. Such a radical step seems unlikely to be needed, but if one wishes to take it there are two possibilities. One is to invoke topological defects to be the seeds of structure formation; the gauge cosmic strings discussed by Vachaspati in his lectures might work if the dark matter is hot, and as with the CDM model there is continuity as one goes up in scale from galaxies to the scale $\sim H_0^{-1}$ observed by COBE. The other is to use a primeval adiabatic or isocurvature density spectrum, but to give it small scale power so that the first objects to form have a mass of, say $1M_\odot$. This would lead to a radically different scenario, with no *a priori* connection between structure formation and the cmb anisotropy.

6 Inflation

It is widely supposed that the very early universe experienced an era of inflation [41, 1, 3]. By 'inflation' one means that the scale factor has positive acceleration, $\ddot{a} > 0$, corresponding to repulsive gravity. During inflation $aH = \dot{a}$ is increasing, so that comoving scales are leaving the horizon (Hubble distance) rather than entering it, and it is supposed that at the beginning of inflation the observable universe was well within the horizon.

Within the context of Einstein gravity, inflation requires negative pressure $p < \rho/3$ (Eq. (45)). This is achieved if p and ρ are dominated by a homogeneous, slowly varying 'inflaton field', $\phi(t)$, because one knows from standard field theory that

$$\rho = V + \frac{1}{2}\dot{\phi}^2 \tag{187}$$

$$p = -V + \frac{1}{2}\dot{\phi}^2 \tag{188}$$

The condition for inflation is clearly $\dot{\phi}^2 < V$. An alternative is to modify Einstein gravity, with or without invoking a scalar field, but in that case one can usually redefine the metric so as to recover the Einstein gravity during inflation. Until the end of this section I focus on the Einstein gravity case.

It is natural to assume that the field is sufficiently slowly varying that the condition $p < \rho/3$ is amply satisfied, so that the potential dominates ρ. One then has $\rho \simeq -p$, which from Eq. (43) makes ρ practically constant,

$$|H^{-1}\dot{\rho}/\rho| = 1 + p/\rho \ll 1 \tag{189}$$

6.1 General features of inflation

The most interesting prediction of inflation concerns the adiabatic density perturbation and the gravitational waves, which at some level it inevitably generates through vacuum fluctuations. Before examining them let us look at the general picture.

Chaotic initial conditions?

First let us ask how inflation is supposed to begin, and more generally what happened before it. A simple proposal, and the only one that has been explored in detail, is Linde's proposal of 'chaotic' initial conditions [3]. At some initial epoch a patch of the universe which includes our own as a small part, is roughly homogeneous and expanding roughly isotropically. Presumably this initial epoch occurs at the Planck scale, with energy density and expansion rate $\rho \sim H \sim m_{Pl}^4$, and it is reasonable to suppose that the size of the patch is of order H^{-1}.

Inflation is supposed to be achieved in the following way. The energy density comes from the various scalar fields existing in nature. The initial values of these fields are random, subject to the constraint that their potential $V(\phi, \psi, ..)$ is of order m_{Pl}^4 (it is supposed that their other contribution to ρ, coming from the spacetime dependence of the fields, is of the same order rather than much bigger.) The inflaton field ϕ is distinguished from the non-inflaton fields $\psi, ...$ by the fact that the potential is relatively flat in its direction. Before ϕ has had time to change much, the non-inflaton fields quickly adjust themselves to minimise the potential at fixed ϕ, after which inflation occurs as ϕ rolls slowly down the potential.

An alternative might be to suppose that the universe is initially *everywhere* roughly homogeneous and isotropic, with positive curvature so that there is no boundary. The chaotic hypothesis is more flexible in regard to model building, but otherwise the outcome is much the same.

The $\Omega = 1$ prediction

We now come to the first virtue of inflation. From Eq. (7), the time dependence of the density parameter is given by

$$\Omega - 1 = \left(\frac{K}{aH}\right)^2 \tag{190}$$

Without inflation, the patch that we are considering will typically either collapse in a Hubble time or so ($\Omega \to \infty$) or will become practically empty ($\Omega \to 0$). With inflation, in contrast, Ω is driven towards 1. Since Ω has the same value when observable universe leaves the horizon as it does at present, this leads to the expectation that *the present value of* Ω *is very close to 1.* This conclusion might be avoided if the observable universe leaves the horizon near the beginning of inflation, which as we remark in a moment is difficult to arrange, or if the initial value of Ω is rather finely tuned to zero [42]. Whether such escape routes are actually viable is not yet clear, but rightly or wrongly the 'prediction' $\Omega_0 = 1$ is widely regarded as one of the firmest consequences of inflation.

To summarise, *we need inflation so that the observable universe is sufficiently old, and is yet sufficiently dense to form structure.* Naturally, the thought suggests itself that non-inflating parts of the universe are common, but simply too inhospitable for us to inhabit.

Let us note, for future reference, that after Ω has been driven to a value fairly close to 1, the expectation Eq. (189) that ρ is slowly varying on the Hubble timescale implies the same for H, leading to an approximately exponential growth of the scale factor, $a \propto e^{Ht}$. We shall see that to have a density perturbation of the right magnitude, ρ at the end of inflation must be less than about $(10^{16}\,\mathrm{GeV})^4$, so if exponential inflation begins near the Planck scale it must last a very large number of Hubble times. This in turn would imply that the epoch at which the observable universe leaves the horizon occurs long after the beginning of inflation, because as we shall see this epoch typically occurs only 60 or so Hubble times before the end of inflation.

Perfect classical homogeneity

So far we need only be dealing only with a patch that is *roughly* homogeneous and isotropic. As in Section 1, the extreme homogeneity and isotropy of the observable universe is 'explained' by asserting that it is a very small part of the whole, and that the whole has a non-fractal nature.[17] However, inflation makes this explanation particularly attractive for two reasons. First, the statement that the observable universe is a very small part of the whole patch follows from the exponential growth of a combined with the constancy of H, provided that the observable universe leaves the horizon at least a few Hubble times after inflation begins [2].

Second and most crucially, we are free to go all the way and assert that the observable universe is *absolutely homogeneous* at the classical level, or in other words that every Fourier mode of every field is in the vacuum state. In fact, not only are we free to make this statement, it is mandatory on each scale well before horizon exit. The reason is that even one particle per momentum state on such scales would spoil inflation, by dominating the energy density [2].

[17] The inflationary 'explanation' of homogeneity and isotropy is often stated incorrectly. Namely, it is said that inflation 'smooths out' or even 'stretches' the observable universe. This seems to me to be very misleading. Before a scale leaves the horizon during inflation the expansion of the universe is not important, and we can think of field oscillations as particles. On the reasonable assumption that interactions are unimportant the the occupation number of each momentum state is *time independent.* After the scale leaves the horizon the field inhomogeneity on that scale is frozen in because causal processes cease to operate. This means that (except perhaps for rotational velocity perturbations, which would decay anyway) the observable universe is no more and no less homogeneous at the beginning of inflation than it is now.

If there is no inhomogeneity at the classical level, how is it generated? The answer, as we discuss in the next section, is that an inhomogeneity of well defined magnitude is generated in a perfectly natural way by the quantum fluctuation of the inflaton field.

Avoiding unwanted relics

Without inflation, thermal equilibrium starts at $T \sim 10^{16}\,\mathrm{GeV}$. Practically every particle is abundantly produced, and there can be many phase transitions creating defects. Some of these relics are definitely unwanted because they would be too abundant, notably examples being magnetic monopoles produced at $T \gtrsim 10^{11}\,\mathrm{GeV}$ and relic thermal gravitinos. Such relics can be avoided by inflation, because the temperature after inflation (reheat temperature) can be made low enough that the offending objects are not produced. Of course, wanted relics might also be killed; in particular structure forming cosmic strings need to have energy per unit length of $\sim 2 \times 10^{16}\,\mathrm{GeV}$, and are ruled out by most inflationary models (but see [18]).

Summary

There are two simple reasons for believing that the early universe started with an era of inflation. First, without inflation a given patch either collapses ($\Omega \to \infty$) or becomes practically empty ($\Omega \to 0$) within a Hubble time, unless its density parameter is very finely tuned to the value $\Omega = 1$. In contrast, inflation drives Ω towards 1 starting with an arbitrary initial value. Second, inflation can postpone the epoch of thermalisation of the universe, so that the temperature is too low to produce cosmologically disasterous relics.

In addition, there is the more complicated issue of homogeneity and isotropy. In order for inflation to occur, the energy density of the inflating patch must not be dominated by small scale inhomogeneities. In fact, on scales much less than the horizon (Hubble distance) even one particle per momentum state would be enough to spoil inflation; in other words the inflationary hypothesis requires that the universe is perfectly homogeneous at the classical level, on very small scales. As I now explain, the quantum fluctuation on a given comoving scale then generates a well defined inhomogeneity and anisotropy, which can be regarded as classical once the scale leaves the horizon, and can explain the perturbation the inhomogeneity and anisotropy of the observable universe.

6.2 The spectrum of the density perturbation

Inflation predicts a well defined adiabatic density perturbation in the following way. Well before horizon entry each Fourier mode of the inflaton field is described by quantum free field theory in (practically) flat spacetime. Its mean square quantum fluctuation can be calculated in a straightforward way, being analogous to the Casimir effect for the electromagnetic field which has been verified in the laboratory. Corresponding to the fluctuation is a well defined adiabatic density perturbation. Primordial gravitational waves are also generated, through the quantum fluctuation of the gravitational wave amplitude. In this section I explain concisely how to do these calculations.

The slow roll conditions

A homogeneous scalar field $\phi(\mathbf{x}, t)$ with minimal coupling to gravity has the equation of motion

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0 \tag{191}$$

Practically all of the usually considered models of inflation satisfy three conditions, usually termed the 'slow roll' conditions. Although the calculations can done without the slow roll conditions, they become much more complicated.

The first slow roll condition is that the motion of the field is overdamped, so that the 'force' V' balances the 'friction term' $3H\dot{\phi}$,

$$\dot{\phi} \simeq -\frac{1}{3H}V' \tag{192}$$

The second condition is that

$$\epsilon \equiv \frac{m_{Pl}^2}{16\pi}\left(\frac{V'}{V}\right)^2 \ll 1 \tag{193}$$

which means that the inflationary requirement $\dot{\phi}^2 < V$ is well satisfied and (assuming $\Omega = 1$)

$$H^2 \simeq \frac{1}{3}\frac{8\pi}{m_{Pl}^2}V \tag{194}$$

These two conditions imply that H is slowly varying, and that the scale factor increases more or less exponentially,

$$a \propto e^{Ht} \tag{195}$$

The third condition is that

$$|\eta| \ll 1 \tag{196}$$

where

$$\eta \equiv \frac{m_{Pl}^2}{8\pi}\frac{V''}{V} \tag{197}$$

It can be 'derived' from the other two by differentiating the approximation Eq. (57) for $\dot{\phi}$ and noting that consistency with the exact expression Eq. (45) requires $\ddot{\phi} \ll V'$ is satisfied. However there is no logical necessity for the derivative of an approximation to be itself a valid approximation, so this third condition is logically independent of the others.

It should be noted that the first slow-roll condition is on a quite different footing from the other two, being a statement about the *solution* of the field equation as opposed to a statement about the potential that defines this equation. What we are saying is that in the usually considered models of inflation, one can show that the first condition is an attractor solution, in a regime typically characterised by the other two conditions, and that moreover reasonable initial conditions on ϕ will ensure that this solution is achieved well before the observable universe leaves the horizon.

Classical equations

To define the perturbation $\delta\phi$ of the inflaton field, one has to choose a slicing of space-time into spacelike hypersurfaces, just as was the case for the perturbations $\delta\rho$, δp and δH that we studied earlier. Since the inflaton field is supposed to dominate the energy-momentum tensor, the momentum density vanishes if its spatial gradients vanish [43]. In other words, $\delta\phi$ vanishes if the hypersurfaces are chosen to be comoving!

In this situation one can proceed heuristically in the following way [29, 2]. First one notes, from Eq. (94), that in the extreme slow roll limit $p/\rho \to -1$ (corresponding to $\dot{\rho} \to 0$), the curvature perturbation $\mathcal{R}$ of comoving hypersurfaces becomes infinite, if the comoving density perturbation is finite. One should therefore define the inflaton field perturbation $\delta\phi$ on a family of hypersurfaces which remains undistorted in the slow roll limit. The time displacement δt of the comoving hypersurfaces from the undistorted ones is determined by the fact that $\delta\phi$ vanishes on the latter,

$$\delta t = -\delta\phi/\dot{\phi} \tag{198}$$

Finally, one can show either by brute force [21] or by an elegant geometrical argument [2] that the distortion of the comoving hypersurfaces caused by this time displacement implies a curvature perturbation $\mathcal{R} = -H\delta t$, which in terms of the field perturbation is

$$\mathcal{R} = H\delta\phi/\dot{\phi} \tag{199}$$

Alternatively one can derive this equation without any mention of the spatial geometry, working entirely with the density perturbation plus the definition Eqs. (91) and (93) of $\mathcal{R}$ [29].

Perturbing the inflaton field equation Eq. (191) leads to

$$(\delta\phi_{\mathbf{k}})^{\cdot\cdot} + 3H(\delta\phi_{\mathbf{k}})^{\cdot} + \left[\left(\frac{k}{a}\right)^2 + V''\right]\delta\phi_{\mathbf{k}} = 0 \tag{200}$$

Until a few Hubble times after horizon exit Eq. (62) ensures that V'' can be dropped, so that the equation becomes

$$(\delta\phi_{\mathbf{k}})^{\cdot\cdot} + 3H(\delta\phi_{\mathbf{k}})^{\cdot} + \left(\frac{k}{a}\right)^2\delta\phi_{\mathbf{k}} = 0 \tag{201}$$

A rigorous treatment [44, 45] is to define $\delta\phi$ on hypersurfaces with zero curvature perturbation, Eq. (199) then being an exact expression for the curvature perturbation of the comoving hypersurfaces [21]. One can show that to first order in the cosmological perturbations the field equation for $\delta\phi$ is

$$u_{\mathbf{k}}'' + (k^2 - z'' u_{\mathbf{k}}) = 0 \tag{202}$$

where $u = a\delta\phi$, $z = a\dot{\phi}/H$ and a prime denotes $a(\mathrm{d}/\mathrm{d}t)$. Finally, one can show that this equation reduces to Eq. (201) if the slow roll conditions are satisfied. (The correction can be calculated explicitly, and has been shown to be small [46].)

The quantum fluctuation

Well before horizon entry $\delta\phi_{\mathbf{k}}$ is a massless field living in (practically) flat space-time, since its wavenumber k/a is much bigger than the Hubble parameter H. It can be quantised in the usual way so that its quantum state is labelled by the number of inflaton particles present with momentum $\mathbf{k}$, and we are assuming that there are no particles, corresponding to the vacuum state.

Working in the Heisenberg representation, $\phi_{\mathbf{k}}$ is associated with an operator

$$\hat{\phi}_{\mathbf{k}}(t) = w_k(t)\hat{a}_{\mathbf{k}} + w_k^*(t)\hat{a}_{-\mathbf{k}}^\dagger \tag{203}$$

The annihilation operator $\hat{a}_{\mathbf{k}}$ satisfies the commutation relation

$$[\hat{a}_{\mathbf{k}\ 1}, \hat{a}_{\mathbf{k}\ 2}^\dagger] = \delta_{\mathbf{k}_1,\mathbf{k}_2} \tag{204}$$

and the field satisfies the commutation relation,

$$[\hat{\phi}(\mathbf{x}_1, t), \frac{\partial}{\partial t}\hat{\phi}(\mathbf{x}_2, t)] = i\delta^3(a\mathbf{x}_1 - a\mathbf{x}_2) \tag{205}$$

$$= ia^3\delta^3(\mathbf{x}_1 - \mathbf{x}_2) \tag{206}$$

As a result the functions w_k are given by

$$w_k(t) = a^{-3/2}(2k/a)^{-1/2}e^{-i(\chi+(kt/a))} \tag{207}$$

The phase factor χ is arbitrary, and it has negligible variation on the timescale a/k. The vacuum state is the one annihilated by $\hat{a}_{\mathbf{k}}$, so the vacuum expectation value of the field perturbation is

$$\langle|\delta\phi_{\mathbf{k}}|^2\rangle = |w_k|^2 \tag{208}$$

To extend these results to the epoch of horizon exit and beyond, we have to accept the validity of free field theory in curved space-time. All we need to assume is that there is a Heisenberg picture, in which operators satisfy the classical equations of motion and state vectors are time independent. Then $\hat{\phi}_{\mathbf{k}}$ continues to satisfy the field equation Eq. (200), and Eqs. (203), (204) and (208) still hold, where w_k is the solution of the field equation reducing to Eq. (207) well before horizon entry. As one easily checks, the required solution is

$$w_k(t) = \frac{H}{(2k^3)^{1/2}}\left(i + \frac{k}{aH}\right)e^{ik/aH} \tag{209}$$

A few Hubble times after horizon exit, the vacuum expectation value is therefore

$$\langle|\delta\phi_{\mathbf{k}}|^2\rangle = \frac{H^2}{2k^3} \tag{210}$$

A measurement of the $\phi_{\mathbf{k}}$'s will yield random phases, and a distribution of moduli whose dispersion is given by Eq. (210). Accordingly the spectrum of the inflaton field, defined by Eq. (23), is given a few Hubble times after horizon exit by

$$\mathcal{P}_\phi^{1/2}(k) = \frac{H}{2\pi} \tag{211}$$

Since H is slowly varying on the Hubble timescale it can be written, a few Hubble times after horizon exit,

$$\mathcal{P}_\phi^{1/2}(k) = \frac{H_*}{2\pi} \tag{212}$$

where the star denotes the epoch of horizon exit, and $\mathcal{P}_\phi$ is evaluated a few Hubble times after horizon exit.

The spectrum of $\mathcal{R}$ is given by

$$\mathcal{P}_{\mathcal{R}}^{1/2} = \frac{H}{\dot{\phi}} \mathcal{P}_\phi^{1/2} \tag{213}$$

In Section 2 we learned that $\mathcal{R}_{\mathbf{k}}$ is constant after horizon exit, so $\mathcal{P}_{\mathcal{R}}$ remains constant even though $H/\dot{\phi}$ and $\mathcal{P}_\phi$ might vary separately. It follows that as long as the scale is far outside the horizon,

$$\mathcal{P}_{\mathcal{R}}^{1/2} = \frac{H_*^2}{2\pi\dot{\phi}_*} \tag{214}$$

where the star denotes the epoch of horizon exit. Using Eq. (57), this leads to [29]

$$\delta_H^2(k) = \frac{32}{75} \frac{V_*}{m_{Pl}^4} \epsilon_*^{-1} \tag{215}$$

Primordial gravitational waves?

The gravitational wave amplitude h_{ij} is dynamically equivalent to a minimally coupled, massless scalar field $\psi = (m_{Pl}^2/16\pi)^{1/2} h_{ij}$. Each non-zero Fourier component has the same vacuum fluctuation as a massless field, with the above conversion factor. Thus the spectrum of the gravitational waves is given by Eq. (211),

$$\mathcal{P}_g(k) = 4 \times \frac{16\pi}{m_{Pl}^2} \left(\frac{H_*}{2\pi} \right)^2 \tag{216}$$

Putting this expression into Eq. (150) and dividing it by Eq. (146) gives the ratio Eq. (152) of the gravitational wave and density contributions to the cmb anisotropy.

The classicality of the perturbation

The quantum fluctuation of a field in flat spacetime cannot be regarded as classical. The reason is that each Fourier mode is equivalent to a harmonic oscillator in its ground state; the position and momentum of such an oscillator cannot be accurately measured simultaneously (unless the measured results are far out on the tail of the probability distribution), and as a result the position cannot be well defined over an extended period of time. Returning to the case of the field, this means that a Fourier component cannot be defined with high accuracy over an extended period of time.

After horizon exit though, a given Fourier component is no longer equivalent to a harmonic oscillator and one can show that it *can* have a well defined value [29, 47]. In this sense it is classical, though of course one still has the 'Schrodinger's cat' problem of how and when it is supposed to have acquired this well defined value, drawn from the quantum probablity distribution.

The inflationary energy scale

Assuming that the adiabatic density perturbation dominates the cmb anisotropy, we saw earlier that gives $\delta_H \simeq 2.0 \times 10^{-5}$ on the Hubble scale. Eq. (215) therefore gives the energy scale at the epoch when the observable universe leaves the horizon during inflation,

$$V_1^{1/4} = 7.3 \times 10^{16}\,\mathrm{GeV} \times \left(\frac{\delta_H \epsilon_1}{2.0 \times 10^{-5}} \right)^{1/4} \tag{217}$$

Since $\epsilon_1 \ll 1$ we conclude that $V^{1/4}$ is less than a few times $10^{16}\,\mathrm{GeV}$. Since the energy density decreases with time, this is also an upper limit on the energy density at the end of inflation, and on the reheat temperature.

6.3 Entering and leaving inflation

In order to work out the observational consequences of Eq. (217) for particular inflationary models, we need to do a little more work.

First, we need to ask how inflation ends. In the chaotic picture, two possibilities exist concerning the minimum into which the non-inflaton fields fall. The simplest possibility is that it corresponds to

the true vacuum; that is, that the non-inflaton fields have the same values as in the present universe. Taking the potential to be quadratic for simplicity, it then has the form

$$V = \frac{1}{2}m^2\phi^2 \tag{218}$$

Inflation ends when the inflaton field starts to execute decaying oscillations around its own vacuum value (zero in the above example). This typically occurs when the slow roll conditions fail,

$$\max\{\epsilon_{\rm end}, |\eta_{\rm end}|\} \simeq 1 \tag{219}$$

The hot Big Bang ensues when the vacuum value has been achieved and the decay products have thermalised. Models of this type have been widely explored

The other possibility [16, 2, 17, 18, 19] is that the minimum *at large fixed* ϕ corresponds to a false vacuum, which dominates the ϕ-dependent part of the potential. In the quadratic case the potential is then

$$V = V_0 + \frac{1}{2}m^2\phi^2 \simeq V_0 \tag{220}$$

where V_0 is the false vacuum energy density. In this case inflation ends when the false vacuum is destabilized, as ϕ falls through some critical value ϕ_c.

Of course one can envisage more complicated possibilities. One is that an early epoch of inflation gives way to a hot universe, which finds itself in a false vacuum which is stabilized by the finite temperature. The vacuum energy density V_0 dominates when the temperature falls below $V_0^{1/4}$, and inflation occurs for a more or less brief era as the field rolls slowly away from its minimum, ending when it starts to oscillate around its true vacuum value. This way of beginning inflation is called 'new inflation' [1]. Apart from being more complicated, it is also even more difficult to implement in the context of sensible particle physics than the other two proposals.

Reheating

In any model, the universe (or at least the constituents dominating the energy density) must achieve thermal equilibrium at some point after inflation. Because the first models of inflation invoked a thermal entry, this event is usually called 'reheating'. The corresponding 'reheat temperature' is the biggest temperature ever achieved after inflation and plays a crucial role in cosmology. Unfortunately, no reliable estimate of it is known at present. If inflation ends with a phase transition, reheating is expected to be prompt. If it ends with the inflaton field oscillating, reheating may be long delayed because the couplings of the inflaton field to other fields is then typically quite weak.

The epoch of horizon exit

A first requirement for an inflationary model is that the observable universe should be within the horizon at the beginning of inflation. We therefore need to know the epoch when a given scale leaves the horizon, something which is necessary also to calculate the inflationary perturbations. Denoting it by a star, this epoch is given by

$$a_* H_* = k \tag{221}$$

The epoch of horizon exit is therefore related to the present magnitude of the scale in Hubble units by

$$\frac{a_0 H_0}{k} = \frac{a_0 H_0}{a_* H_*} \tag{222}$$

Let us denote the end of inflation by a subscript 'end' and the epoch of reheating by 'reh', assuming matter domination during the era (if any) between these two epochs. Let us also assume that after 'reh' there is radiation domination, until the epoch 'eq' at which the dark matter density becomes equal to that of the radiation. Throughout the history of the universe the radiation energy density is proportional to a^{-4}, that of the matter is proportional a^{-3}, and the total is proportional to H^2. It follows that

$$\frac{k}{a_0 H_0} = \frac{a_*}{a_{\rm end}}\frac{a_{\rm end}}{a_{\rm reh}}\frac{a_{\rm reh}}{a_0}\frac{H_*}{H_0} \tag{223}$$

$$= e^{-N_*}\left(\frac{\rho_{\rm reh}}{\rho_{\rm end}}\right)^{1/3}\left(\frac{\rho_{0r}}{\rho_{\rm reh}}\right)^{1/4}\left(\frac{\rho_*}{\rho_0}\right)^{1/2} \tag{224}$$

where $\rho_{0r} = (a_{\rm eq}/a_0)\rho_0$ is the present radiation energy density and N_* is the number of Hubble times between horizon exit and the end of inflation,

$$N_* \equiv \ln(a_*/a_{\rm end}) \tag{225}$$

$$\simeq \int_{t_*}^{t_{\rm end}} H dt \tag{226}$$

It follows that

$$N_* = 62 - \ln\frac{k}{a_0 H_0} - \ln\frac{10^{16}\,{\rm GeV}}{V_*^{1/4}} + \ln\frac{V_*^{1/4}}{V_{\rm end}^{1/4}} - \frac{1}{3}\ln\frac{V_{\rm end}^{1/4}}{\rho_{\rm reh}^{1/4}} \tag{227}$$

This equation relates the three energy scales $V_*^{1/4}$, $V_{\rm end}^{1/4}$ and $\rho_{\rm reh}^{1/4}$. The first two scales are related by another expression for N_*, which follows from Eqs. (226), (192), (193) and (194)

$$N_* = \frac{8\pi}{m_{Pl}^2}\int_{\phi_{\rm end}}^{\phi_*}\frac{V}{V'}d\phi = \sqrt{\frac{4\pi}{m_{Pl}^2}}\left|\int_{\phi_{\rm end}}^{\phi_*}\epsilon^{-1/2}d\phi\right| \tag{228}$$

The biggest scale that can be explored is roughly the present Hubble distance, $a_0/k = H_0^{-1} = 3000h^{-1}\,{\rm Mpc}$. I shall refer to the epoch of horizon exit for this scale as 'the epoch when the observable universe leaves the horizon', and denote it by the subscript 1. As displayed in Eq. (217), the COBE observations require that $V_1^{1/4} \lesssim 10^{16}\,{\rm GeV}$, with the equality holding in most models of inflation. Also, Eq. (228) gives in most models of inflation $V_1^{1/4} \simeq V_{\rm end}^{1/4}$. If reheating is prompt, we learn that in most models the observable universe leaves the horizon about 62 e-folds before the end of inflation. If reheating is long delayed N_1 could be considerably reduced, being equal to 32 for the most extreme possibility of $V_{\rm reh}^{1/4} \sim 1000\,{\rm GeV}$ (corresponding to reheating just before the electroweak transition which is presumably the last opportunity for baryogenesis). For most purposes, however, one needs only the order of magnitude of N_1.

The smallest scale on which the primeval perturbation can be probed at present is around 1 Mpc, and one sees that this scale leaves the horizon about 9 Hubble times after the observable universe. Discounting the case of very late reheating, we conclude that in the usually considered inflationary models, scales of cosmological interest leave the horizon 50 to 60 Hubble times before the end of inflation.

The spectral index

The spectral index n of the density perturbation is obtained by differentiating Eq. (215) with the aid of Eqs. (192), (193), (194), (221) and (195)

$$n = 1 + 2\eta_1 - 6\epsilon_1 \tag{229}$$

For definiteness I have taken to be the epoch when the observable universe leaves the horizon.

6.4 Specific models of inflation

Of the many viable inflationary potentials that one could dream up, only a few are reasonable from a particle physics viewpoint. Some of them are described now.

True vacuum chaotic inflation

If the potential is $V \propto \phi^\alpha$, Eq. (219) shows that at the end of inflation $\phi_{\rm end} \simeq \alpha(m_{Pl}^2/8\pi)^{1/2}$. From Eq. (228) with $N_1 = 60$, $\phi_1^2 \simeq 120\alpha m_{Pl}^2/8\pi$. This leads to $1-n = (2+\alpha)/120$ and $R = .05\alpha$, leading to the following relation between the gravitational wave contribution and the spectral index

$$R = 6(1-n) - 0.1 \tag{230}$$

For $\alpha = 2$, 4, 6 and 10 one has $1-n = .033$, .05, .067 and .1, and $R = .10$, .20, .30 and .50. Lazarides and Shafi [48] have shown that a potential of this kind follows from a class of superstring-inspired gauge theories, with an index α varying between 6 and 10.

False vacuum chaotic inflation

The simplest model of false vacuum inflation is [16, 2, 17, 18, 19]

$$V(\psi,\phi)=\frac{1}{4}\lambda(\psi^2-M^2)^2+\frac{1}{2}m^2\phi^2+\frac{1}{2}\lambda'\phi^2\psi^2 \tag{231}$$

The couplings λ and λ' are supposed to be somewhat less than unity, and in the numerical examples I set $\lambda=\lambda'=0.1$. For $\phi^2>\phi_{\rm end}^2=\lambda M^2/\lambda'$, the potential for the ψ field has a local minimum at $\psi=0$, in which the field is assumed to sit (except for the quantum fluctuation). Inflation can occur as ϕ slowly rolls down its potential

$$V(\phi)=\frac{1}{4}\lambda M^4+\frac{1}{2}m^2\phi^2 \tag{232}$$

We will suppose that while observable scales leave the horizon the first term dominates, since in the opposite case we recover the ϕ^2 potential already considered. When ϕ falls below $\phi_{\rm end}$ this local minimum becomes a local maximum, and a second order phase transition to the true vacuum occurs which ends inflation and causes prompt reheating.

The requirement that the first term dominates means that

$$\frac{2}{\lambda}\frac{m^2\phi_1^2}{M^4}\ll 1 \tag{233}$$

Of the parameters ϵ and η which are required to be small, the second is independent of ϕ,

$$\eta=\frac{4}{\lambda}X^2 \tag{234}$$

where

$$X^2\equiv\frac{m_{Pl}^2}{8\pi}\frac{m^2}{M^4} \tag{235}$$

The ratio X must therefore be significantly less than 1. From Eq. (228),

$$\phi_1=\sqrt{\lambda/\lambda'}\,Me^{N_1\eta} \tag{236}$$

Consistency with Eq. (233) requires roughly $\eta\lesssim.1$. The other small quantity ϵ_1 is given by

$$\epsilon_1=\frac{1}{2}\frac{\lambda}{\lambda'}\frac{8\pi}{m_{Pl}^2}M^2\eta^2e^{2N_1\eta} \tag{237}$$

and Eq. (233) requires $\epsilon_1\ll\eta$. It therefore follows that the spectral index n is *bigger* than 1 in the two-scale model.

Setting $\lambda=\lambda'=.1$ and imposing the COBE normalisation $\delta_H=1.7\times10^{-5}$ determines all of the parameters in terms of m. For $m=100\,\mathrm{GeV}$ one has $M=4\times10^{11}\,\mathrm{GeV}$ leading to $\eta=10^{-4}$ and $\epsilon_1=10^{-23}$. The gravitational waves are absolutely negligible, and the spectral index is extremely close to 1. The maximum value of m permitted by Eq. (233) is roughly $m=10^{13}\,\mathrm{GeV}$, giving $M=2\times10^{16}\,\mathrm{GeV}$, $\eta=.07$ and $\epsilon_1=10^{-3}$. The gravitational waves are still negligible, but $n=1.14$, significantly *bigger* than 1.

So far we have considered a single scalar field ψ, possessing the discrete symmetry $\psi\to-\psi$. When inflation ends, domain walls will form along the surfaces in space where ψ is exactly zero, so to make the model cosmologically viable one would have to get rid of the walls by slightly breaking the symmetry. However, it is clear that one can easily consider more fields, with continuous global or gauge symmetries. In particular, if ψ is complex one can use the same potential with the replacement $\psi^2\to|\psi|^2$. This leads to exactly the same inflationary model, but now global strings are formed instead of domain walls.

The case $m\sim100\,\mathrm{GeV}$ and $M\sim10^{11}\,\mathrm{GeV}$ is particularly interesting because these are rather natural scales in particle physics. One possibility is to identify ψ with the Peccei-Quinn field, and it might also be possible to realise the model in the context of superstrings [18]

Natural Inflation.

The 'natural inflation' model [49, 50] has the potential

$$V(\phi) = \Lambda^4 \left(1 \pm \cos(\phi/f)\right) \tag{238}$$

where Λ and f are mass scales. In this model

$$\epsilon = r^{-1} \tan^2 \frac{\phi}{2f} \tag{239}$$

$$\eta = -r^{-1} \left(1 - \tan^2 \frac{\phi}{2f}\right) \tag{240}$$

where $r \equiv (16\pi f^2)/m_{Pl}^2$. Since $\epsilon - \eta = r^{-1}$ one must have $r \gg 1$ to be in the slow-roll regime anywhere on this potential. This requires that f is at the Planck scale, suggesting a connection with superstrings which has been explored in [50].

In this model inflation has to begin with ϕ close to zero. If there is a preceding thermal era, one might expect ϕ to be chosen randomly so that a value close to zero is certainly allowed. With or without a thermal era, additional fields have to be involved to have a chaotic beginning at the Planck scale.

Inflation ends with an oscillatory exit at the epoch given by Eq. (219), $\tan(\phi_{\rm end}/2f) \simeq r^{1/2}$ which is of order 1. Using this result, Eq. (228) with $N_1 = 60$ gives $\phi_1/2f \simeq \exp(-60/r)$, leading to

$$\epsilon_1 = \frac{1}{r} e^{-120/r} \tag{241}$$

$$\eta_1 \simeq -1/r \tag{242}$$

$$1 - n \simeq 2/r \tag{243}$$

Thus, natural inflation makes the gravitational waves negligible but tends to give a significantly tilted spectrum. The observational bound $n > .7$ implies that $r > 6$.

R^2 inflation

The Lagrangian density for Einstein gravity is just $-(m_{Pl}^2/16\pi)R$, where R is the spacetime curvature scalar. Perhaps the simplest model of inflation is to replace R by $R + R^2/(6M^2)$. By redefining the metric one can then recover Einstein gravity, at the expense of modifying the rest of the Lagrangian *and* adding a completely new field ϕ with a potential

$$V(\phi) = \frac{3m_{Pl}^2 M^2}{32\pi} \left[1 - \exp\left(-\left(\frac{16\pi}{3m_{Pl}^2}\right)^{1/2} \phi\right)\right]^2 \tag{244}$$

In the regime $\phi \gtrsim m_{Pl}$, the potential $V(\phi)$ satisfies the slow-roll conditions so that inflation occurs. The non-gravitational sector is irrelevant during inflation, and if the R^2 term quickly becomes negligible afterwards we have a model in which inflation is driven by a scalar field, without any modification of gravity. What has been achieved is to motivate the otherwise bizarre form of the inflationary potential.

In the regime $\phi \gtrsim m_{Pl}$ where they are small, the parameters appearing in the slow-roll conditions are

$$\eta = -\frac{4}{3} \exp\left(-\sqrt{\frac{2}{3}} \frac{\sqrt{8\pi}}{m_{Pl}} \phi\right) \tag{245}$$

$$\epsilon = \frac{3}{4} \eta^2 \tag{246}$$

There is an oscillatory exit to inflation at the epoch given by Eq. (219), $\phi_{\rm end} \sim (m_{Pl}^2/8\pi)^{1/2}$. From Eq. (228) with $N_1 = 60$,

$$\phi_1 \simeq 5 \frac{m_{Pl}}{\sqrt{8\pi}} \tag{247}$$

leading to $\eta_1 \simeq -.02$ and $\epsilon_1 \sim 10^{-4}$. Thus the spectral index is $n = 0.96$, but the gravitational wave contribution is completely negligible.

Justification for this model of inflation has recently been claimed in the context of superstrings [51]

Extended inflation

'Extended inflation' models are based on modifications of Einstein gravity of the Brans-Dicke type [52]. After recovering Einstein gravity by redefining the metric, one typically has an inflationary scalar field, with the potential

$$V = V_0 \exp\left(\sqrt{\frac{16\pi}{pm_{Pl}^2}}\phi\right) \tag{248}$$

Irrespective of when inflation ends, one finds $-n_g = 1 - n(= 2/p)$, and the 'canonical' gravitational wave ratio $R \simeq 6(1-n)$ that we discussed earlier.

Inflation in these models ends with a first order phase transition (caused, in the Einstein gravity picture, by the modification of the non-gravitational Lagrangian induced by the metric transformation). The bubbles that it generates must not cause an unacceptable microwave background anisotropy, which generally leads to the constraint $n \lesssim 0.75$ [53, 33], in conflict with the observational bound $n > .85$ that we noted in Section 5. One needs to make these models quite complex in order to avoid this difficulty [54, 17].

6.5 Summary and prospects

Particle physics, especially in the context of superstrings, is beginning to hint at several viable models of inflation. Each of them makes a distinctive prediction regarding the tilt of the density perturbation spectrum, which will allow one to discriminate between them in the near future provided that some variant of the CDM model continues to fit the observations. One will then, for the first time, have an observational window on energy scales not far below the Planck scale.

References

[1] E. W. Kolb and M. S. Turner, *The Early Universe*, Addison Wesley (1990).
T. Padmanabhan, *Structure Formation in the Universe*, Cambridge University Press (1993).
P. J. E. Peebles, *Principles of Physical Cosmology*, Princeton University Press, N.J. (1993).

[2] A. R. Liddle and D. H. Lyth, Phys. Rep. **231**, 1 (1993).

[3] A. D. Linde, Phys. Lett. **B129**, 177 (1983);
A. D. Linde, *Particle Physics and Inflationary Cosmology*, Harwood Academic, Switzerland (1990).
A. Linde and A. Mezhlumian, Phys. Lett. **307**, 25 (1993).

[4] T. Walker *et al*, Ap. J. 376, 51 (1991).

[5] E. Aubourg *et al*, Nature **365**, 623 (1993); C. Alcock *et al*, *ibid*, 621.

[6] A. Dekel, E. Bertschinger and S. M. Faber, Ap. J. **364**, 349 (1990). A. Dekel *et al*, Astroph. J. **412**, 1 (1993).

[7] S. M. Carroll, W. H. Press and E. L. Turner, Ann. Rev. Astron. and Astroph., 30, 499 (1992).

[8] D. Maoz and H.-W. Rix, to appear in Astro. J **415** (1993).

[9] S. A. Bonometto and O. Pantano, Phys. Reports, **228**, 175 (1993).

[10] D. H. Lyth, Phys. Rev. D**45**, 3394 (1992).
D. H. Lyth and E. D. Stewart, Phys. Rev. D**46**, 532 (1992).

[11] D. H. Lyth, Phys Rev D **48**, 4523 (1993).

[12] S. A. Bonometto, F. Gabbiani and A. Masiero, (1993)

[13] E. J. Chun, N. B. Kim and J. E. Kim (1993).

[14] J. Madsen, Phys. Rev. Lett. **69**, 571 (1993).
N. Kaiser, R. A. Malaney and G. D. Starkman, Phys. Rev. Lett. **71**, 1128 (1993).

[15] E. W. Kolb and I. I. Tkachev, Phys. Rev. Letts. **71**, 3051 (1993).

[16] A. D. Linde, Phys. Lett. **B259**, 38 (1991).

[17] A. D. Linde, "Hybrid Inflation", Stanford preprint SU-ITP-93-17 (1993).

[18] E. J. Copeland, A. R. Liddle, D. H. Lyth, E. D. Stewart and D. Wands, 'False vacuum inflation with Einstein gravity', preprint (1993).

[19] S. Mollerach, S. Matarrese and F. Lucchin, "Blue Perturbation Spectra from Inflation", CERN preprint (1993).

[20] E. M. Lifshitz J. Phys. (Moscow) **10**, 116 (1946).

[21] J. M. Bardeen, Phys. Rev. D **22**, 1882 (1980).
V.F. Mukhanov, H. A. Feldman and R. H. Brandenberger, Phys. Rep. **215**, 203 (1992).

[22] H. Kodama and M. Sasaki, Prog. Theor. Phys., **78**, 1 (1984), and references traced from N. Sugiyama, N. Gouda and M. Sasaki, Astrophys. J. **365**, 432 (1990). R. K. Schaefer, Int. J. Mod. Phys. **A6**, 2075 (1991). R. K. Schaefer and Q. Shafi, Phys. Rev. D **47**, 1333 (1993).

[23] Hawking, S. W., Astrophys. J., **145**, 544 (1966). D. W. Olson, Phys. Rev. D, **14**, 327 (1976).

[24] D. H. Lyth and M. Mukherjee, Phys. Rev. D **38**, 485 (1988).

[25] G. F. R. Ellis and M. Bruni, Phys. Rev. D, **40**, 1804 (1989).

[26] D. H. Lyth and E. D. Stewart, Astrophys. J., **361**, 343 (1990).

[27] M. Bruni, P. K. S. Dunsby and G. F. R. Ellis, Astrophys. J. **395**, 34 (1992).

[28] P. K. S. Dunsby, M. Bruni and G. F. R. Ellis, Astrophys. J. **395**, 54 (1992).

[29] D. H. Lyth, Phys. Rev. D**31**, 1792 1985.

[30] S. A. Bonometto and R. Valdarnini, Phys. Lett. **103A**, 369 (1984).
Q. Shafi and F. W. Stecker, Phys. Rev. Lett. **53**, 1292 (1984).
L. Z. Fang, S. X. Li and S. P. Xiang, Astron. Astrophys. **140**, 77 (1984).
R. Valdarnini and S. A. Bonometto, *Astron Astrophys* **146**, 235 (1985).
S. Achilli, F. Occhionero & R. Scaramella, *Astrophys J* **299**, 577 (1985).
S. Ikeuchi, C. Norman & Y. Zhan *Astrophys J* **324**, 33 (1988).
R. K. Schaefer, Q. Shafi and F. Stecker *Astrophys J* **347**, 575 (1989).
J. Holtzman, *Astrophys J Supp* **71**, 1 (1989).
E. L. Wright *et al*, *Astrophys J Lett* **396**, L13 (1992).
R. K. Schaefer and Q. Shafi, Nature, **359**, 199, (1992).
M. Davis, F. J. Summers and D. Schlegel, Nature, **359**, 393 (1992).
A. N. Taylor and M. Rowan-Robinson, Nature, **359**, 396 (1992).
T. van Dalen and R. K. Schaefer, ApJ, **398**, 33 (1992).
R. K. Schaefer and Q. Shafi, Phys. Rev., **D47**, 1333, (1993).
J. Holtzman and J. Primack, *Astrophys J* **405**, 428 (1993).
A. Klypin, J. Holtzman, J. R. Primack, and E. Regös, Astrophys J, **416**, 1 (1993).
D. Yu. Pogosyan and A. A. Starobinsky, "Confrontation of the CDM+HDM Model with Observational Data", Cambridge preprint (1993).
Y. P. Jing, H. J. Mo, G. Börner & L. Z. Fang, to appear in Astron. and Astrophys. (1993).
S. A. Bonometto, S. Borgani, S. Ghigna, A. Klypin and J. R. Primack, submitted to Mon. Not. Roy. Ast. Soc. (1993).

[31] M. Bruni and D. H. Lyth, Lancaster preprint (1993).

[32] M. E. Machacek, 'Growth of perturbations in self-interacting dark matter', preprint (1993).

[33] A. R. Liddle and D. H. Lyth, Phys. Lett. **B291**, 391 (1992).

[34] R. Crittenden, J. R. Bond, R. L. Davis, G. Efstathiou, P. J. Steinhardt and M. S. Turner, Phys. Rev. Letts., **71**, 324 (1993).

[35] A. A. Starobinsky, Sov. Astron. Lett. **11**, 133 (1985).

[36] G. F. Smoot *et. al.*, Astrophys. J. Letts. **396**, L1 (1992).

[37] E. L. Wright *et al*, "Comments on the Statistical Analysis of Excess Variance in the COBE DMR Maps", to appear, Astrophys. J. (1994).

[38] A. R. Liddle and D. H. Lyth, Mon. Not. Roy. Astr. Soc. **265**, 379 (1993).

[39] Haehnelt, M. G., 1993, "High redshift constraints on alternative spectra for primeval density perturbations", Cambridge Institute of Astronomy preprint

[40] R. K. Schaefer and Q. Shafi, *A Simple Model of Large Scale Structure Formation*, Bartol preprint (1993).

[41] A. H. Guth, Phys. Rev. D**23**, 347 (1981).

[42] D. H. Lyth and E. D. Stewart, Phys Lett. **B252**, 336 (1993).

[43] J. M. Bardeen, P. S. Steinhardt and M. S. Turner, Phys. Rev. D**28**, 679 (1993).

[44] V. F. Mukhanov, *JETP Lett* **41**, 493 (1985).

[45] M. Sasaki, *Prog Theor Phys* **76**, 1036 (1986).

[46] Stewart, E. D., Lyth, D. H., Phys. Lett., **B302**, 171, (1993).

[47] A. H Guth and S.-Y. Pi, Phys. Rev. D**32**, 1899 (1985).

[48] G. Lazarides and Q. Shafi, Phys. Lett. B (in press) (1993)

[49] K. Freese, J. A. Frieman and A. Olinto, Phys. Rev. Letts. *65*, 3235 (1993).

[50] F. C. Adams, J. R. Bond, K. Freese, J. A. Frieman and A. V. Olinto, Phys. Rev. D**47**, 426 (1993).

[51] G. L. Cardoso and B. A. Ovrut, Phys. Letts. **B298**, 292 (1993).

[52] D. La and P. J. Steinhardt, Phys. Rev. Lett. **62**, 376 (1989).
P. J. Steinhardt and F. S. Accetta, Phys. Rev. Lett. **64**, 2740 (1990).
J. D. Barrow and K. Maeda, Nucl. Phys. **B341**, 294 (1990).

[53] E. J. Weinberg, Phys. Rev. D**40**, 3950 (1989);
A. R. Liddle and D. Wands, Mon. Not. Roy. astr. Soc. **253**, 637 (1991).
A. R. Liddle and D. Wands, Phys. Rev. D**45**, 2665 (1992);
M. S. Turner, E. J. Weinberg and L. M. Widrow, Phys. Rev. D**46**, 2384 (1992).

[54] A. M. Laycock and A. R. Liddle, "Extended Inflation with a Curvature-Coupled Inflaton", Sussex preprint SUSSEX-AST 93/6-1, astro-ph/9306030 (1993).