The Early Universe ... according to David Wands

Contents

1 Introduction

Cosmology is not a new subject, but it is a new science. The questions are old — how big is the world, what is our place in the universe, how old is it — but the idea that we can attempt to answer these questions by scientific theory and observations is a new one. The aim of this course will be to try and give you some of the current answers to these profound questions.

I say current answers, because some current answers you probably should believe and some you probably should not. This is not a cause for despair, but rather an introduction to the uncertain world of scientific research. Science does not grant us absolute knowledge of how the world is, but instead enables us to build up working models through which we hope to make sense of complex observations. Astronomical observations often rely on a lot of model-dependent processing to turn them into the "observational results" that appear in textbooks and this is especially true in cosmology where we are reaching out to the most distant objects in the universe, beyond interfering junk like our atmosphere, our Solar System and our Galaxy.

I will try to give you some feeling for which answers I think you should believe. There is good observational evidence for the standard model of the hot big bang. The first part of the course will cover this standard model and the theoretical framework it gives us for studies of the early universe. There are some solid observational branches to cling onto here. As we move onto the upper reaches of the cosmological tree the current answers are liable to have a shorter life-expectancy. The "observational results" become more sparse and indirect and the extrapolation of physical theories becomes more extreme. But the temptations are too great to resist and theorists often venture out on flimsy support. Some twigs snap, but others survive, strengthen and grow.

Over recent years technological advances, satellite experiments and space telescopes have dramatically increased the quantity and quality of observational evidence with which to test our theories and it is a useful to start by reviewing the observational cornerstones of the standard model and assess their current status.

1.1 The Hubble expansion.

In 1929 Hubble published his results on the relation between the distance to nearby galaxies and their velocities relative to us, determined by the Doppler shift (or "redshift") of their spectral lines. It was pretty rough data, but he showed an approximately linear relationship between recessional velocity, v, and distance r,

$$v = H_0 r \tag{1}$$

where H_0 has become known as Hubble's constant. This "observational result" established the picture of an expanding universe which has been reinforced by numerous surveys since.

Hubble was less successful in establishing the rate of expansion. He estimated that a galaxy's recessional velocity was about 500kms^{-1} times its distance in mega-parsecs ($1\text{Mpc} \equiv 3 \times 10^{24}\text{cm}$). Determining this of course requires an accurate estimate of the distance of a galaxy, i.e. its absolute as well as apparent magnitude. Successive revisions of the absolute magnitudes and remaining uncertainties in their correct values leaves the present-day estimates of the Hubble constant somewhere between 40 and 80 $\mathrm{kms^{-1}Mpc^{-1}}$.

One of the key projects of the Hubble Space Telescope has been to try and definitively pin down the actual value. A team lead by Wendy Freedman finally reported a value of 72 ± 7 $\mathrm{kms^{-1}Mpc^{-1}}$, and this range is now accepted by most, if not all astronomers. Nonetheless it is still important to bear in mind that there are many different ways of estimating H_0 which yield almost as many different values and they can't all be right! Our uncertainty is usually signified by the parameter, h. Throughout this course I will use

$$H_0 = 100\ h \text{ km s}^{-1} \text{ Mpc}^{-1} \tag{2}$$

with $h = \{0.4, 0.8\}$.

The Hubble constant also gives a characteristic timescale,

$$H_0^{-1} = 10^{10} \left(\frac{100\text{km s}^{-1}\text{Mpc}^{-1}}{H_0} \right) \text{ years,} \tag{3}$$

and in a Big Bang model the age of the universe is predicted to be of order H_0^{-1}. At the same time, stellar evolution models estimate the age to be at least 1.2×10^{10}years, so we're in the right ball-park.

The simplest cosmological model, where the present expansion is driven by ordinary, non-relativistic matter such as the stuff that makes up galaxies and stars, would have an age of $(2/3)H_0^{-1}$, $\mathrm{kms^{-1}Mpc^{-1}}$, so would be incompatible with H_0 any bigger than $55\mathrm{kms^{-1}Mpc^{-1}}$. Until a few years ago this simple model was most theoreticians' favourite (mine too!) and Freedman's value of 72 seemed therefore a rather disconcertingly large value.

However about the same time observations of distant supernovae (from the Supernovae Cosmology Project and the High-redshift Supernovae Team) seemed to show that the cosmos wasn't as simple as we had imagined it should be. By plotting the apparent luminosity of type IA supernovae versus their redshift, they presented evidence that the present expansion of the Universe is accelerating. The gravity of ordinary matter should slow the expansion of the universe, so instead we require some exotic form of dark energy to drive the present expansion of universe. While the theorists can dream up possible candidates, none so far seem very natural (to me). But only an accelerating universe can have an age greater than H_0^{-1}.

1.1.1 Large-scale structure

Nowadays the redshifts of hundreds of thousands of galaxies are known (not to me personally, of course). Combined with independent distance estimators this provides not only a strong confirmation of Hubble's law but also of theories of how this structure has formed. Indeed we have so much good information about the positions (on the sky) and redshifts of galaxies that even without independent estimates of their distance (which are notoriously unreliable) there is a huge industry now busy using redshift surveys to test cosmological theories.

The failure of a galaxy's observed velocity to obey this relationship exactly defines the "peculiar velocity" relative to the Hubble flow

$$v_{obs} = H_0 r + v_{pec} \,. \tag{4}$$

But if you look far enough out ($H_0 r \gg v_{pec}$) this peculiar motion should become a small perturbation. The peculiar velocities are supposed to be due to the gravitational interactions. If the "redshift-space" locations give you a rough estimate of the distribution of mass then this should predict the peculiar velocities. If you have a rough estimate of the peculiar velocities you can construct a rough picture of the real-space distribution, with which you can estimate the peculiar velocities, and so on. Major surveys are currently underway (notably the 2-degree field, 2dF, and Sloan Digital Sky Survey, or SDSS) which will measure over a million galaxy redshifts. The sheer weight of data will turn what sounds like a rather shaky procedure into a formidable test of our models.

Combined with data on the primordial perturbation spectra seen as anisotropies in the cosmic microwave background (see below) the formation of structure in our Universe via gravitational instability is emerging as a new pillar of the Hot Big Bang.

Models of structure formation also seem to work best with some exotic form of dark energy. So dark energy, as well as dark matter, has become widely accepted as part of the standard cosmological model in recent years.

1.2 The cosmic microwave background.

Just because our universe is expanding doesn't mean there had to be a big bang. Throughout the 1950's the Steady State theory was a perfectly viable alternative explanation for the expanding universe. It proposed that matter was continually being created so the overall state of the universe was uniform in time as well as space.

The primary observational reason for believing the universe did emerge from a hot big bang is the presence of the cosmic microwave background (CMB). Today these microwave photons are part of a blackbody spectrum of radiation corresponding to a temperature of 2.735 Kelvin. But we know that radiation is redshifted by the expansion of the universe and so temperature is inversely proportional to the size of the universe.

$$\frac{T}{T_0} = \frac{a_0}{a} \tag{5}$$

So the universe must become exceedingly hot at early times!

The CMB was discovered by Penzias and Wilson in 1965 and since then there has been a great deal of effort to observe deviations from isotropy or from a perfect blackbody spectrum. Apart from a small dipole component the radiation is observed to be fantastically isotropic - about 1 part in 10^5. It was only in 1992 that the first evidence for intrinsic anisotropies were discovered by the COsmic Background Explorer (COBE) satellite. This lead to much media hype, but also a new era in observational cosmology. There are now a large number of balloon and ground-based experiments that have confirmed COBE's measurements and extended them to smaller angular scales. These observations will give us a map, or at least a statistical picture, of the way the universe was when it was only a few hundred thousand years old.

The specific features of this map provide key tests of different cosmological parameters, but the general picture of an isotropic early universe very close to thermal equilibrium is the basis of all our ideas about the early universe.

1.3 Primordial nucleosynthesis.

About the same time the microwave background was first discovered, theorists were beginning to take seriously another prediction of the hot big bang model. If temperatures really were so high in the early universe there must have been an epoch at which nuclear reactions occurred. A hot plasma of electrons and protons, photons and neutrons and neutrinos, all interacting and cooling as the universe expanded must make a definite prediction of some primordial abundance of the elements. In practice it might be rather complicated to calculate, but in principle at least there must be well-defined answer. Later in the course we will tackle some of the issues involved.

In fact the early universe turns out only to be very successful at producing nuclei of the lightest elements: helium and lithium, as well as, of course, hydrogen. The heavier elements are only produced later in stars. By studying the abundance of the light elements and their different isotopes, and plotting this against the abundance of heavier elements one can extrapolate back to a primordial value (where there were no heavy elements). The relative abundances of the light elements are sensitive to the cosmology only one second after the big bang. Changing the rate of cosmological expansion or the relative number photons to protons, say, affects these yields.

In particular nucleosynthesis places a bound on the density of baryons (protons and neutrons) relative to the microwave background. Because the photon density today can be easily obtained from the temperature of the microwave background, this can be interpreted as a measurement of the present baryon density:

$$1 \times 10^{-31} < \rho_B < 4 \times 10^{-31} \text{ g cm}^{-3} \tag{6}$$

This is a good example of how models of the early universe can offer potentially very precise information about a number that is very hard to obtain directly today. On the other hand there is a lot of careful modelling that has to be applied before we get there.

2 Friedmann-Robertson-Walker cosmology

2.1 Geometry

The infinitesimal displacement between neighbouring points in the an isotropic and spatially homogeneous spacetime in D-dimensions can be written as

$$ds^2 = -c^2 dt^2 + a^2(t) d\Omega^2_{D-1,K} \,, \tag{7}$$

where t is the proper (cosmic) time at fixed spatial coordinates and $d\Omega^2_{D-1,K}$ describes the spatial displacement in a maximally symmetric $(D-1)$-dimensional space of constant curvature K.

The $(D-1)$-dimensional spatial hypersurfaces at fixed cosmic time t have time-dependent scale described by the scale factor, $a(t)$, and constant curvature K:

- S^{D-1} spherical space for $K > 0$: compact
- H^{D-1} hyperbolic space for $K < 0$: compact or non-compact
- R^{D-1} flat space for $K = 0$: compact or non-compact

In maximally symmetric 2-dimensional space we can write

$$d\Omega^2_{2,K} = d\theta^2 + \sin^2_K(\theta) d\phi^2 \,, \tag{8}$$

where

$$\sin_K(\theta) = \begin{cases} \frac{1}{\sqrt{K}} \sin\left(\sqrt{K}\theta\right) & \text{for } K > 0 \\ \frac{1}{\sqrt{-K}} \sinh\left(\sqrt{-K}\theta\right) & \text{for } K < 0 \\ \theta & \text{for } K = 0 \end{cases} \,. \tag{9}$$

In maximally symmetric 3-dimensional space we can write

$$d\Omega^2_{3,K} = d\chi^2 + \sin^2_K(\chi) d\Omega^2_{2,+1} \,. \tag{10}$$

Sometimes it is more useful to work in terms of the angular diameter distance $r \equiv \sin_K(\chi)$ as the radial coordinate, which gives the alternative form

$$d\Omega^2_{3,K} = \frac{dr^2}{1 - Kr^2} + r^2 d\Omega^2_{2,+1} \,. \tag{11}$$

Only in flat space, $K = 0$, are χ and r the same thing. Still more generally one can write

$$d\Omega^2_{3,K} = \gamma_{ab} dx^a dx^b \tag{12}$$

where γ_{ab} is the metric on the maximally symmetric 3-dimensional space[1] .

[1] I will use latin indices $a, b, ...$ to run from 1 to 3, greek indices $\mu, \nu, ..$ run from 0 to 3 and repeated indices should be summed.

In a four-dimensional spacetime, where at fixed time t the 3D spatial hypersurfaces are homogeneous and isotropic, the line element is given by

$$ds^2 = g_{\mu\nu} x^\mu x^\nu = -c^2 dt^2 + a^2(t) d\Omega^2_{3,K} \,, \tag{13}$$

and I will refer to this as an FRW (Friedmann-Robertson-Walker) spacetime. (Some include Lemaitre to make it FLRW.)

Observers following worldlines with fixed spatial coordinates see an isotropic expansion, $\Theta = 3H$, where $H \equiv \dot{a}/a$ is the Hubble rate. (The Hubble constant, H_0, is the present value of this variable in our Universe.)

2.2 Conformal time

It is often useful to extract the scale factor as an overall (conformal) scaling to give

$$ds^2 = a^2(\tau) \left[-c^2 d\tau^2 + d\Omega^2_{3,K} \right] \,, \tag{14}$$

where τ is the conformal time

$$\tau \equiv \int \frac{dt}{at} \,. \tag{15}$$

Note that the conformal time is a measure of the coordinate distance travelled by anything moving at the speed of light, $ds^2 = 0$.

For $K = 0$ we have $g_{\mu\nu} = a^2 \eta_{\mu\nu}$, where $\eta_{\mu\nu}$ is the metric for flat Minkowski spacetime, so the metric is said to be *conformally flat*.

For $K = 0$ and a =constant the spacetime *is* flat Minkowski spacetime.

2.3 Einstein equations

The simplest model for gravity consistent with all experiments is still Einstein's Theory of Relativity. Spacetime is curved by matter according to Einstein's equations:

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi G_N T_{\mu\nu} - \Lambda g_{\mu\nu} \,, \tag{16}$$

where $G_{\mu\nu}$ is the Einstein tensor, G_N Newton's constant and $T_{\mu\nu}$ the energy-momentum tensor for matter. I have included the cosmological constant Λ which is all the rage these days.

For the FRW metric (13) this yields two field equations: the evolution equation

$$\frac{\ddot{a}}{a} = -\frac{4\pi G_N}{3} \left(\rho + \frac{3P}{c^2} \right) + \frac{\Lambda}{3} \,, \tag{17}$$

and the Friedmann constraint equation

$$H^2 \equiv \left(\frac{\dot{a}}{a} \right)^2 = \frac{8\pi G_N}{3} \rho - \frac{Kc^2}{a^2} + \frac{\Lambda}{3} \,, \tag{18}$$

where ρ is the matter density and P the pressure.

The Einstein tensor is isotropic for a FRW metric so the matter pressure must also be isotropic for consistency, with energy momentum tensor $T^\mu_\nu = \text{diag}(-\rho, P, P, P)$. Local energy-momentum conservation

$$\nabla^\nu T^\mu_\nu = 0\,, \tag{19}$$

follows from Einstein's equations plus the Bianchi identities in curved spacetime which require $\nabla^\nu G^\mu_\nu = 0$. But one expects local conservation of energy-momentum to hold in any relativistic theory of gravity (i.e., one without a preferred frame). In the FRW spacetime this yields only one equation:

$$\dot\rho + 3H(\rho + \frac{P}{c^2}) = 0\,, \tag{20}$$

i.e., local energy decrease due to dilution plus work done by expansion. The local momentum vanishes due to isotropy.

Any vacuum energy $\Lambda \sim 8\pi G_N \rho_{c0}$ remains undiluted by the cosmological expansion so would have had a negligible effect in the early universe. It only begins to affect the cosmological dynamics close to present time. Even if we find a satisfactory explanation for the smallness of the cosmological constant, there remains the 'cosmic coincidence' problem of why we happen to be observing the universe just around the transition from matter to vacuum dominated epochs.

We seem to live in a surprisingly complicated cosmos where many different forms of matter - baryons, cold(?) dark matter, photons, neutrinos, and dark (vacuum?) energy contribute to Ω_{tot}.

2.4 Cosmological matter

Matter in the universe includes:

- Relativistic particles travelling at (or close to) the speed of light with respect to comoving observers at fixed spatial coordinates in the FRW spacetime. This is usually referred to as *radiation*, such as the photons of the cosmic microwave background, and (probably) analogous backgrounds of light neutrinos and gravitons (as yet undetected). An isotropic distribution of relativistic particles exerts an average pressure $P = \rho c^2/3$ in all directions.

- Non-relativistic matter travelling at much less than the speed of light, such as galaxies, stars, and most baryons (protons and neutrons) and (probably) some form(s) of weakly interacting massive particles (WIMPs), usually referred to as cold dark matter (CDM). Non-relativistic matter exerts negligible pressure and is often referred to as *dust*, or sometimes just as 'matter'.

- Vacuum energy (probably) exerting a negative effective pressure and causing the presently observed expansion of the universe. Could be described by self-interacting scalar fields, or just a cosmological constant, Λ.

All of these can be modelled as a barotropic fluid with equation of state $P = (\gamma - 1)\rho c^2$ with constant barotropic index, γ. In this case the continuity equation (20) can be integrated to give

$$\rho = \rho_0 \left(\frac{a_0}{a}\right)^{3\gamma} . \tag{21}$$

For $P = -\rho c^2$ and $\gamma = 0$ we have $\rho = \rho_0$ =constant and this kind of 'vacuum energy', undiluted by the expansion of the universe, is equivalent to a cosmological constant $\Lambda = 8\pi G_N \rho_0$ appearing in the Einstein equations.

The contribution of each fluid density ρ_i to the Hubble expansion is often written in terms of a dimensionless density parameter:

$$\Omega_i \equiv \frac{8\pi G_N \rho_i}{3H^2} . \tag{22}$$

The Friedmann equation (18) then becomes the dimensionless constraint equation

$$\Omega_{\rm tot} \equiv \sum_i \Omega_i = 1 + \frac{Kc^2}{a^2 H^2} . \tag{23}$$

If the universe was spatially flat ($K = 0$ and $\Omega_{\rm tot} = 1$), the observed expansion rate would correspond to a *critical density*

$$\rho_c \equiv \frac{3H^2}{8\pi G} . \tag{24}$$

For the present-day value of the Hubble constant this yields

$$\rho_{c0} = 1.9 \times 10^{-29}\ h^2 \text{g cm}^{-3} \tag{25}$$

where h was defined in equation (2).

The present density of non-relativistic matter in the universe is observed to lie somewhere between $10^{-32}h$ and $2 \times 10^{-29} h^2 \text{gcm}^{-3}$. (See for instance Padmanabhan, Chapter 11.) This corresponds to $0.001h^{-1} < \Omega_0 \leq 1$. The lower end of this range represents the mass of luminous matter seen in galaxies, i.e. stars, principally baryonic matter (atomic nuclei) like our Sun. However dynamical studies of galaxies and clusters of galaxies suggest that the real mass of matter in galaxies is probably greater than this by a factor of at least ten, requiring some form of *dark matter*. The largest estimates of the density of matter come from studies of the motions of galaxies on the very largest scales we can observe.

Whatever the actual value it is certainly larger than the energy density of photons which can be calculated precisely from the observed black-body spectrum of the microwave background radiation [see equation (35), coming up soon!]

$$\rho_\gamma = 4.8 \times 10^{-34}\ \text{g cm}^{-3} , \tag{26}$$

which corresponds to

$$\Omega_{\gamma 0} = 2.6 \times 10^{-5} h^{-2} . \tag{27}$$

However remembering that the density of a relativistic fluid like photons decreases as a^{-4} as the universe expands while dust decreases as a^{-3} [equation (21)] the two densities must have been equal at some time in the past t_{eq} where,

$$\left(\frac{a_0}{a_{eq}}\right) = \frac{\Omega_0}{\Omega_{\gamma 0}} \simeq 3.9 \times 10^4 \; \Omega_0 h^2 \; . \tag{28}$$

At earlier times it must have been relativistic rather than non-relativistic matter that dominated the energy density of the universe. Similarly any curvature term in the constraint equation decays only as a^{-2} so must also have become negligible so that Ω_γ was very close to one at that time. This is the basis of the hot big bang.

3 The Hot Big Bang

3.1 Temperatures and times

Let us consider in more detail how a relativistic fluid, and specifically photons, evolve in an expanding universe.

A photon gas in thermal equilibrium has a black-body or Planck spectrum, which means that the number of photons per unit volume in the range energy range E to $E + dE$ is given by

$$dn = \frac{1}{\pi^2 \hbar^3 c^3} \, \frac{E^2}{\exp(E/k_B T) - 1} \, dE \tag{29}$$

where k_B is the Boltzmann constant and T is the temperature of the distribution.

Note that as the universe expands the energy of the photons, $E = h\nu = hc/\lambda$, is redshifted to a lower energy

$$E' = \frac{a}{a'} \, E \tag{30}$$

The number of photons in the same coordinate volume now lies in the energy range from E' to $E' + dE'$ and

$$dn = \frac{1}{\pi^2 \hbar^3 c^3} \left(\frac{a'}{a} \right)^3 \frac{E'^2}{\exp(a'E'/ak_B T) - 1} \, dE' \tag{31}$$

However the energy density per unit physical volume is decreased by the expansion $dn' = (a/a')^3 dn$, so we have

$$dn' = \frac{1}{\pi^2 \hbar^3 c^3} \, \frac{E'^2}{\exp(a'E'/ak_B T) - 1} \, dE' \; . \tag{32}$$

which is still a black-body spectrum with a new temperature

$$T' = \frac{a}{a'} \, T \; . \tag{33}$$

Thus once a black-body spectrum has been established by thermal equilibrium it remains a black-body spectrum even if it is no longer in thermal equilibrium, and its temperature is inversely proportional to the scale factor. Hence the microwave background we see today still has an almost perfect black-body spectrum even though the photons have a temperature of only 2.735K and have long since decoupled from the other matter in the universe.

Integrating the black-body spectrum (29) over all energies gives the total number of photons per unit volume

$$n_\gamma = \int dn = \frac{2.4}{\pi^2} \left(\frac{k_B T}{\hbar c} \right)^3 \tag{34}$$

and the density per unit volume

$$\rho_\gamma = \int \frac{E}{c^2} dn = \frac{\pi^2}{15 \hbar^3 c^5} (k_B T)^4 \; . \tag{35}$$

At sufficiently high temperatures ($k_B T \gg mc^2$, the rest mass energy of the particle) we expect all types of elementary particles to become relativistic. If they interact with the photons they will share the same equilibrium temperature and so all types of relativistic particles will have a density proportional to T^4. We can write the density as

$$\rho = g_{\text{eff}} \frac{\pi^2}{30} \frac{(k_B T)^4}{\hbar^3 c^5} \tag{36}$$

where g_{eff} is a sum over the effective number of degrees of freedom of particles. g is 2 for photons alone, corresponding to the two different polarizations. All other relativistic bosons (with integer spin) counts one per degree of freedom, while for fermions (half-integer spins, e.g., neutrinos) it is 7/8 per degree of freedom due to their Fermi statistics rather than Bose-Einstein statistics [see Eq.(47)].

Remember, however that the energy density in a $K = 0$ radiation dominated model can also be written in terms of the expansion rate, and thus the time since the big bang:

$$\rho = \frac{3H^2}{8\pi G} = \frac{3}{32\pi G (t - t_*)^2} \tag{37}$$

Thus we have a relation between temperature and time

$$t - t_* = \frac{1}{2H} = \sqrt{\frac{3}{32\pi G} \frac{30\hbar^3 c^5}{g_{\text{eff}} \pi^2}} \frac{1}{(k_B T)^2} \, , \tag{38}$$

which can be written as

$$\frac{t - t_*}{1 \text{ sec}} \approx \frac{1}{\sqrt{g_{\text{eff}}}} \left(\frac{1 \text{ MeV}}{k_B T} \right)^2 \, . \tag{39}$$

Thus, for instance, the average energy of relativistic particles after 1 second was about the same as typical nuclear binding energies. Only below this temperature would we expect atomic nuclei not to be destroyed by energetic photons. The only uncertain parameter in this relation is g_{eff}, though it is fairly weakly dependent on the number of degrees of freedom. In the standard model of particle physics $g_{\text{eff}} = 106.75$ at high temperatures, but extensions to the standard model (e.g., supersymmetric theories) include more degrees of freedom.

Notice that we can also write down the entropy density of a relativistic fluid

$$s = \frac{\rho c^2 + p}{T} = \frac{2\pi^2 g_{\text{eff}}}{45} \frac{k_B}{\hbar^3 c^3} (k_B T)^3 \tag{40}$$

which for photons is simply proportional to the number of particles, $s_\gamma \simeq 3.6 k_B n_\gamma$.

This relation between temperature, time and density forms the basis of our models of the early universe.

3.2 Particle survival

In a static universe ($\dot{a} = 0$) we would expect all interactions between particles to produce a thermal equilibrium distribution as $t \to \infty$. However in an expanding universe we can only expect thermal equilibrium to be achieved if the rate of interactions, Γ, is greater than the rate of expansion, H. Typically we find that particles are in equilibrium at high temperatures and hence early times, dropping out of equilibrium as the temperature drops. How and when species drop out of equilibrium determines the abundance at late cosmic times of massive particles.

We will assume that we can treat the particles as an ideal gas which means that the forces between particles are short range and their mean free paths between interactions, $\langle\sigma v\rangle$, are much larger than the inter-particle spacing, $n^{-1/3}$. This proves to be a good approximation for most of the history of the universe. The interaction rate

$$\Gamma = n \langle\sigma v\rangle , \tag{41}$$

where v is the velocity and σ is the interaction cross-section which will depend on the type of interaction and the centre-of-mass energy too.

3.2.1 Relativistic neutrinos

Neutrinos are elusive particles that only interact through the weak interaction. As a result their masses are not known. Indeed only very recently has evidence been found that they indeed have any mass at all. Neutrinos come in three types associated with the three families of leptons; electron, muon or tau. Certainly the electron neutrino is very light with a rest-mass energy less than 13eV. They have no electric charge and so don't interact with photons.

At temperatures below about 10^{12}K the only relativistic particle species are photons, neutrinos and electrons and positrons [whose rest mass-energies $m_e c^2 \simeq 0.5$MeV$\simeq k_B(6\times 10^9\text{K})$]. Thus the only interactions keeping the neutrinos and anti-neutrinos in equilibrium are

1. *scattering:* $e^-\nu_e \to e^-\nu_e$, $e^+\nu_e \to e^+\nu_e$, etc;
2. *creation:* $e^-e^+ \to \nu_e\bar{\nu}_e$;
3. *annihilation:* $\nu_e\bar{\nu}_e \to e^-e^+$.
4. *inter-conversion:* $\nu_e\bar{\nu}_e \to \nu_\mu\bar{\nu}_\mu$, etc.

The cross-sections for all these weak interactions are of order $\hbar^2c^2G_F^2E^2$, where $G_F \simeq 1.2\times10^{-5}\text{GeV}^{-2}$ is the Fermi constant, and E is the particles' centre-of-mass energy. The interactions are mediated by the W or Z bosons, which are massive particles ($m_Wc^2 \sim m_Zc^2 \sim 90$GeV) and the interaction strength $G_F \sim (m_Wc^2)^{-2}$ for $E < m_W/2$.

Using the equilibrium distribution given in equation (29) and the temperature time relation in equation (38), we find

$$\frac{\Gamma}{H} \simeq \left(\frac{T}{1.6\times10^{10}\text{ K}}\right)^3 , \tag{42}$$

for neutrino interactions. Thus the neutrinos decouple from the photon background at a temperature $T_{\nu d} \simeq 1.6\times10^{10}$K equivalent to an energy scale of 1.4MeV.

If neutrinos are massless, they should maintain their thermal equilibrium spectrum as the universe expands with a temperature

$$T_{\nu 0} = \frac{a_{\nu d}}{a_{\nu 0}} T_{\nu d} \ , \tag{43}$$

as shown earlier for a non-interacting photon gas. If the photon background remained decoupled too then the neutrinos and photons should have the same temperature today. However the photons are still coupled to the electrons and positrons until the temperature falls below 6×10^9K at which point the electrons and positrons annihilate without being regenerated, as $k_B T$ falls below their rest-mass energy. All the entropy density that was stored in the combined photons-electron-positron plasma where $g_{\rm eff} = 2 + (7/8) \times 2 \times 2 = 11/2$ is dumped into the photons alone, $g = 2$, increasing the temperature by a factor of $(11/4)^{1/3}$. Thus the neutrino background today would have a temperature

$$T_{\nu 0} = \left(\frac{4}{11} \right)^{1/3} T_{\gamma 0} = 1.9\text{K} \ . \tag{44}$$

Even if neutrinos are not exactly massless this temperature still determines the number density of each family of light neutrinos in the universe today, as the total number in a co-moving volume will be conserved after decoupling

$$n_{\nu 0} = \frac{3}{4} \times \frac{2.4}{\pi^2} \left(\frac{k_B T_{\nu 0}}{\hbar c} \right)^3 \simeq 108 \ \text{cm}^{-3} \ . \tag{45}$$

(The factor 3/4 occurs because neutrinos are fermions, not bosons.) If we could detect the presence of such a cosmic background of neutrinos and even measure its temperature this would be an unparalleled test of the hot big bang scenario. But neutrinos are so weakly interacting such a direct detection may never be possible.

However, the neutrino background could still have important cosmological consequences. If they are massless their energy density today would be very similar to that of the microwave background photons. But even if neutrinos have only a very small mass, $m_\nu c^2 > 10^{-4}$eV, the number density calculated above implies a mass density, most usefully expressed as a fraction of the critical density:

$$\Omega_{\nu 0} = \left(\frac{m_\nu c^2}{90 \ \text{eV}} \right) \ h^{-2} \ . \tag{46}$$

If any, or all, of the neutrino families have a masses of a few eV, they would have a density close to the critical density. Such a mass is tantalizingly close to the current upper limit on the mass of the electron neutrino. On the other hand if a neutrino was found to have a mass of, say, a few keV this would be a cosmological disaster incompatible with our current models.

Our calculation does rely on the neutrinos still being relativistic at the time they decouple, so this places an upper limit on the mass $m_\nu c^2 < 1$MeV for which the result holds. Such matter is referred to as *hot dark matter*, and neutrinos are the most obvious candidate for such a type of dark matter. Note that it need not be relativistic today. In the next section we will see how non-relativistic matter decouples.

3.2.2 Non-relativistic particles

Massive particles interacting with the photon background at a temperature T have a thermal equilibrium distribution

$$\frac{dn}{dE} = \frac{1}{c^3\hbar^3\pi^2} \frac{E\sqrt{E^2 - m_0^2}}{\exp(E/k_BT) \mp 1} , \tag{47}$$

where the plus or minus sign corresponds to bosons or fermions. For non-relativistic particles, $k_BT \ll m_0c^2$, the number density of massive particles becomes exponentially suppressed and this gives a Boltzmann distribution (for bosons or fermions)

$$n = \frac{g}{\hbar^3} \left(\frac{m_0 k_B T}{2\pi} \right)^{3/2} \exp\left(-\frac{m_0c^2}{k_BT} \right) . \tag{48}$$

The relic density of particles today should be given by $n_0 = (a_d/a_0)^3 n_d$ where n_d is the number density when annihilation effectively ceases.

For massive neutrinos (1MeV$< m_\nu c^2 <$ 45GeV), the annihilation cross-section $\langle \sigma v \rangle \simeq \hbar^2 c^3 G_F^2 E^2 \simeq \hbar^2 c^7 G_F^2 m_\nu^2$ if they are non-relativistic. Thus we can find the temperature at which the annihilation rate for the above Boltzmann distribution equals the Hubble expansion as given in equation (38). We find

$$k_B T_d \simeq 70 \left(\frac{m_\nu c^2}{1 \text{ GeV}} \right) \text{MeV} . \tag{49}$$

Inserting this into equation (48) yields a number density at decoupling and thus a present day density which can be written as

$$\Omega_\nu \sim \left(\frac{1 \text{ GeV}}{m_\nu c^2} \right)^2 h^{-2} \tag{50}$$

This yields a lower limit on the allowed mass for a heavy neutrino species. In fact for neutrino masses above 90GeV$= m_Z c^2/2$ the interaction cross-section changes and we must do yet another calculation of the present day relic density which yields

$$\Omega_\nu \sim \left(\frac{m_\nu c^2}{1 \text{ TeV}} \right)^2 h^{-2} . \tag{51}$$

The combined bounds on neutrino masses are shown in figure 3.1 of Padmanabhan, p.99.

As we consider different types of particle species with different masses, and there are no shortage of candidates in models beyond the standard model of particle physics, we find a range of different constraints cosmology places upon the allowed masses. Massive particles that are non-relativistic when they decouple are generally referred to as *cold dark matter*. Those that are relativistic, like light neutrinos, are called *hot dark matter*. Substantial density of either type have important implications for models of how structure forms in the universe as we shall see later.

3.2.3 Baryon asymmetry

Having studied the relic abundances of neutrino or other more exotic particles, it is reasonable to ask whether the standard hot big bang yields sensible numbers of more easily detectable matter such as protons and neutrons (baryons) or electrons. Unfortunately the simple answer to this question is no, it doesn't.

Because the annihilation rate for protons and neutrons remains greater than the Hubble expansion rate until a temperature $k_B T \approx 20$MeV, while their rest-mass energy is ≈ 1GeV, their number density is exponentially suppressed. The number escaping annihilation in a symmetric baryon–anti-baryon universe would correspond to only one baryon for every 10^{19} photons. Instead the observed photon to baryon ratio is observed to be about 10^9. In any case there is no evidence for the equal numbers of anti-baryons as the symmetric model would predict. The only way we can explain this is to assume that there was a pre-existing asymmetry when baryons and anti-baryons annihilated leaving the observed number density of baryons (and effectively no anti-baryons).

In the standard model of particle physics baryon number is always conserved in (perturbative) particle interactions so there is no way we could create such an asymmetry. On the other hand we also have to produce an identical asymmetry between electrons and positrons so one would hope to find interactions that violate baryon and lepton number together in some Grand Unified Theory. This is one of the strongest cosmological reasons for seeking such a theory.

Many years ago Sakharov set down a number of conditions that must be satisfied for *baryogenesis* to occur in the early universe:

1. baryon number non-conservation (obviously!);

2. CP violation (where C is charge symmetry and P parity) which implies time asymmetry;

3. non-equilibrium.

Recently there has been considerable interest in whether baryogenesis might be possible within the standard model of electroweak interactions. There are non-perturbative processes that violate baryon and lepton number, although these are strongly suppressed at low energies. Near the electroweak transition these are not suppressed and a first-order phase transition could also provide the required non-equilibrium conditions. However the consensus seems to be that the phase transition is not first-order and CP violation is too small to produced the required asymmetry.

3.3 Primordial Nucleosynthesis

Having failed to understand the origin of the observed number of baryons in the universe, should we really believe that we know enough about the early universe to describe the abundances of other particles far harder to detect? If the microwave background photons were the only other relic particles we could observe, we might begin to wonder whether they really had some other origin.

However we do have more information. We can observe not only the total number of baryons, but also the relative number of different atomic nuclei. We know from the work of Hoyle and others that the heavier elements (all called "metals" by astronomers) can be synthesised in the interior of stars. But there is no way to produce sufficient quantities of helium, which is known to account for about 25% of nuclear matter (by mass). In fact Hoyle's work on the synthesis of the elements in stars complements the results of the hot big bang, as Fred Hoyle himself discovered in a paper in 1964 with Roger Tayler. The early universe turns out to be an efficient manufacturer of only the lightest elements: hydrogen, helium and lithium, and the isotopes deuterium and helium-3.

If we go back to the era when neutrinos decoupled, $k_B T \sim 1\text{MeV}$ or $t \sim 1\text{s}$, we find the non-relativistic baryons are kept in statistical equilibrium with one another by weak interactions:

1. scattering: $n + e^+ \rightarrow p + \bar{\nu}_e$, $p + e^- \rightarrow n + \nu_e$;

2. neutron decay: $n \rightarrow p + e^- + \bar{\nu}_e$.

and the relative number of neutrons to protons is then just given by the Boltzmann factor

$$\frac{n_n}{n_p} \simeq \exp\left(-\frac{Q}{k_B T}\right) \tag{52}$$

where $Q = (m_n - m_p)c^2 = 1.3\text{MeV}$ is the rest-mass energy difference between the neutron and proton.

At high temperatures ($k_B T \gg Q$, $m_e c^2$) the scattering rates are of order $\hbar^{-1} G_F^2 (k_B T)^5$, as in the case of neutrino scattering discussed earlier. As this falls with temperature, the interaction rate for protons becomes less than the expansion rate at a temperature $k_B T \simeq 0.8\text{MeV}$. Thereafter the only important interaction is the neutron decay which has a half-life $\tau = \Gamma^{-1}$ where $\Gamma \simeq 0.05\hbar^{-1} G_F^2 (m_e c^2)^5$. We will use this as the basis of our estimate of nuclear abundances. The neutron to proton ratio is given by the equilibrium abundance at $k_B T \simeq 0.8\text{MeV}$, which gives $n_n/n_p \simeq 1/6$. The number of neutrons is then depleted by neutron decay.

What about the light elements? The binding energies for the nuclei of deuterium, helium-3 and helium-4 are 2.2MeV, 7.7MeV and 28MeV. Should these not already have formed stable nuclei? In fact the equilibrium abundances of the light nuclei remain very small at this stage because of the large number of photons relative to the baryons. Although the temperature of the photon distribution has fallen to $k_B T = 0.8\text{MeV}$, there is still a sufficient number of energetic photons to disintegrate the nuclei. Put another way, entropy favours free baryons and the entropy of the universe is very high.

Only when $k_B T$ falls to 0.1MeV does the deuterium become stable against dissociation. By this time the neutron to proton ratio has dropped to 1/7 due to neutron decay. However, once deuterium, the first building block, becomes stable, both helium-3 and helium-4 also form rapidly. Because helium-4 has much the greatest binding energy per nucleon essentially all the neutrons available at this time are processed into helium-4, and we can estimate the primordial ^{4}He abundance ratio (by mass) as

$$Y_P \simeq \left(\frac{4(n_n/2)}{n_n + n_p}\right)_{\text{nuc}} \tag{53}$$

where the subscript nuc denotes the time when $k_B T_D = 0.1\text{MeV}$ and deuterium becomes stable against dissociation. Hence we find $Y_P \simeq 0.25$, which should be compared with an observational value of $0.22 < Y_P < 0.25$.

Interactions between hydrogen and helium nuclei could also, in principle, form elements with atomic masses of 5 or 8, but there are no tightly bound nuclei with these masses so primordial nucleosynthesis effectively stops with helium. Small amounts of lithium-7 are produced, with a relative abundance of about 10^{-10}, but the density of helium is too small for the "triple-α" reaction ($^4\text{He}+^4\text{He}+^4\text{He}\to^{12}\text{C}$) which can occur in the interiors of some stars. A thorough calculation of the primordial abundances of the light elements requires a detailed numerical calculation using the complete nuclear reaction network with well determined reaction rates. The primordial abundances of ^{4}He, D, ^{3}He and ^{7}Li all provide key tests of our model of the early universe and all provide good support. It may not be so remarkable that we can find a model that predicts the primordial abundance of, say, helium, but it is remarkable that it can also yield the correct abundances of the other light elements which range over many orders of magnitude. So maybe there is some truth in the big bang model after all!

The relic abundances of deuterium and helium-3 are the abundances when the processes forming helium-4 eventually become inefficient. These reaction rates are dependent on the number of nucleons present, $\Gamma \propto n_B$, and so the relic abundances of deuterium and helium-3 are smaller if n_B is larger at a given temperature. The observed deuterium and helium-3 abundances require the ratio of the number baryons to photons to lie between

$$2 \times 10^{-10} < \frac{n_B}{n_\gamma} < 6 \times 10^{-10} \, , \tag{54}$$

which, assuming the number of baryons and photons are conserved up until the present, implies

$$0.007 < \Omega_{B0} h^2 < 0.02 \, . \tag{55}$$

This seems to be compatible with the observed density of matter today assuming some, but probably not all, the dark matter is baryonic.

The primordial abundance of helium on the other hand should be fairly insensitive to the baryon-to-photon ratio. Instead it depends on the neutron-proton ratio at temperature T_{nuc}. This in turn is a fixed fraction (due to neutron decay) of the ratio at the freeze-out temperature, T_d, when the neutrino scattering ceased to maintain the equilibrium, $\Gamma_d = H_d$. We have seen that in the radiation dominated era we have a simple relation between temperature and expansion rate given in equation (38), where the only unknown is the number of degrees of freedom, g_{eff}. This has been used to constrain the contribution to g_{eff} of light ($m_\nu c^2 < 1\text{MeV}$) neutrinos, which gives $N_\nu \leq 3$. Note that nucleosynthesis occurs at about the same time as the neutrinos decouple (see section 3.2.1) so "light" neutrinos usually means relativistic both at decoupling and nucleosynthesis. In practice both free theoretical parameters (n_B/n_γ and g_{eff}) should be fitted in a combined likelihood analysis constrained by all the different abundances.

More generally cosmologists take the success of primordial nucleosynthesis as good evidence in support of the validity of the standard Hot Big Bang cosmology at energies of order 1MeV and as a strong constraint on possible deviations from the standard radiation dominated FRW models at this time. It is used to place constraints on any change in the expansion rate due to exotic matter or non-standard gravity during nucleosynthesis.

3.4 The Microwave Background

After all the excitement of nucleosynthesis, we have a universe where the only relativistic particles are the photons and light neutrinos. The latter have decoupled but the photons are still coupled to the charged particles: electrons and ions (principally protons, although one should really also consider helium ions in a more careful treatment).

3.4.1 Spectral distortions

There are various processes responsible for coupling the photons to the electrons:

1. *Thomson scattering*: in the limit that the energy of a photon is much less than the electron rest-mass ($k_B T \ll m_e c^2$) the photon scatters elastically and no energy is transfered.

2. *Compton scattering*: an electron absorbs and then re-emits a photon with different energy, $\delta E/E \simeq k_B T/m_e c^2$. Note that this does not change the photon number, which requires double-Compton scattering when two photons are emitted (or absorbed).

3. *Free-free emission/absorption*: as two electrons scatter off one another (by exchanging a photon) they may emit or absorb a photon.

As in the previous examples of particle interactions, we can establish the time at which each of these interaction rates become longer than the Hubble time. Estimates of the relevant times are calculated in Padmanabhan, for instance. Here I will just quote the basic results.

The particle interactions that change the total number of photons (free-free and double Compton scattering) become ineffective at a red-shift of order $z_{\rm th} \sim 10^7$. At earlier times any violent non-equilibrium processes that might distort the black-body photon spectrum can be thermalised. However below this redshift the photon number remains fixed although a *statistical equilibrium* (rather than thermal equilibrium) can still be established by Compton scattering. In statistical equilibrium the photon distribution is given by

$$dn = \frac{1}{\pi^2 \hbar^3 c^3} \, \frac{E^2}{\exp[(E/k_B T) + \mu] - 1} \, dE \tag{56}$$

where μ is the *chemical potential*, equal to zero for the pure black-body spectrum given in equation (29). Eventually the Compton scattering becomes unable to maintain even the statistical equilibrium at a redshift of about $z_{\rm st} \sim 10^5$

Any non-adiabatic release of energy between $z_{\rm th}$ and $z_{\rm st}$, such as the damping of pressure waves or more exotic effects including the decay of unstable particles or the evaporation of primordial black holes should produce distortions in the observed photon spectrum. The absence of any such deviation, constrained by the COBE satellite to be $\mu < 10^{-3}$, is an important constraint. It confirms the simple picture of a homogeneous, placid universe cooling adiabatically at these times.

3.4.2 Recombination and Decoupling

Note that Thomson scattering, the low temperature limit of Compton scattering, still remains effective ($t_{\rm T} \ll H^{-1}$) even below $z_{\rm st}$, though there is no significant momentum transfer in these interactions.

Eventually it becomes energetically favourable for the free electrons and ions to form neutral hydrogen atoms (which have a binding energy 13.6eV). As in the case of deuterium nuclei, the high entropy of the universe delays *recombination*, defined to be the time when the ionisation fraction for electrons x_e becomes 0.1, until the temperature has fallen to $k_B T_{\rm rec} \simeq 0.29$eV.

With the rapid drop in the number of charged particles, Thomson scattering of the photons rapidly becomes ineffective and photons finally decouple from non-relativistic matter at a $k_B T_{\rm dec} \simeq 0.26$eV. The photons of the microwave background that we observe today last scattered at this epoch. Numerical calculations show that the *last-scattering surface* lies at a redshift of $z \simeq 1070$, with a width $\delta z \simeq 80$. Our microwave background sky is a relic of the early universe, last scattered when the temperature was $T \simeq 3000$K, giving a view of the universe as it was about one hundred thousand years after the big bang.

3.5 Enigmas and Conundrums

We have seen that the hot big bang model based on established models of particle interaction in a homogeneous and isotropic universe obeying the evolution equations of Einstein's general relativity, provides a good description of many aspects of the observed universe. But at the same time there are a number of unanswered questions. We have seen for instance that although it can explain the abundances of the light elements it cannot explain where the baryons that make up these nuclei came from. Of course they could just be there, put in as an initial condition. As cosmologists became more confident in this basic picture through the '60s and '70s, they also became more worried about the peculiar initial conditions it seemed to require.

3.5.1 The smoothness problem

Just why is our universe so homogeneous? For instance the microwave background is sensitive to perturbations in the metric at the time of decoupling. Perturbations in the gravitational potential produce temperature anisotropies, yet the microwave sky is uniform to about one part in 10^5 (except for the dipole moment). On the other hand it can't be completely homogeneous or there would be no structures in the universe. Gravity tends to make matter clump together so should amplify any initial inhomogeneities. As we shall see later, small density perturbations in the dust dominated era grow proportional to the scale factor, so if the density perturbation $\delta\rho/\rho$ is only of order unity today as we observe on scales of about $8h^{-1}$Mpc, these must have grown from an initial density perturbation $\sim 10^{-4}$ at t_{eq}, see equation (28). Why such small initial inhomogeneities and where did they come from?

It was realised, too, that in order to avoid producing large metric perturbations on small scales (which would collapse to form black holes) and to avoid inhomogeneities on large scales (which would been seen on the microwave sky) a particular form of initial density perturbation spectrum, called the Harrison-Zel'dovich spectrum, was required in order to produce structure across a wide range of scales. The Harrison-Zel'dovich spectrum corresponds to a fixed gravitational potential $G\delta\rho \times L^2 \sim 10^{-4}$ on all scales L. This is often referred to as a *scale-invariant* spectrum. But how does the universe produce such a spectrum?

3.5.2 The horizon problem

The smoothness problem gets even worse if we think about the size of causally connected regions in the early universe. Starting from the big bang at $t = 0$ a light ray travels a coordinate distance equal to $\int_0^t cdt'/a(t')$ which at time t corresponds to a physical distance

$$d(t) = a(t) \int_0^t \frac{cdt'}{a(t')} = 2ct = cH^{-1} \tag{57}$$

in the radiation dominated era. This is called the horizon distance. At the time of nucleosynthesis, say, this corresponds to a physical distance of about 10^{12}cm. This region expanded up to the present corresponds to about 10^{21}cm which is only 1kpc.

Assuming that nothing travels faster than the speed of light there would seem to be no way that causally disconnected regions could have established homogeneity by that time, and yet there is no evidence for different primordial nuclear abundances in different parts of our galaxy, or even different galaxies. I have discussed thermal equilibrium implicitly assuming homogeneity but causally disconnected regions could not even know of one another's existence much less establish homogeneity. Similarly, if we estimate the size of causally connected regions on the microwave background sky at the time of decoupling, they are only about 2° across. The initial conditions must have been set up so that the universe was not only approximately homogeneous over causally disconnected regions, but also contained a Harrison-Zel'dovich spectrum of small but non-zero density perturbations.

3.5.3 The flatness problem

If curvature tends to dominate over any matter density at late times, how come we live in a universe where the matter density is still comparable to the curvature? Surely the curvature should have come to dominate ages ago. This is sometimes put forward as an argument for why Ω_0 must actually be 1 even if some observations suggest that maybe it is only 0.01 or 0.1. If it isn't exactly 1 why should it still be as close as 0.01 or 0.1? If the curvature had been comparable to the matter density in the early universe it would have dominated within a few Hubble times. A closed model would recollapse, an open model would expand so that the matter density rapidly became negligible.

Let's be a bit more precise. The constraint equation (18) can be written as

$$\begin{aligned} \Omega - 1 &= \frac{Kc^2}{\dot{a}^2} && (58) \\ &\propto a^2 \text{ for radiation dominated} && (59) \\ &\propto a \text{ for dust dominated} && (60) \end{aligned}$$

The evolution equation (17) shows that $\dot{a}$ decreases as the universe expands with either radiation or dust and $|\Omega - 1|$ must grow with time. Thus we have $\dot{\Omega} < 0$ for $\Omega < 1$ (and $\kappa < 0$) or $\dot{\Omega} > 0$ for $\Omega > 1$ (and $\kappa > 0$). Only if $\Omega = 1$ and $\kappa = 0$ does Ω stay constant. If we use the matter dominated solutions we see that if $|1 - \Omega_0|$ is 0.1 today, then $|1 - \Omega|$ would have to be $0.1 \times (T_0/T_{eq}) \times (T_{eq}/10^{10}\mathrm{K})^2 \approx 10^{-15}$ at the time of nucleosynthesis. Even if Ω is only 0.1 or 0.01 today we still require an initial value fantastically close to unity. Why do we have such precise fine-tuning?

3.5.4 The entropy problem

Another re-statement of the flatness problem is to ask why the universe is so big or so old. If we look for a dimensionless measure of how big the universe is, we might consider the total entropy of the observable universe. The entropy density today is related to the number density of microwave background photons (which make up most of the particles in the universe) by equation (40) which gives $s = 3.6 k_B n_\gamma$. They have a current density of approximately $400\mathrm{cm}^{-3}$. On the other hand the size of the the present horizon

$$d_0 = a_0 \int_0^{t_0} \frac{c dt'}{a(t')} = 3ct_0 = 2cH_0^{-1} \tag{61}$$

is approximately $6000h^{-1}$Mpc. Thus the number of photons, or equivalently the entropy of the observable universe is about 10^{88}.

3.6 Natural Units

At this point, having tried to dutifully carry through all the various fundamental constants that have cropped up in the equations, I want to get rid of them. It is standard practice to do this in cosmology by setting

$$\hbar = k_B = c = 1 \ . \tag{62}$$

This allows us to discuss lengths in terms of time, or temperatures in terms of energies. This may seem a bit disconcerting at first but you get used to it, and it makes the equations much simpler. After all, you are already used to the idea that you can measure distances in terms of time. When you say a star is so many light-years away you know to interpret the *years* as being the *distance* light would travel in that time. You have just set $c = 1$. Similarly I have already begun to talk about a temperature in terms not of Kelvin but in terms of $k_B T$ being so many MeV, or whatever. Henceforth I'll drop the k_B and just refer to temperatures in MeV. Similarly a mass in MeV just means mc^2, where $c = 1$.

All this leaves only one fundamental unit given in terms of the only fundamental constant with dimensions that I have left, G. The Planck mass:

$$M_P = \sqrt{\frac{\hbar c}{G}} = \frac{1}{\sqrt{G}} = 2.177 \times 10^{-5} \text{ g} . \tag{63}$$

The corresponding energy $M_P c^2 = 1.223 \times 10^{19}$GeV. The Planck temperature is, obviously enough,

$$T_P = \frac{M_P c^2}{k_B} = \frac{1}{\sqrt{G}} = 1.42 \times 10^{32} \text{ K} , \tag{64}$$

and the corresponding Compton wavelength gives the Planck length

$$L_P = \frac{\hbar}{M_P c^2} = \sqrt{G} = 1.616 \times 10^{-33} \text{ cm} , \tag{65}$$

or Planck time

$$t_P = \frac{L_P}{c} = \sqrt{G} = 5.390 \times 10^{-44} \text{ sec} . \tag{66}$$

The Planck scale is the scale at which quantum gravitational effects can no longer be ignored. For example, the uncertainty in the energy within a Planck length, $\Delta E \sim \hbar/L_P \sim M_P c^2$ is the same as the energy required to form a black hole with the same Schwarzschild radius, $M_P c^2 = L_P c^2/G$. To understand physics, or cosmology on the Planck length we would need to understand the physics of virtual black holes, i.e., we need a theory of quantum gravity.

4 Scalar fields

4.1 Introduction

Scalar fields have come to play a dominant role in many studies of the early universe, and will do so in the rest of this course. The basic reason is that a homogeneous scalar field, $\varphi(t)$, is immediately compatible with a homogeneous and isotropic universe. At the same time, scalar fields offer a novel form of matter which changes fundamentally the cosmological dynamics. In particular self-interacting scalar fields can exert a negative pressure, quite unlike what one would expect from 'ordinary' matter.

One should be aware that no fundamental scalar field is known to exist in nature. For instance, the electromagnetic field is described by a four-dimensional vector potential, as are the W and Z vector bosons of the electroweak theory. Nonetheless scalars are commonly used in effective descriptions of phenomena, e.g., order parameters in solid state physics. So bear in mind that even in the early universe a scalar field may be an effective description with limited range of validity, if only because of our ignorance of the fundamental nature of quantum gravity.

That said, there are many theoretical candidates for fundamental scalar fields of nature. Two of the most commonly invoked are:

- The Higgs boson: The vacuum expectation value of the Higgs field gives the W and Z bosons their mass after spontaneous symmetry breaking in the electroweak phase transition. Experimental particle physicists are confident that the Higgs particle will be discovered when the Large Hadron Collider starts operating at CERN.

- The dilaton: String theory, it is sometimes said, predicts gravity. In the language of particle physics, this means that a low energy effective description of the excitations of strings includes the graviton. However along with this tensor field comes a scalar field called the dilaton. In the low energy limit, this is a massless scalar field, with a non-minimal coupling to the spacetime curvature.

When cosmologists first started introducing scalar fields into their models of the early universe in the early 1980's, it was usually introduced as a Higgs-type field associated with symmetry breaking in one or other Grand Unified Theory uniting the strong theory with the electromagnetic and weak interactions. These fields typically possess a symmetric Mexican hat type self-interaction potential, $V(\varphi)$, so that at low temperatures/late times the field settles into the global minimum of the potential at $\varphi \neq 0$, where the underlying symmetry of the full potential is hidden.

Since then the number and variety of scalar fields considered, their couplings and interactions, has exploded. This is partly due to the use of supersymmetric theories where every fermionic (half-integer spin) state is related to bosonic (whole-integer spin) states, including spin-0 scalars. But is also due to the growing interest in string theory (and M-theory) as a fundamental theory uniting all the interactions including gravity in a higher-dimensional spacetime. In order to recover the four-dimensional world we see around us requires some process of dimensional reduction whereby the various shapes and sizes of the extra dimensions are typically described by scalar moduli fields in a four-dimensional effective field theory.

The plethora of possibilities have left cosmologists free to try almost any form they wish for the scalar field and its interactions.

4.2 Scalar field dynamics

The scalar field equations of motion can be derived from the canonical Lagrange density

$$\begin{aligned} \mathcal{L}_\varphi &= -\frac{1}{2}(\nabla\varphi)^2 - V(\varphi)\,, \\ &= -\frac{1}{2}g^{\mu\nu}\varphi_{,\mu}\varphi_{,\nu} - V(\varphi)\,. \end{aligned} \tag{67}$$

Equivalently we can integrate Eq. (67) over some region of spacetime to give the dimensionless action

$$S_\varphi = \int dV \mathcal{L}_\varphi = \int d^4x\sqrt{-g}\left[-\frac{1}{2}g^{\mu\nu}\varphi_{,\mu}\varphi_{,\nu} - V(\varphi)\right]\,. \tag{68}$$

The scalar field equation of motion is obtained by minimising the action with respect to variations $\varphi \to \varphi + \delta\varphi$. This yields the Euler-Lagrange equations

$$\frac{\partial}{\partial x^\mu}\left(\frac{\partial \mathcal{L}_\varphi}{\partial \varphi_{,\mu}}\right) = \frac{\partial \mathcal{L}_\varphi}{\partial \varphi}\,, \tag{69}$$

which, for the canonical scalar field Lagrangian (67), yields

$$\Box\varphi \equiv g^{\mu\nu}\nabla_\mu\nabla_\nu\varphi = V'\,. \tag{70}$$

where $V' \equiv dV/d\varphi$.

Equation (70) is sufficient to study the dynamics of a scalar field in a fixed background spacetime. But as cosmologists we will also be interested in the gravitational backreaction of the scalar field energy upon the cosmological expansion. Combining the scalar field action, given in Eq. (68), with the Einstein-Hilbert action of general relativity yields

$$S = \int d^4x\sqrt{-g}\left[\frac{1}{16\pi G}R - \frac{1}{2}g^{\mu\nu}\varphi_{,\mu}\varphi_{,\nu} - V(\varphi)\right]\,. \tag{71}$$

where R is the Ricci scalar curvature and G_N is Newton's gravitational constant.

Minimising the Einstein-Hilbert action with respect to variations in the metric tensor $g_{\mu\nu} \to g_{\mu\nu} + \delta g_{\mu\nu}$ yields the Einstein field equations:

$$G^{\mu\nu} \equiv R^{\mu\nu} - \frac{1}{2}g^{\mu\nu}R = 8\pi G T^{\mu\nu}\,. \tag{72}$$

The energy-momentum (or stress-energy) tensor for matter is obtained by varying the action with respect to the metric tensor.

$$T^{\mu\nu} \equiv \frac{2}{\sqrt{-g}}\frac{\partial}{\partial g_{\mu\nu}}\left(\sqrt{-g}\mathcal{L}\right)\,. \tag{73}$$

For a canonical scalar this gives

$$T^{\mu\nu} = \varphi_{,\mu}\varphi_{,\nu} - g_{\mu\nu}\left(\frac{1}{2}g^{\lambda\kappa}\varphi_{,\lambda}\varphi_{,\kappa} + V\right)\,. \tag{74}$$

4.2.1 Scalar fields in Minkowski spacetime

In flat (Minkowski) spacetime the d'Alembertian corresponds to the operator

$$\Box \equiv -\frac{\partial^2}{\partial t^2} + \sum_i \frac{\partial^2}{\partial x_i^2}\,, \tag{75}$$

so that the scalar field equation of motion is a wave equation, driven by the potential gradient:

$$\frac{\partial^2 \varphi}{\partial t^2} - \sum_i \frac{\partial^2 \varphi}{\partial x_i^2} = -V'\,. \tag{76}$$

For a homogeneous scalar field $\varphi = \varphi_0(t)$ the field equation (76) can be integrated for arbitrary $V(\varphi)$ to give

$$\frac{1}{2}\dot{\varphi}^2 + V(\varphi) = C\,. \tag{77}$$

where C is an arbitrary constant of integration. Thus the field obeys an energy-conservation equation (or Hamiltonian constraint) and one can interpret the dynamics of the field as simply being like that of a ball (with 'position', φ, and kinetic energy, $\dot{\varphi}^2/2$) rolling along a hillside (with a potential energy, $V(\varphi)$) maintaining a constant total energy (C).

To study the dynamics of arbitrary spatial field configurations in a 3-dimensional Euclidean space it is often useful to decompose the field into (linearly independent) Fourier modes:

$$\varphi(t, x^i) = \sum_{\mathbf{k}} \varphi_{\mathbf{k}}(t) \exp(i\mathbf{k}.\mathbf{x})\,, \tag{78}$$

where the each Fourier mode is an eigenmode of the spatial Laplacian

$$\sum_i \frac{\partial^2}{\partial x_i^2} \left[\varphi_{\mathbf{k}}(t) \exp(i\mathbf{k}.\mathbf{x})\right] = -k^2 \left[\varphi_{\mathbf{k}}(t) \exp(i\mathbf{k}.\mathbf{x})\right] \tag{79}$$

with eigenvalue $-k^2$, and wavelength $2\pi/|\mathbf{k}|$. (A homogeneous field, $\varphi_0(t)$, has only one Fourier mode with infinite wavelength.)

For a quadratic potential, $V = m^2\varphi^2/2+$constant, the field equation is linear, that is we can solve for the evolution of each Fourier mode separately:

$$\ddot{\varphi}_{\mathbf{k}} + [k^2 + m^2]\varphi_{\mathbf{k}} = 0\,, \tag{80}$$

This is the equation of motion for a simple harmonic oscillator with $\varphi_{\mathbf{k}} \propto \exp(\pm i\omega_{\mathbf{k}} t)$ where the angular frequency $\omega_{\mathbf{k}}^2 = k^2 + m^2$. Thus the general solution of the wave equation for a free scalar field is

$$\varphi(t, x^i) = \sum_{\mathbf{k}} \left[A_{\mathbf{k}} e^{-i\omega_{\mathbf{k}} t} + B_{\mathbf{k}} e^{+i\omega_{\mathbf{k}} t}\right] \exp(i\mathbf{k}.\mathbf{x})\,, \tag{81}$$

Thus Fourier modes of a massless field (with $m = 0$) all propagate with the speed of light, whereas for a massive field ($m^2 > 0$) only modes with $k^2 \gg m^2$ are relativistic. If $m^2 < 0$ the frequency of long-wavelength modes ($k^2 < -m^2$) becomes imaginary, i.e., there is an instability which allows the exponential growth of long-wavelength modes, and the field is said to be tachyonic.

Cubic and higher-order terms in the potential leads to couplings between the different Fourier modes. (In a quantum theory one would describe this as interactions between φ-particles.) The field equation is no longer a linear differential equation so one can no longer solve the Fourier mode evolution equations separately in order to construct the full solution and one needs to make some other simplifying assumptions to make any analytic progress.

4.2.2 Scalar fields in FRW spacetime

In an unperturbed FRW spacetime with scale factor $a(t)$ the scalar field equation of motion (70) becomes

$$\ddot{\varphi} + 3H\dot{\varphi} - \frac{1}{a^2}\gamma^{ij}\partial_i\partial_j\varphi + V' = 0\,. \tag{82}$$

where γ_{ij} is the metric on the comoving 3-dimensional space.

For a homogeneous scalar field $\varphi = \varphi_0(t)$ we have the ordinary (rather than partial) differential equation

$$\ddot{\varphi}_0 + 3H\dot{\varphi}_0 + V' = 0\,, \tag{83}$$

which again one can interpret as describing the acceleration ($\ddot{\varphi}$) of a ball (with 'position') rolling along a hillside (with a potential energy, $V(\varphi)$) but now subject to 'Hubble damping' proportional to its velocity. The characteristic time-scale on which the damping has an effect is the Hubble time, H^{-1}. The time-dependence of the scale factor means that there is no conserved total energy as was given in equation (77) in Minkowski spacetime.

In a spatially flat FRW universe arbitrary spatial dependence of the scalar field can be decomposed into Fourier modes as in Minkowski spacetime. But in curved space the spatial modes must be generalised to the appropriate harmonic functions which are eigenvalues of the corresponding spatial Laplacian. On S^3 these are the 3-dimensional analogues of the spherical harmonics on the surface of a 2-dimensional sphere. In H^3 they are the analytic continuation of the hyperspherical harmonics. Fortunately one can often do without the explicit form of the mode functions and just label them by their eigenvalue so that

$$\varphi(t,x^i) = \sum_{\mathbf{k}} \varphi_{\mathbf{k}}(t) Q_{\mathbf{k}}(\mathbf{x})\,, \tag{84}$$

where

$$\gamma^{ij}\partial_i\partial_j Q_{\mathbf{k}}(\mathbf{x}) = -k^2 Q_{\mathbf{k}}(\mathbf{x})\,. \tag{85}$$

Substituting this form for an inhomogeneous field into the equation of motion (82) for a non-self-interacting field (i.e., quadratic potential) yields independent equations for each mode:

$$\ddot{\varphi}_{\mathbf{k}} + 3H\dot{\varphi}_{\mathbf{k}} + \left[\frac{k^2}{a^2} + m^2\right]\varphi_{\mathbf{k}} = 0\,, \tag{86}$$

The Hubble expansion leads to a damping term, and the k^2/a^2 term, due to spatial gradients, decreases with increasing a.

If we change to conformal time, $\tau \equiv \int dt/a$, and introduce the conformally rescaled field, $u = a\varphi$, the field equation (86) for each mode becomes

$$u''_{\mathbf{k}} + \left(k^2 + m^2 a^2 - \frac{a''}{a}\right) u_{\mathbf{k}} = 0\,. \tag{87}$$

This is analogous to the field equation for a canonical scalar field in flat (Minkowski) spacetime but with a time-dependent mass

$$m^2 \to m^2 a^2 - \frac{a''}{a} \tag{88}$$

In particular, at very short wavelength $k^2 \gg |m^2 a^2 - (a''/a)|$ the evolution of $u_{\mathbf{k}}$ is the same as a massless field in flat spacetime. This allows us to employ results from quantum field theory in flat spacetime to fields in an expanding FRW universe on sufficiently short wavelengths.

4.3 Spontaneous Symmetry Breaking

Spontaneous symmetry breaking seems to give a good model for the electroweak interactions. While the massless photon mediates electro-magnetic forces, massive bosons (the W and Z) mediate the weak interactions. It is the mass of these intermediate particles that make the weak interaction short ranged and "weak". The Fermi coupling constant is $G_F \sim (m_W c^2)^{-2}$. However at high temperatures $\gg 10^{15}$K all the bosons become effectively massless and the full symmetry is restored.

The masses of the W and Z bosons are supposed to be due to their interaction with the famous, but as yet undiscovered, Higgs boson. This has a non-zero vacuum expectation value, ϕ, at low temperatures due to its self-interaction which leads to a potential energy density $V(\phi)$ with a minimum at $\phi \neq 0$. The masses of the vector bosons are of order $|\phi|$. But in a thermal background there are temperature-dependent corrections to the effective potential which include terms of order $T^2\phi^2$. Thus at high enough temperatures $\phi = 0$ can become a stable minimum where the bosons become massless.

There are many different GUTs which seek to extend this unification to all the forces, but almost all invoke some form of spontaneous symmetry breaking associated with a scalar field.

4.3.1 Topological defects

At very early times in the universe, the Higgs field is in the symmetric $\phi = 0$ "false vacuum" phase. As the universe expands and cools this symmetric state is eventually broken when $\phi = 0$ becomes unstable and ϕ rolls down to the "true vacuum". However, there may not be a unique true vacuum state. If there are many, or even a continuum of different true vacuum states, the horizon problem demonstrates that there is a limit to the scale on which the choice of broken symmetry vacuum state can be aligned. This is not just a question of initial conditions. Thermal fluctuations make the final true vacuum state truly random. Depending on the topology of the "vacuum manifold" defects form between regions of different alignment which may be topologically stable.

Consider the symmetric potential

$$V(\phi) = \frac{\lambda}{4}(\phi^2 - M^2)^2 \ , \tag{89}$$

where ϕ is a real scalar field, M is the GUT mass-scale, and λ is a dimensionless coupling constant. There are two equivalent true vacuum states: $\phi = \pm M$. If a thermal background leads to a finite temperature effective potential of the form

$$V_{\text{eff}}(\phi, T) = V(\phi) + \frac{\alpha}{2}T^2\phi^2 \ , \tag{90}$$

$\phi = 0$ will be a stable minimum while $d^2V_{\text{eff}}/d\phi^2 > 0$ which requires $T^2 > T_c^2 \equiv \lambda M^2/2\alpha$.

At lower temperatures the field must roll into one or other of the true vacua, spontaneously breaking the symmetry of the underlying potential. Between regions of different vacuum states a domain wall is formed which will have an approximate form

$$\phi(x) \simeq M \tanh(x\sqrt{2}/M\sqrt{\lambda}) \ . \tag{91}$$

This is actually the solution to Eq. (76) for a planar wall in flat-spacetime. The domain wall's thickness is $\delta = [(\sqrt{\lambda/2})M]^{-1} \simeq 10^{-30}\lambda^{-1/2}(10^{16}\text{GeV}/M)\text{cm}$ and we can calculate the surface density of the wall to be

$$\sigma = \frac{2\sqrt{2}}{\sqrt{\lambda}}M^3 \simeq 10^{71}\lambda^{-1/2}\left(\frac{M}{10^{16}\text{GeV}}\right)^3 \ \text{gcm}^{-2} \ . \tag{92}$$

As we might expect, these defects containing regions of the unbroken GUT phase are fantastically thin but have a huge energy density. Note that $H^{-1} \gg \delta$ when $T < T_c$ for $M \ll M_P$, so the flat-spacetime values for δ and σ should be reasonable estimates.

In the radiation dominated era the critical temperature is reached when the horizon length is d_c, given by equations (57) and (38) as

$$d_c = 2H_c^{-1} \simeq \sqrt{\frac{45}{4\pi^3 G g_{\text{eff}}}\frac{2\alpha}{\lambda M^2}} \ . \tag{93}$$

This represents the largest scales on which the vacuum state can be correlated at this time. Thus when symmetry breaking occurs at this temperature we must expect to find at least one domain wall per horizon volume, which implies an energy density $\rho_W \gtrsim (\sigma d_c^2)/d_c^3$.

The trouble is that this energy density represents a threat to our successful radiation dominated model of the early universe. Although initially this density is less than the radiation density, the energy stored in walls cannot be dissipated fast enough to avoid coming to dominate the universe. The correlation length of the network of walls cannot grow faster than the horizon, $\rho_W \propto d^{-1} \propto t^{-1}$, and so ρ_W must grow relative to the background radiation density which we know decreases as t^{-2}. A GUT that produces stable domain walls is not cosmologically viable.

On the other hand this argument suggests that linear defects which had a line-density μ and a correlation length that grew with the horizon size, could decay away sufficiently fast if $\rho_s \sim (\mu t)/t^3 \propto t^{-2}$. These objects, called cosmic strings, can form when the vacuum manifold has the topology of a loop, such as occurs for a two-component scalar field $\phi = \{\phi_1, \phi_2\}$ with potential energy density

$$V(\phi) = \frac{\lambda}{4}(\phi_1^2 + \phi_2^2 - M^2)^2 \; . \tag{94}$$

This is the basis of the cosmic string model for structure formation, whereby inhomogeneities are produced in an otherwise homogeneous universe simply due to the production of topological defects after spontaneous symmetry breaking in the early universe.

The type of defects that are produced are purely a consequence of the symmetries of the broken and unbroken phase. In particular, any symmetry breaking that includes the $U(1)$ symmetry of electromagnetism must produce topologically stable point-like defects which are magnetic monopoles.

4.3.2 The Monopole problem

Magnetic monopoles were first proposed as theoretically possible particles by Dirac in the 1930's. However it was only with the advent of GUT's based on spontaneous symmetry breaking that they became apparently inevitably present in the universe. The fact that our presently observed universe was made up of many causally disconnected horizon volumes at earlier times meant that there should be many monopoles formed. Using the horizon size, d_c, as an upper limit on the correlation length at the critical temperature calculated in equation (93) the number density of monopoles relative to the entropy density [given in equation (40)] must be

$$\frac{n_M}{s} > 10^{-9}\lambda^{3/2}\alpha^{-3/2}g_{\rm eff}^{1/2}\left(\frac{M}{10^{16}\ {\rm GeV}}\right) \tag{95}$$

which corresponds to a comoving number density (the number per coordinate volume, expanded up to a present-day physical volume) of about

$$n_{M0} > 10^{-6}\lambda^{3/2}\alpha^{-3/2}\left(\frac{M}{10^{16}{\rm GeV}}\right)^3 \tag{96}$$

The simple scaling argument used above for the evolution of a network of walls and strings, no longer applies in the case of disconnected point-like defects and instead the calculation is more analogous to our earlier particle survival calculation. Monopoles are never in thermal equilibrium but we can calculate the efficiency of monopole-anti-monopole annihilations for a given initial number density. In fact annihilation is only effective in reducing the number density to about $n_M/s \sim 10^{-10}$. This is similar to the number density of baryons. However, unlike baryons, the monopoles have a mass of order the GUT scale! Simply requiring that $\Omega h^2 < 1$ today requires

$$\frac{n_M}{s} < 10^{-24}\ \left(\frac{m_{\rm mon}}{10^{16}\ {\rm GeV}}\right) \tag{97}$$

This is totally incompatible with our calculated density of GUT monopoles.

This basic inability of the standard cosmological model to incorporate GUT physics was the final spur that lead cosmologists in the early 1980's to develop a radically different picture of the very early universe.

5 Inflation

5.1 Scalar field cosmology

For simplicity lets consider a spatially homogeneous field so that $\varphi = \varphi(t)$. In a homogeneous metric the field then has a total energy density

$$\rho_\varphi = \frac{1}{2}\dot{\varphi}^2 + V(\varphi) \tag{98}$$

Note that the self-interaction energy, V, is not diminished by the universal expansion as it is simply a function of φ. The only decrease in the scalar field's energy density is due to the frictional damping of the motion of the field due to the Hubble expansion. The continuity equation is

$$\dot{\rho}_\varphi = -3H\dot{\varphi}^2 \ . \tag{99}$$

Comparing this with equation (20) we see that the pressure can be written as

$$p_\varphi = \frac{1}{2}\dot{\varphi}^2 - V(\varphi) \ . \tag{100}$$

If the scalar field rolls slowly enough, the pressure can become negative. Indeed, in the limit that the field becomes stationary, the pressure is equal to minus the density and the continuity equation shows that the energy density remains a constant even in an expanding universe. The scalar field exerts a tension and, as the physical volume of a given coordinate space increases, the total energy increases.

In fact, the continuity equation also follows from the equation of motion for the field

$$\ddot{\varphi} + 3H\dot{\varphi} = -\frac{dV}{d\varphi} \tag{101}$$

[Multiply this by $\dot{\varphi}$ and you get equation (99).] This is the same as the equation of motion for a particle with coordinate φ, moving in a potential $V(\varphi)$, subject to frictional damping by the Hubble expansion. Such an analogy will get you a long way in understanding the cosmological evolution of a scalar field!

Despite their alluring simplicity, it is important to emphasize that no scalar particle has yet been discovered experimentally. It is not so surprising that the cosmological consequences were not seriously investigated until the 1980's. The Higgs mechanism seems to be a very good model for electroweak physics but the Higgs boson is yet to be discovered and it is worth remembering that if it did not exist, much of the motivation for the models I will now describe would be lost.

5.1.1 de Sitter expansion

If $V = V_0$ =constant then the Hubble damping in an expanding universe brings the field to rest and $\rho_\varphi \to V_0$. The same will happen if $V(\varphi)$ has a minimum at $\varphi = \varphi_0$. Even if the field initially oscillates about this point, the Hubble expansions dissipates the kinetic energy and again $\rho_\varphi \to$constant.

The energy density of radiation (or dust) is always diluted by the Hubble expansion so if $V_0 > 0$ the potential energy density of the scalar field eventually dominates. Note that as $T \to 0$ temperature corrections to the effective potential can be neglected. Because $\rho_\varphi \to V_0$ and $p_\varphi \to -\rho_\varphi$ we have in effect a barotropic fluid with barotropic index $\gamma = 0$.

Consider now the effect this matter has on the expansion. The constraint equation (18) becomes

$$\left(\frac{\dot{a}}{a}\right)^2 = H_\infty^2 - \frac{\kappa c^2}{a^2} \tag{102}$$

where $H_\infty^2 \equiv 8\pi G V_0/3$. This can be integrated to give

$$a(t) = \begin{cases} H_\infty^{-1} \cosh\left[H_\infty(t-t_0)\right] & \text{for } \kappa = +1 \\ a_0 \exp\left[H_\infty(t-t_0)\right] & \text{for } \kappa = 0 \\ H_\infty^{-1} \sinh\left[H_\infty(t-t_0)\right] & \text{for } \kappa = -1 \end{cases} \tag{103}$$

Note how as $t \to \infty$ all the solutions approach exponential expansion. In particular the acceleration of the scale factor $\ddot{a}$ is always positive, in contrast with all the models we have previously considered. This runaway expansion of the scale factor has become known as inflation.

This is the de Sitter spacetime first derived by de Sitter for a universe dominated by a *cosmological constant*. The constant potential energy density of the scalar field acts just like a cosmological constant, $\Lambda = 8\pi G V_0$, in the constraint equation (18).

5.1.2 Slow-roll inflation

A constant potential energy density $V = V_0$ provides the simplest case of inflationary expansion, but any sufficiently flat potential will do. If the Hubble damping is sufficiently strong ($3H\dot{\varphi} \gg \ddot{\varphi}$), we can use the approximate *slow-roll equation* for the scalar field

$$3H\dot{\varphi} \simeq -\frac{dV}{d\varphi} \tag{104}$$

and assume the Hubble expansion is dominated by the potential ($\dot{\varphi}^2 \ll V$)

$$H^2 \simeq \frac{8\pi G}{3} V \tag{105}$$

(I will discuss shortly why we might neglect spatial curvature.)

Though the slow roll approximation might appear a crude approximation, one can show that it is at least a self-consistent approximation so long as both the slope and curvature of the potential are small. We write these in terms of the *slow-roll parameters*:

$$\begin{aligned} \epsilon_V &\equiv \frac{M_P^2}{16\pi}\left(\frac{V'}{V}\right)^2 , \\ \eta_V &\equiv \frac{M_P^2}{8\pi}\left(\frac{V''}{V}\right) . \end{aligned} \tag{106}$$

A necessary condition for the slow-roll approximation to hold is then $\max\{\epsilon_V, |\eta_V|\} \ll 1$.

While this approximation holds, $\dot{H} \ll H^2$ and the universe undergoes a period of quasi-exponential expansion. It is useful to quantify the amount of inflation by the *number of e-foldings*

$$N \equiv \ln\left(\frac{a}{a_i}\right) = \int_{t_i}^{t} H dt \simeq -8\pi G \int_{\varphi_i}^{\varphi} \frac{V d\varphi}{V'} \tag{107}$$

I'll give a specific example later.

Unlike domination by a network of domain walls which produces an inhomogeneous and anisotropic universe, a universe dominated by a homogeneous scalar field will be both homogeneous and isotropic. Of course somehow we have to get back to the radiation dominated early universe we know and love, but before that lets see how a scalar-field might resolve some of the enigmas and conundrums I posed earlier...

5.2 Generic Inflation

We have seen that a homogeneous scalar field can lead to an accelerated expansion

$$\ddot{a} > 0 \tag{108}$$

and it is this that I will take as my definition of *inflation*. Note that this implies that $\dot{a} = aH$ is an increasing function of time. The quantity $(aH)^{-1}$ is a coordinate, or *comoving*, measure of the Hubble length. Thus we can state that the comoving Hubble length decreases during inflation, but increases during all other eras.

A scalar field provides one possible source of inflation but one should bear in mind that there may be others. We see from the evolution equation (17) that inflation occurs whenever

$$p < -\frac{1}{3}\rho \tag{109}$$

Returning for a moment to the case of a scalar field, we can see from equations (100) and (109) that the condition for inflation is satisfied when

$$\dot{\varphi}^2 < V(\varphi) \tag{110}$$

This is evidently satisfied when the slow-roll approximation is valid. Such a scalar field is commonly referred to as the *inflaton*.

5.2.1 Flatness

If we look again at equation (58) we see that Ω falls away from 1 whenever $\dot{a}^2$ decreases. However during inflation, by my definition, $\dot{a}$ is increasing and $\Omega - 1 \to 0$. Inflation drives Ω towards unity. As seen in the case of de Sitter spacetime, as $t \to \infty$ the spatial curvature becomes irrelevant and all models approach the flat-space model.

As long as inflation lasts long enough there is no trouble in driving Ω arbitrarily close to one. Assuming we (re)start a radiation dominated era at some time, t_r, Ω once again begins to fall away from unity and by the present

$$1 - \Omega_0 = \left(\frac{T_r}{T_{eq}}\right)^2 \left(\frac{T_{eq}}{T_0}\right)(1 - \Omega_r) \tag{111}$$

So to force Ω to within 0.1 of unity today requires $|1 - \Omega_r| < 10^{-60}(T_r/T_P)^2$.

5.2.2 Horizon length

In calculating the horizon size during the radiation or dust dominated eras I implicitly assumed that there was no significant contribution from any very early era. Suppose there is an initial non-radiation dominated era, then we have to split the integral in equation (57) to give

$$d(t) = a(t)\left[\int_{t_f}^{t} \frac{dt'}{a(t')} + \int_{t_i}^{t_f} \frac{dt'}{a(t')}\right] = 2(t - t_f) + \left(\frac{t}{t_f}\right)^{1/2} d_f \tag{112}$$

The second term can be neglected for conventional (non-inflationary) evolution if $t_f \ll t$. For instance, for power-law evolution $a \propto (t - t_i)^n$ with $n < 1$, we have $d_f = (t_f - t_i)/(1 - n) = (n/(1-n))H_f^{-1}$.

However after a period of inflation this may no longer be the case. We can write

$$d_f = a_f \int_{t_i}^{t_f} \frac{dt'}{a(t')} = a_f \int_{a_i}^{a_f} \frac{1}{\dot{a}} \frac{da}{a} \tag{113}$$

During inflation $\dot{a}$ increases so $\dot{a}_f > \dot{a}$ at earlier times and hence

$$d_f > a_f \int_{a_i}^{a_f} \frac{1}{\dot{a}_f} \frac{da}{a} = H_f^{-1} \ln\left(\frac{a_f}{a_i}\right) \tag{114}$$

Thus in the limit $a_i \to 0$ horizon length is divergent. Put another way, as the duration of inflation, represented here by the number of e-folds $N_{tot} = \ln(a_f/a_i)$ becomes arbitrarily large, so physical scale of initially causally connected regions becomes arbitrarily large.

5.2.3 Monopoles

The hugely increased horizon size can easily solve the monopole problem. If the temperature in the subsequent radiation dominated era is not high enough to restore the symmetric phase, the Higgs field is can be correlated on arbitrarily large scales. Even if there was a monopole forming transition before (or during) inflation, the number density will be diluted by the expansion and can easily become negligible.

Note however that the increased horizon size does not necessarily solve the monopole problem. If we (re-)start the radiation dominated era at temperatures above the critical temperature for the monopole-forming symmetry breaking, then thermal fluctuations can wash out any large-scale correlations, and the relevant upper limit on correlations again becomes the particle horizon during the radiation dominated era.

5.2.4 Entropy and reheating

So far I have cavalierly assuming that we can return to a conventional radiation dominated solution at some point after inflation. This is clearly necessary as I have spent so long arguing how successful this standard model is. The simplest assumption would be that at some instant, all the scalar field energy density is converted into "ordinary" relativistic matter (radiation). The spatial curvature and total energy density, or equivalently the scale factor and expansion rate, should would then be continuous at the transition, giving the instantaneous reheat temperature

$$T_{R*} = \left(\frac{30}{\pi^2 g_{\text{eff}}} V_0 \right)^{1/4} \tag{115}$$

This must be a non-adiabatic, irreversible process during which the large entropy of our observed universe can be generated.

How is this *reheating* to be achieved? Inflation is so effective at diluting the density of other matter, the effective of curvature, etc. that we have to invoke on the spontaneous decay of the self-interaction energy of the inflaton field. This simply never happens in the example given above where $V = V_0$ =constant. However if $V(\varphi)$ after a period of slow-rolling along a flat potential, the scalar field falls into a minimum it may indeed decay into other matter. If the inflaton field is coupled to other matter, as it oscillates about the minimum the coherent oscillations may excite oscillations in other fields. In terms of particle interactions, the inflaton particle decays into some other particles species.

Suppose the lifetime of the inflaton is given by Γ_φ^{-1}, then we can (very crudely) represent this decay by an interaction between the scalar field and the radiation field

$$\dot{\rho}_\varphi = -3H\dot{\varphi}^2 - \Gamma_\varphi \dot{\varphi}^2 \tag{116}$$

$$\dot{\rho}_\gamma = -4H\rho_\gamma + \Gamma_\varphi \dot{\varphi}^2 \tag{117}$$

In the equation of motion for the scalar field this looks like another frictional term damping the motion. It gives a characteristic time-scale for the decay. During this time the time-averaged pressure of the oscillating inflaton is zero and the universe expands like a dust-dominated model, $\rho \propto t^{-2}$. Thus assuming $\Gamma_\varphi \ll H$ at the end of inflation, the energy density after the inflaton has decayed is approximately given by

$$\rho_R \sim \frac{1}{6\pi} M_P^2 \Gamma_\varphi^2 \tag{118}$$

and the reheat temperature is

$$T_R \simeq 0.6 g_{\text{eff}}^{-1/4} \left(M_P \Gamma_\varphi \right)^{1/2} \tag{119}$$

$$= \epsilon\, T_{R*} \tag{120}$$

where the "efficiency" $\epsilon \simeq (\Gamma_\varphi / H)^{1/2}$.

It is important to emphasize that this is just a phenomenological representation of the decay of coherent oscillations, and in particular this decay mode isn't present during the slow-roll phase. Indeed the whole issue of reheating is still rather uncertain. Only recently has the importance of resonant effects been properly appreciated and their detailed effects are still being investigated.

T_R^3. Putting all this together gives an entropy in the final

6 Cosmological perturbations

The single most remarkable feature of inflation, and one quite unexpected when inflation was first proposed by Guth, is that the large-scale structure of the universe can originate from the zero-point vacuum fluctuations of the field(s) driving inflation.

To investigate this we need to consider the evolution of inhomogeneities in an inflating universe, and even flirt with quantum gravity.

6.1 Scalar field fluctuations

Consider a first-order inhomogeneous perturbation about a homogeneous scalar field in an (unperturbed) FRW spacetime:

$$\varphi \to \varphi(t) + \left(\frac{1}{2\pi}\right)^3 \int d^3k \delta\varphi_k(t) Q_k(x) \tag{121}$$

where I have split the inhomogeneous perturbations into harmonics on the maximally symmetric 3-space, Q_k, with eigenvalues $-k^2$.

For a massless, non-interacting field ($V'' \simeq 0$), and neglecting for the moment the gravitational back-reaction, the inhomogeneous modes obey the wave equation:

$$\delta\ddot{\varphi}_k + 3H\delta\dot{\varphi}_k + \frac{k^2}{a^2}\delta\varphi = 0\,. \tag{122}$$

This is the equation for a damped oscillator with two characteristic time-scales:

1. damping timescale: $t_{\rm damp} \sim H^{-1}$,
2. oscillation timescale: $t_{\rm osc} \sim \lambda \sim a/k$.

Hence we model the evolution into two regimes:

1. under-damped: $k > aH \to$ "sub-horizon",
2. over-damped: $k < aH \to$ "super-horizon".

The defining feature of inflation is that the comoving Hubble length decreases ($\ddot{a} > 0$) and modes *exit the horizon* as $aH = \dot{a}$ increases. Individual modes evolve from the sub- to super-horizon regimes, i.e., from under-damped oscillations on small scales to become over-damped (or more loosely "frozen-in") on large scales.

After inflation, in a conventional radiation or dust dominated era the opposite happens and modes *re-enter the horizon*, evolving from the super- to sub-horizon regime.

6.1.1 Quantum fluctuations

I will now give a slightly heuristic derivation of the amplitude of quantum fluctuations during inflation starting from the effective action which yields the equation of motion (122) for each mode of the linearly perturbed scalar field in the previous sub-section:

$$\delta S_k = \int dt\, a^3 \left[\frac{1}{2}\delta\dot{\varphi}_k^2 - \frac{1}{2}k^2\delta\varphi_k^2\right]\,. \tag{123}$$

Introducing the rescaled field perturbation

$$u \equiv a\delta\varphi \,, \tag{124}$$

and using conformal time, the effective action becomes

$$\delta S_k = \int d\eta \left[\frac{1}{2} u_k'^2 - \frac{1}{2} \left(k^2 - \frac{a''}{a} \right) u_k^2 \right] , \tag{125}$$

where a prime denotes a derivative with respect to conformal time. This is the effective action for a simple oscillator with time-dependent mass $-a''/a$. The corresponding equation of motion is

$$u_k'' + \left(k^2 - \frac{a''}{a} \right) u_k = 0 \,. \tag{126}$$

In the short-wavelength limit ($k^2/|a''/a| \to \infty$) we have a simple harmonic oscillator with solution $u_k \propto e^{\pm ik\eta}$. As such it is a textbook system to quantise!

In the quantised theory u_k and u_k' become canonically conjugate operators $\hat{q}_k$ and $\hat{p}_k$ which obey the canonical commutation relations. In the Heisenberg representation we write the time-dependent operators in terms of time-dependent mode functions and time-independent annihilation and creation operators where the mode function $u_k(\eta)$ is a solution of the classical equation of motion. The commutation relations impose the normalisation condition

$$u_k^* u_k' - u_k u_k^{*\prime} = -i \,. \tag{127}$$

In effect this determines the amplitude of the zero-point fluctuations in the quantum vacuum.

We impose the vacuum normalisation (127) in the short wavelength limit ($k^2/|a''/a| \to \infty$) where the solution is a simple harmonic oscillator. By convention we describe the vacuum state as positive frequency modes and hence we obtain

$$u_k = \frac{e^{-ik\eta}}{\sqrt{2k}} \tag{128}$$

in this limit. The choice of phase of the vacuum fluctuations is arbitrary, but the normalisation is crucial.

Remember that in an inflationary era (of sufficient duration) any given mode starts within the horizon so the short-wavelength limit coincides with the early-time limit. By fixing the amplitude of short-wavelength fluctuations to coincide with the zero-point fluctuations of the quantum vacuum one obtains a prediction for the amplitude of perturbations on arbitrarily long-wavelengths at late times. The large-scale structure of the post-inflationary universe can be predicted by applying quantum field theory in a curved spacetime - an audacious claim.

In order to study the average value of the perturbed field in real space we must integrate over the vacuum fluctuations in the harmonic k-space (e.g., Fourier space):

$$\langle \delta\varphi^2 \rangle_{\text{realspace}} = \left(\frac{1}{2\pi} \right)^3 \int d^3k \, \frac{|u_k|^2}{a^2} \,. \tag{129}$$

For the vacuum solution (128), valid on short-wavelengths, this would give

$$\langle \delta\varphi^2 \rangle_{\rm realspace} = \left(\frac{1}{2\pi}\right)^3 \int 4\pi k^2 \frac{1}{2ka^2} dk \,. \tag{130}$$

I will write the power-spectrum of fluctuations as the contribution per logarithmic interval in k-space to the real-space fluctuations, i.e.,

$$\langle \delta\varphi^2 \rangle_{\rm realspace} = \int d(\ln k)\, \mathcal{P}_{\delta\varphi}(k) \,. \tag{131}$$

Hence for vacuum fluctuations, Eq. (130) this yields

$$\mathcal{P}_{\delta\varphi}(k) = \left(\frac{k}{2\pi a}\right)^2 \,. \tag{132}$$

This is only valid on short-wavelengths, typically up to about the horizon scale, after which the fluctuations become over-damped (or "frozen-in"). Thus vacuum fluctuations on short-wavelengths lead to perturbations of the field at horizon-exit

$$\mathcal{P}_{\delta\varphi}(k = aH) = \left(\frac{H}{2\pi}\right)^2 \,. \tag{133}$$

6.2 Semi-classical gravity

The process described above whereby short-wavelength vacuum fluctuations are stretched up to arbitrarily long-wavelengths during inflation is closely related to Hawking radiation from a black hole. In both cases it is the presence of an event horizon that allows short-wavelength fluctuations at the horizon to produce long-wavelength perturbations for the distant observer. Both phenomena are manifestations of quantum gravity. Although we do not have a consistent theory of quantum gravity, we can apply the standard rules of quantum field theory to linear perturbations in curved spacetime. This is referred to as semi-classical gravity.

It seems reasonable to expect any complete theory of quantum gravity to reproduce semi-classical results, just as any relativistic theory of gravity must reproduce the correct Newtonian limit. Indeed, one of the recent 'successes' of string theory – a putative theory of everything – has been to reproduce semi-classical calculations of Hawking radiation from extremal black holes.

There wasn't much gravity involved in the calculation I've just presented for scalar field fluctuations in a fixed (FRW) background spacetime. But if one wishes to discuss the perturbations of the field driving the inflationary expansion one should allow for the gravitational back-reaction of the field. Indeed, one cannot in general unambiguously separate out the field perturbations from those of the spacetime. This is the gauge problem.

Here is a crash course in the gauge-problem for curvature perturbations about an FRW universe...

6.2.1 Metric perturbations

There are many different ways of characterizing cosmological perturbations, reflecting the arbitrariness in the choice of coordinates (gauge), which in turn determines the slicing of spacetime into spatial hypersurfaces, and its threading into timelike worldlines. The line element allowing arbitrary linear scalar perturbations of a Friedmann–Robertson–Walker (FRW) background can be written [2, 1, 3, 4]

$$\begin{aligned} ds^2 &= -(1+2A)dt^2 + 2a^2(t)\nabla_i B\, dx^i dt \\ &\quad + a^2(t)\left[(1-2\psi)\gamma_{ij} + 2\nabla_i\nabla_j E\right] dx^i dx^j \,. \end{aligned} \tag{134}$$

The unperturbed spatial metric for a space of constant curvature κ is given by γ_{ij} and covariant derivatives with respect to this metric are denoted by ∇_i. [2] The intrinsic curvature of a spatial hypersurface, ${}^{(3)}R$, is usually described by the dimensionless curvature perturbation[3] ψ, where

$${}^{(3)}R = \frac{6\kappa}{a^2} + \frac{12\kappa}{a^2}\psi + \frac{4}{a^2}\nabla^2\psi \,. \tag{136}$$

Of primary interest to us, and much of modern cosmology, is the evolution of the curvature perturbation, ψ, on the constant-time hypersurfaces defined in Eq. (134). These constant-time hypersurfaces are orthogonal to the unit time-like vector field [3]

$$n^\mu = (1-A, -\nabla^i B) \,. \tag{137}$$

The expansion of the spatial hypersurfaces with respect to the proper time, $d\tau \equiv (1+A)dt$, of observers with 4-velocity n^μ, is given by

$$\theta \equiv n^\mu{}_{;\mu} = 3H\,(1-A) - 3\dot\psi + \nabla^2\sigma \,, \tag{138}$$

where the scalar describing the shear is

$$\sigma = \dot E - B \,. \tag{139}$$

However it is useful to define the expansion rate with respect to the coordinate time

$$\tilde\theta = (1+A)\theta = 3H - 3\dot\psi + \nabla^2\sigma \,. \tag{140}$$

We can write this as an equation for the time evolution of ψ in terms of the perturbed expansion, $\delta\tilde\theta \equiv \tilde\theta - 3H$, and the shear:

$$\dot\psi = -\frac{1}{3}\delta\tilde\theta + \frac{1}{3}\nabla^2\sigma \,. \tag{141}$$

Note that this is independent of the field equations and follows simply from the geometry.

[2] For comparison with the notation of Bardeen [1] note that

$$\begin{aligned} A &\equiv A_B Q^{(0)} \,, & \psi &\equiv -\left(H_L + \frac{1}{3}H_T\right) Q^{(0)} \,, \\ B &\equiv \frac{B_B Q^{(0)}}{ka} \,, & E &\equiv \frac{H_T Q^{(0)}}{k^2} \,, \end{aligned} \tag{135}$$

where Bardeen explicitly included $Q^{(0)}(x^i)$, the eigenmodes of the spatial Laplacian, ∇^2, with eigenvalue $-k^2$.

[3] This quantity is denoted $\mathcal{R}$ in Refs. [5, 6].

6.2.2 Gauge dependence

Once you break the maximal symmetry of the constant-time hypersurfaces (or time-slicing) in the FRW model, there is no longer a preferred time coordinate. Under a first-order coordinate transformation

$$t \to t + \delta t \,, \tag{142}$$

(where δt may be space and time dependent), the scalar field perturbation at given coordinate changes:

$$\delta\varphi \to \delta\varphi - \dot{\varphi}\delta t \,, \tag{143}$$

so does the density

$$\delta\rho \to \delta\rho - \dot{\rho}\delta t \,, \tag{144}$$

and so do the geometrical perturbations, such as the intrinsic spatial curvature perturbation

$$\psi \to \psi + H\delta t \,, \tag{145}$$

or the shear

$$\sigma \to \sigma - \delta t \,. \tag{146}$$

The field perturbation $\delta\varphi$ or the curvature perturbation ψ is a gauge-dependent quantity. Perform a gauge transformation $\delta t = \delta\varphi/\dot{\varphi}$ and you find $\delta\varphi \to 0$, i.e., you are working on uniform-field hypersurfaces. Or pick $\delta t = -\psi/H$ and you have $\psi \to 0$ on uniform-curvature hypersurfaces.

One can however construct *gauge-invariant perturbations* such as

$$Q \equiv \delta\varphi + \frac{\dot{\varphi}}{H}\psi \,, \tag{147}$$

which is the field perturbation on uniform-curvature hypersurfaces, or

$$\mathcal{R} \equiv \psi + \frac{H}{\dot{\varphi}}\delta\varphi \,, \tag{148}$$

which is the curvature perturbation on uniform-field hypersurfaces. The two are proportional

$$HQ = \dot{\varphi}\mathcal{R} \,, \tag{149}$$

as they both measure the physical displacement between uniform-field and uniform-curvature hypersurfaces.

6.2.3 Quantum fluctuations from inflation

Mukhanov [15] was the first to give a gauge-invariant description of the quantum fluctuations of the inflaton field including linear metric perturbations. He showed that the effective action for the fluctuations can be written as

$$\delta S_k = \int d\eta \left[\frac{1}{2} u_k'^2 - \frac{1}{2}\left(k^2 - \frac{z''}{z}\right) u_k^2 \right] \,, \tag{150}$$

where

$$z \equiv \frac{a\dot{\varphi}}{H} \tag{151}$$

and

$$u \equiv aQ = a\left[\delta\varphi + \frac{\dot{\varphi}}{H}\psi\right] \tag{152}$$

is a gauge-invariant definition of the inflaton field fluctuations on uniform-curvature hypersurfaces.

The analysis is then identical to that presented for the effective action (125). In particular, in the slow-roll approximation we have

$$\frac{z''}{z} \simeq 2a^2H^2\left[1 + \epsilon - \frac{3}{2}\eta\right] \tag{153}$$

and the power spectrum of field perturbations on uniform-curvature hypersurfaces in the long-wavelength regime ($k^2 \ll a^2H^2$) is related to the Hubble rate at horizon crossing:

$$\mathcal{P}_Q(k = aH) = \left(\frac{H}{2\pi}\right)^2 \tag{154}$$

6.3 Large-scale perturbations

The curvature perturbation on uniform-density hypersurfaces, can be written as[4]

$$-\zeta = H\xi \ , \tag{155}$$

where the displacement between the uniform-density ($\delta\rho = 0$) hypersurface and the uniform-curvature ($\psi = 0$) hypersurface has the gauge-invariant definition:

$$\xi \equiv \frac{\psi}{H} + \frac{\delta\rho}{\dot{\rho}} \, . \tag{156}$$

Alternatively one can work in terms of the density perturbation on uniform-curvature hypersurfaces

$$\delta\rho_\psi = \dot{\rho}\,\xi \, , \tag{157}$$

where the subscript ψ indicates the uniform-curvature hypersurface.

The curvature perturbation on uniform-density hypersurfaces, ζ, is often chosen as a convenient gauge-invariant definition of the scalar metric perturbation on large scales. These hypersurfaces become ill-defined if the density is not strictly decreasing, as can occur in a scalar field dominated universe when the kinetic energy of the scalar field vanishes. In this case one can instead work in terms of the density perturbation on uniform-curvature hypersurfaces, $\delta\rho_\psi$, which remains finite.

During slow-roll inflation the density perturbation is directly related to the field perturbation

$$\delta\rho \simeq V'\delta\varphi \, , \tag{158}$$

[4] The sign of ζ is chosen here to coincide with Refs. [7, 8].

and hence the uniform-density and uniform-field hypersurfaces coincide

$$\zeta = -\mathcal{R}\,. \tag{159}$$

The pressure perturbation (in any gauge) can be split into adiabatic and entropic (non-adiabatic) parts, by writing

$$\delta p = c_{\rm s}^2 \delta\rho + \dot{p}\Gamma\,, \tag{160}$$

where $c_s^2 \equiv \dot{p}/\dot{\rho}$. The non-adiabatic part is $\delta p_{\rm nad} \equiv \dot{p}\Gamma$, and

$$\Gamma \equiv \frac{\delta p}{\dot{p}} - \frac{\delta\rho}{\dot{\rho}}\,. \tag{161}$$

The entropy perturbation Γ is *gauge-independent*, and represents the displacement between hypersurfaces of uniform pressure and uniform density.

In the slow-roll approximation (with a single scalar field) the scalar field value determines both the density and the pressure so there can be no entropy perturbation: $\Gamma \simeq 0$.

6.3.1 Evolution of the curvature perturbation

Irrespective of the gravitational field equations we can derive important results from the local conservation of the energy–momentum tensor $T^\mu_{\nu;\mu} = 0$. The energy conservation equation $n^\nu T^\mu_{\nu;\mu} = 0$ for first-order density perturbations gives

$$\dot{\delta\rho} = -3H(\delta\rho + \delta p) + (\rho + p)\left[3\dot{\psi} - \nabla^2\left(\sigma + v + B\right)\right]\,, \tag{162}$$

where $\nabla^i v$ is the perturbed 3-velocity of the fluid. In the uniform-density gauge, where $\delta\rho = 0$ and $\psi = -\zeta$, the energy conservation equation (162) immediately gives

$$\dot{\zeta} = -\frac{H}{\rho + p}\delta p_{\rm nad} - \frac{1}{3}\nabla^2\left(\sigma + v + B\right)\,. \tag{163}$$

This result is derived without invoking any gravitational field equations. We see that ζ is constant if (i) there is no non-adiabatic pressure perturbation, and (ii) the divergence of the 3-momentum on zero-shear hypersurfaces, $\nabla^2(v + B + \sigma)$, is negligible.

On sufficiently large scales, gradient terms can be neglected and [11, 9]

$$\dot{\zeta} = -\frac{H}{\rho + p}\delta p_{\rm nad}\,, \tag{164}$$

which implies that ζ is constant if the pressure perturbation is adiabatic. It has been argued [9] that the divergence is likely to be negligible on all super-horizon scales, and in the following we shall make that assumption.

It is this constancy of ζ on super-Hubble scales that allows us to relate vacuum fluctuations in some abstract scalar field in the very early universe to density perturbations in the radiation or dust-dominated era.

In the dust-dominated era this constant value for ζ on large-scales is directly related to the density perturbation at horizon re-entry:

$$\delta_H \equiv \left(\frac{\delta\rho}{\rho}\right)_{k=aH} = \frac{2}{5}\zeta\,. \tag{165}$$

6.3.2 Matter plus radiation

In a multi-fluid system we can define uniform-density hypersurfaces for each fluid and a corresponding curvature perturbation on these hypersurfaces, $\zeta_{(i)} \equiv -\psi - \delta\rho_{(i)}/\dot{\rho}_{(i)}$. Equation (163) then shows that $\zeta_{(i)}$ *remains constant for adiabatic perturbations in any fluid whose energy–momentum is locally conserved:* $n^\nu T_{(i)\ \nu;\mu}^{\ \mu} = 0$. Thus, for example, in a universe containing non-interacting cold dark matter plus radiation, which both have well-defined equations of state ($p_{\rm m} = 0$ and $p_\gamma = \rho_\gamma/3$), the curvatures of uniform-matter-density hypersurfaces, ζ_m, and of uniform-radiation-density hypersurfaces, ζ_γ, remain constant on super-horizon scales. The curvature perturbation on the uniform-total-density hypersurfaces is given by

$$\zeta = \frac{(4/3)\rho_\gamma\zeta_\gamma + \rho_m\zeta_{\rm m}}{(4/3)\rho_\gamma + \rho_m} . \tag{166}$$

At early times in the radiation dominated era ($\rho_\gamma \gg \rho_{\rm m}$) we have $\zeta_{\rm init} \simeq \zeta_\gamma$, while at late times ($\rho_{\rm m} \gg \rho_\gamma$) we have $\zeta_{\rm fin} \simeq \zeta_{\rm m}$. ζ remains constant throughout only for adiabatic perturbations where the uniform-matter-density and uniform-radiation-density hypersurfaces coincide, ensuring $\zeta_\gamma = \zeta_{\rm m}$. The isocurvature (or entropy) perturbation is conventionally denoted by the perturbation in the ratio of the photon and matter number densities

$$S = \frac{\delta n_\gamma}{n_\gamma} - \frac{\delta n_{\rm m}}{n_{\rm m}} = 3\left(\zeta_\gamma - \zeta_{\rm m}\right) . \tag{167}$$

Hence the entropy perturbation for any two non-interacting fluids always remains constant on large scales independent of the gravitational field equations. Hence we recover the standard result for the final curvature perturbation in terms of the initial curvature and entropy perturbation[5]

$$\zeta_{\rm fin} = \zeta_{\rm ini} - \frac{1}{3}S . \tag{168}$$

6.3.3 The separate universe approach

There is a particularly simple approach to studying the evolution of perturbations on large scales [6] , which has been employed in some multi-component inflation models [12, 13, 5, 14, 6, 9]. This considers each super-horizon sized region of the Universe to be evolving like a separate Robertson–Walker universe where density and pressure may take different values, but are locally homogeneous. After patching together the different regions, this can be used to follow the evolution of the curvature perturbation with time.

Consider two such locally homogeneous regions (a) and (b) at fixed spatial coordinates, separated by a coordinate distance λ, on an initial hypersurface (e.g., uniform-density hypersurface) specified by a fixed coordinate time, $t = t_1$, in the appropriate gauge (e.g., uniform-density gauge). The initial large-scale curvature perturbation on the scale λ can then be defined (independently of the background) as

$$\delta\psi_1 \equiv \psi_{a1} - \psi_{b1} . \tag{169}$$

[5] This result was derived first by solving a differential equation [10], and then [9] by integrating Eq. (164) using Eq. (166). We have here demonstrated that even the integration is unnecessary.

[6] The section is taken from Ref. [17].

On a subsequent hypersurface defined by $t = t_2$ the curvature perturbation at (a) or (b) can be evaluated using Eq. (141) [but neglecting $\nabla^2\sigma$] to give [5]

$$\psi_{a2} = \psi_{a1} - \delta N_a \,, \tag{170}$$

where the integrated expansion between the two hypersurfaces along the world-line followed by region (a) is given by $N_a = N + \delta N_a$, with $N \equiv \ln a$ the expansion in the unperturbed background and

$$\delta N_a = \int_1^2 \frac{1}{3}\delta\tilde{\theta}_a dt \,. \tag{171}$$

The curvature perturbation when $t = t_2$ on the comoving scale λ is thus given by

$$\delta\psi_2 \equiv \psi_{a2} - \psi_{b2} = \delta\psi_1 - (N_a - N_b) \;. \tag{172}$$

In order to calculate the change in the curvature perturbation in any gauge on very large scales it is thus sufficient to evaluate the difference in the integrated expansion between the initial and final hypersurface along different world-lines.

In particular, using Eq. (172), one can evolve the curvature perturbation, ζ, on super-horizon scales, knowing only the evolution of the family of Robertson–Walker universes, which according to the separate Universe assumption describe the evolution of the Universe on super-horizon scales:

$$\Delta\zeta = \Delta N \,, \tag{173}$$

where $\Delta\zeta = -\psi_a + \psi_b$ on uniform-density hypersurfaces and $\Delta N = N_a - N_b$ in Eq. (172). This evolution is in turn specified by the values of the relevant fields during inflation, and as a result one can calculate ζ at horizon re-entry from the vacuum fluctuations of these fields.

While it is a non-trivial assumption to suppose that every comoving region *well outside the horizon* evolves like an unperturbed universe, there has to be some scale $\lambda_{\rm s}$ for which that assumption is true to useful accuracy. If there were not, the concept of an unperturbed (Robertson–Walker) background would make no sense. I use the phrase 'background' to describe the evolution on a much larger scale λ_0, which should be much bigger even than our present horizon size, with respect to which the perturbations in section 6.2 were defined. It is important to distinguish this from regions of size $\lambda_{\rm s}$ large enough to be treated as locally homogeneous, but which when pieced together over a larger scale, λ, represent the long-wavelength perturbations under consideration. Thus we require a hierarchy of scales:

$$\lambda_0 \gg \lambda \gg \lambda_{\rm s} > cH^{-1} \;. \tag{174}$$

Ideally λ_0 would be taken to be infinite. However it may be that the Universe becomes highly inhomogeneous on some very much larger scale, $\lambda_{\rm e} \gg \lambda_0$, where effects such as stochastic or eternal inflation determine the dynamical evolution. Nevertheless, this will not prevent us from defining an effectively homogeneous background in our observable Universe, which is governed by the local Einstein equations and hence impervious to anything happening on vast scales. Specifically we will assume that it is possible to foliate spacetime on this large scale λ_0 with spatial hypersurfaces.

When we use homogeneous equations to describe separate regions on length scales greater than λ_s, we are implicitly assuming that the evolution on these scales is independent of shorter wavelength perturbations. This is true within linear perturbation theory in which the evolution of each Fourier mode can be considered independently, but any non-linear interaction introduces mode-mode coupling which undermines the separate universes picture. The separate universe model may still be used for the evolution of linear metric perturbation if the perturbations in the total density and pressure remain small, but a suitable model (possibly a thermodynamic description) of the effect of the non-linear evolution of matter fields on smaller scales may be necessary in some cases.

Adiabatic perturbations in the density and pressure correspond to shifts forwards or backwards in time along the background solution, $\delta p/\delta\rho = \dot{p}/\dot{\rho} \equiv c_s^2$, and hence $\Gamma = 0$ in Eq. (161). For example, in a universe containing only baryonic matter plus radiation, the density of baryons or photons may vary locally, but the perturbations are adiabatic if the ratio of photons to aryons remains unperturbed. Different regions are compelled to undergo the same evolution along a unique trajectory in field space, separated only by a shift in the expansion. The pressure p thus remains a unique function of the density ρ and the energy conservation equation, $d\rho/dN = -3(\rho + p)$, determines ρ as a function of the integrated expansion, N. Under these conditions, uniform-density hypersurfaces are separated by a uniform expansion and hence the curvature perturbation, ζ, remains constant.

For $\Gamma \neq 0$ it is no longer possible to define a simple shift to describe both the density and pressure perturbation. The existence of a non-zero pressure perturbation on uniform-density hypersurfaces changes the equation of state in different regions of the Universe and hence leads to perturbations in the expansion along different worldlines between uniform-density hypersurfaces. This is consistent with Eq. (163) which quantifies how the non-adiabatic pressure perturbation determines the variation of ζ on large scales [11, 9].

The entropy perturbation between any two quantities (which are spatially homogeneous in the background) has a naturally gauge-invariant definition [which follows from the obvious extension of Eq. (161)]

$$\Gamma_{xy} \equiv \frac{\delta x}{\dot{x}} - \frac{\delta y}{\dot{y}} \,. \tag{175}$$

We define a generalized adiabatic condition which requires $\Gamma_{xy} = 0$ for any physical scalars x and y. In the separate universes picture this condition ensures that if all field perturbations are adiabatic at any one time (i.e. on any spatial hypersurface), then they must remain so at any subsequent time. Purely adiabatic perturbations can never give rise to entropy perturbations on large scales as all fields share the same time shift, $\delta t = \delta x/\dot{x}$, along a single phase-space trajectory.

6.4 Primordial perturbation spectra

The curvature perturbation ζ on large-scales provides a link between the inflaton field perturbations during inflation and density perturbations in the radiation and dust dominated eras:

$$\delta_H = \frac{2}{5}\zeta = \left[\frac{2}{5}\frac{H}{\dot{\varphi}}Q\right]_{k=aH} . \tag{176}$$

I will use the notation of Lidsey et al [19] to give the contribution of scalar and tensor perturbations to the cmb anisotropies.

$$\begin{aligned}
A_S^2 &\equiv \mathcal{P}_{\delta_H} && (177)\\
&\simeq \frac{4}{25}\left(\frac{H^2}{2\pi\dot{\varphi}}\right)^2_{k=aH} && (178)\\
&\simeq \frac{512\pi G_N^3}{75}\left(\frac{V^3}{V'^2}\right)_{k=aH} && (179)\\
A_T^2 &\equiv \frac{1}{\rho_{\rm rad}}\frac{d\ln\rho_{\rm gw}}{d\ln k} && (180)\\
&\simeq \frac{32G_N^2}{75}\left(V\right)_{k=aH} && (181)
\end{aligned}$$

Note that the observed isotropy of the CMB places a direct upper limit on the amplitude of tensor fluctuations, $A_T^2 < 10^{-10}$, and hence the energy scale during inflation.

Observations can constrain not only the amplitude but also the tilt of the spectra, defined as

$$\begin{aligned}
n_S - 1 &\equiv \frac{d\ln A_S^2}{d\ln k} && (182)\\
n_T &\equiv \frac{d\ln A_T^2}{d\ln k} && (183)
\end{aligned}$$

The scale dependence of the spectra arises due to the time-dependence of the potential and its derivative when these modes leave the horizon during inflation, when $k = aH$, which we can translate into a dependence on the field value

$$\frac{d\ln k}{d\varphi} = \frac{d\ln aH}{d\varphi} \simeq \frac{H}{\dot{\varphi}} \tag{184}$$

and thus we find

$$\begin{aligned}
n_S - 1 &= -\frac{V'}{8\pi GV}\left(3\frac{V'}{V} - 2\frac{V''}{V'}\right) , && (185)\\
n_T &= -\frac{V'}{8\pi GV}\left(\frac{V'}{V}\right) . && (186)
\end{aligned}$$

These can be conveniently written in terms of the slow-roll parameters introduced (106)

$$\begin{aligned}
n_S - 1 &= -6\epsilon + 2\eta , && (187)\\
n_T &= -2\epsilon . && (188)
\end{aligned}$$

If we can measure n_S and n_T we can directly measure the slow-roll parameters during inflation.

This also offers a possible test of slow-roll models of (single-field) inflation, as the ratio between the scalar and tensor contributions to the large-angle cmb anisotropy is also determined by one of the slow-roll parameters:

$$\frac{A_T^2}{A_S^2} = -2\epsilon \tag{189}$$

Comparing this ratio to tilt of the tensor spectrum offers a potential test of the simple slow-roll scenario, but one that may prove impossible to test, unless tensors contribute an unexpectedly large fraction of the observed anisotropy. It is quite possible that there is a completely negligible contribution from tensors and the only information we can extract will be from the scalar sector.

6.5 Non-adiabatic perturbations from inflation

Consider N scalar fields [7] with Lagrangian density:

$$\mathcal{L} = -V(\varphi_1, \cdots, \varphi_N) - \frac{1}{2}\sum_{I=1}^{N} g^{\mu\nu}\varphi_{I,\mu}\varphi_{I,\nu}\,, \tag{190}$$

and minimal coupling to gravity. The field equations, derived from Eq. (190) for the background homogeneous fields, are

$$\ddot{\varphi}_I + 3H\dot{\varphi}_I + V_{\varphi_I} = 0\,, \tag{191}$$

where $V_x = \partial V/\partial x$, and the Hubble rate, H, in a spatially flat Friedmann-Robertson-Walker (FRW) universe, is determined by the Friedman equation:

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\left[V(\varphi_I) + \frac{1}{2}\sum_I \dot{\varphi}_I^{\,2}\right]\,, \tag{192}$$

with $a(t)$ the FRW scale factor.

Scalar field perturbations, with comoving wavenumber $k = 2\pi a/\lambda$ for a mode with physical wavelength λ, then obey the perturbation equations

$$\begin{aligned}\delta\ddot{\varphi}_I + 3H\delta\dot{\varphi}_I + \frac{k^2}{a^2}\delta\varphi_I + \sum_J V_{\varphi_I\varphi_J}\delta\varphi_J \\ = -2V_{\varphi_I}A + \dot{\varphi}_I\left[\dot{A} + 3\dot{\psi} + k^2\sigma\right]\,.\end{aligned} \tag{193}$$

The metric terms on the right-hand-side, induced by the scalar field perturbations, obey the energy and momentum constraints

$$3H\left(\dot{\psi} + HA\right) + \frac{k^2}{a^2}\left[\psi + Ha^2\sigma\right] = -4\pi G\delta\rho\,, \tag{194}$$

$$\dot{\psi} + HA = -4\pi G\delta q\,. \tag{195}$$

[7] This section is taken from Ref. [18].

The total energy and momentum perturbations are given in terms of the scalar field perturbations by

$$\begin{aligned} \delta\rho &= \sum_I \left[\dot{\varphi}_I\left(\dot{\delta\varphi_I} - \dot{\varphi}_I A\right) + V_{\varphi_I}\delta\varphi_I\right] && (196) \\ \delta q_{,i} &= -\sum_I \dot{\varphi}_I \delta\varphi_{I,i} \,. && (197) \end{aligned}$$

These two equations can be combined to construct a gauge-invariant quantity, the comoving density perturbation [1]

$$\begin{aligned} \epsilon_m &\equiv \delta\rho - 3H\delta q \\ &= \sum_I \left[\dot{\varphi}_I\left(\dot{\delta\varphi_I} - \dot{\varphi}_I A\right) - \ddot{\varphi}_I \delta\varphi_I\right] \,, && (198) \end{aligned}$$

which is sometimes used to represent the total matter perturbation.

Because the anisotropic stress vanishes to linear order for scalar fields minimally coupled to gravity, we have a further constraint on the metric perturbations:

$$a^2\left(\dot{\sigma} + 3H\sigma\right) + \psi - A = 0 \,. \qquad (199)$$

The coupled perturbation equations (193)–(197) and (199) are probably most often solved in the zero-shear (or longitudinal or conformal Newtonian) gauge, in which $\sigma_\ell = 0$ [4]. The two remaining metric perturbation variables which appear in the scalar field perturbation equation, $A_\ell = \Phi$ and $\psi_\ell = \Psi$, are then equal in the absence of any anisotropic stress by Eq. (199).

We have seen that another useful choice is the spatially flat gauge, in which $\psi_Q = 0$ [3, 28]. The scalar field perturbations in this gauge are sometimes referred to as the Sasaki or Mukhanov variables [20], which have the gauge-invariant definition

$$Q_I \equiv \delta\varphi_I + \frac{\dot{\varphi}_I}{H}\psi \,. \qquad (200)$$

The shear perturbation in the spatially flat gauge is simply related to the curvature perturbation, Ψ, in the zero-shear gauge:

$$\sigma_Q = \sigma + \frac{1}{a^2 H}\psi = \frac{1}{a^2 H}\Psi \,. \qquad (201)$$

The energy and momentum constraints, Eqs. (194) and (195), in the spatially flat gauge thus yield

$$\begin{aligned} \frac{k^2}{a^2}\Psi &= -4\pi G\epsilon_m \,, && (202) \\ HA_Q &= -4\pi G\delta q_Q \,, && (203) \end{aligned}$$

where ϵ_m is given in Eq. (198), and from Eq. (197) we have $\delta q_Q = -\sum_I \dot{\varphi}_I Q_I$.

The equations of motion, Eq. (193), rewritten in terms of the Sasaki-Mukhanov variables, and using Eqs. (202) and (203) to eliminate the metric perturbation terms in the spatially flat gauge, become [21]:

$$\begin{aligned}\ddot{Q}_I + 3H\dot{Q}_I + \frac{k^2}{a^2}Q_I \\ + \sum_J \left[V_{\varphi_I\varphi_J} - \frac{8\pi G}{a^3}\left(\frac{a^3}{H}\dot{\varphi}_I\dot{\varphi}_J\right)^{\cdot}\right] Q_J = 0\,. \end{aligned} \tag{204}$$

6.5.1 Curvature and entropy perturbations

The comoving curvature perturbation [22, 16] is given by

$$\begin{aligned}\mathcal{R} &\equiv \psi - \frac{H}{\rho+p}\delta q \\ &= \sum_I \left(\frac{\dot{\varphi}_I}{\sum_J \dot{\varphi}_J^2}\right) Q_I\,. \end{aligned} \tag{205}$$

This can also be given in terms of the metric perturbations in the longitudinal gauge as [4]

$$\mathcal{R} = \Psi - \frac{H}{\dot{H}}\left(\dot{\Psi} + H\Phi\right)\,. \tag{206}$$

For comparison we give the curvature perturbation on uniform-density hypersurfaces,

$$-\zeta \equiv \psi + H\frac{\delta\rho}{\dot{\rho}}\,, \tag{207}$$

first introduced by Bardeen, Steinhardt and Turner [7] as a conserved quantity for adiabatic perturbations on large scales [23, 17]. It is related to the comoving curvature perturbation in Eq. (205) by a gauge transformation

$$-\zeta = \mathcal{R} + \frac{2\rho}{3(\rho+p)}\left(\frac{k}{aH}\right)^2 \Psi\,, \tag{208}$$

where we have used to the constraint equation (202) to eliminate the comoving density perturbation, ϵ_m. Note that $\mathcal{R}$ and $-\zeta$ thus coincide in the limit $k \to 0$.

Both $\mathcal{R}$ and $-\zeta$ are commonly used to characterise the amplitude of adiabatic perturbations as both remain constant for purely adiabatic perturbations on sufficiently large scales as a direct consequence of local energy-momentum conservation [17], allowing one to relate the perturbation spectrum on large scales to quantities at the Hubble scale crossing during inflation in the simplest inflation models [7, 24].

A dimensionless definition of the total entropy perturbation (which is automatically gauge-invariant) is given by

$$\mathcal{S} = H\left(\frac{\delta p}{\dot{p}} - \frac{\delta\rho}{\dot{\rho}}\right)\,, \tag{209}$$

which can be extended to define a generalised entropy perturbation between any two matter quantities x and y:

$$\mathcal{S}_{xy} = H\left(\frac{\delta x}{\dot{x}} - \frac{\delta y}{\dot{y}}\right). \tag{210}$$

The total entropy perturbation in Eq. (209) for N scalar fields is given by

$$\mathcal{S} = \frac{2\dot{V}\sum_I \dot{\varphi}_I(\dot{\delta\varphi}_I - \dot{\varphi}_I A) + \left(\dot{V} + 3H\sum_J \dot{\varphi}_J^2\right)\delta V}{3\left(2\dot{V} + 3H\sum_J \dot{\varphi}_J^2\right)\sum_I \dot{\varphi}_I^2}, \tag{211}$$

where the perturbation in the total potential energy is given by $\delta V = \sum_I V_{\varphi_I}\delta\varphi_I$.

The change in $\mathcal{R}$ on large scales (i.e., neglecting spatial gradient terms) can be directly related to the non-adiabatic part of the pressure perturbation [11, 17, 25]

$$\dot{\mathcal{R}} \approx -3H\frac{\dot{p}}{\dot{\rho}}\mathcal{S}. \tag{212}$$

We will thus now consider the evolution of the adiabatic and entropy perturbations in both one- and two-field models of inflation.

6.5.2 Single field

Perturbations in a single self-interacting scalar field obey the gauge-dependent equation of motion

$$\begin{aligned} &\ddot{\delta\varphi} + 3H\dot{\delta\varphi} + \left(\frac{k^2}{a^2} + V_{\varphi\varphi}\right)\delta\varphi \\ &\quad = -2V_\varphi A + \dot{\varphi}\left[\dot{A} + 3\dot{\psi} + \frac{k^2}{a^2}(a^2\dot{E} - aB)\right], \end{aligned} \tag{213}$$

subject to the energy and momentum constraint equations given in Eqs. (194) and (195), where the energy and momentum perturbations for a single scalar field are given by

$$\delta\rho = \dot{\varphi}\left(\dot{\delta\varphi} - \dot{\varphi}A\right) + V_\varphi\delta\varphi, \tag{214}$$

$$\delta q_{,i} = -\dot{\varphi}\delta\varphi_{,i}. \tag{215}$$

The scalar field perturbation in the spatially flat gauge has the gauge-invariant definition, Eq. (200),

$$Q_\varphi \equiv \delta\varphi + \frac{\dot{\varphi}}{H}\psi. \tag{216}$$

For a single field this is directly related to the curvature perturbation in the comoving gauge, where the momentum vanishes and hence, from Eq. (215), $\delta\varphi_m = 0$,

$$\mathcal{R} = \psi + \frac{H}{\dot{\varphi}}\delta\varphi = \frac{H}{\dot{\varphi}}Q_\varphi. \tag{217}$$

From Eq. (198) we have the comoving density perturbation

$$\epsilon_m = \dot\varphi\left(\dot{\delta\varphi} - \dot\varphi A\right) - \ddot\varphi\delta\varphi\,, \tag{218}$$

and hence, from Eqs. (214) and (215), we obtain the constraints

$$\begin{aligned} \frac{k^2}{a^2}\Psi &= -4\pi G\left[\dot\varphi\left(\dot{\delta\varphi} - \dot\varphi A\right) - \ddot\varphi\delta\varphi\right]\,, \qquad (219)\\ A_Q &= 4\pi G\frac{\dot\varphi}{H}Q_\varphi\,. \qquad (220) \end{aligned}$$

Substituting these expressions for the metric perturbations into Eq. (213), yields the decoupled equation of motion for the scalar field perturbation,

$$\ddot Q_\varphi + 3H\dot Q_\varphi + \left[\frac{k^2}{a^2} + V_{\varphi\varphi} - \frac{8\pi G}{a^3}\left(\frac{a^3}{H}\dot\varphi^2\right)^{\displaystyle\cdot}\right]Q_\varphi = 0\,. \tag{221}$$

It is not obvious that the intrinsic entropy perturbation for a single scalar field, obtained from Eq. (211),

$$\mathcal{S} = \frac{2V_\varphi}{3\dot\varphi^2(3H\dot\varphi + 2V_\varphi)}\left[\dot\varphi\left(\dot{\delta\varphi} - \dot\varphi A\right) - \ddot\varphi\delta\varphi\right]\,, \tag{222}$$

should vanish on large scales. Because the scalar field obeys a second-order equation of motion, its general solution contains two arbitrary constants of integration, which can describe both adiabatic and entropy perturbations. However $\mathcal{S}$ for a single scalar field is proportional to the comoving density perturbation given in Eq. (218), and this in turn is related to the metric perturbation, Ψ, via Eq. (202), so that [26]

$$\mathcal{S} = -\frac{V_\varphi}{6\pi G\dot\varphi^2[3H\dot\varphi + 2V_\varphi]}\left(\frac{k^2}{a^2}\Psi\right)\,. \tag{223}$$

In the absence of anisotropic stresses, Ψ must be of order A_Q, by Eq. (199), and hence the non-adiabatic pressure becomes small on large scales [29, 26, 9]. The amplitude of the asymptotic solution for the scalar field at late times (and hence large scales) during inflation thus determines the amplitude of an adiabatic perturbation.

The change in the comoving curvature perturbation is given by

$$\dot{\mathcal{R}} = \frac{H}{\dot H}\frac{k^2}{a^2}\Psi\,, \tag{224}$$

and hence the rate of change of the curvature perturbation, given by $d\ln\mathcal{R}/d\ln a \sim (k/aH)^2$, becomes negligible on large scales during single-field inflation.

6.5.3 Two fields

In this section we will consider two interacting scalar fields, $\phi \equiv \varphi_1$ and $\chi \equiv \varphi_2$. The analysis developed here should be straightforward to extend to include additional scalar fields, but we do not expect to see any qualitatively new features in this case, so for clarity we restrict our discussion here to two fields.

In order to clarify the role of adiabatic and entropy perturbations, their evolution and their inter-relation, we define new adiabatic and entropy fields by a rotation in field space. The "adiabatic field", Σ, represents the evolution along the classical trajectory, such that

$$d\Sigma = (\cos\theta)d\phi + (\sin\theta)d\chi\,, \tag{225}$$

where

$$\cos\theta = \frac{\dot\phi}{\sqrt{\dot\phi^2+\dot\chi^2}}, \quad \sin\theta = \frac{\dot\chi}{\sqrt{\dot\phi^2+\dot\chi^2}}\,. \tag{226}$$

This definition, plus the original equations of motion for ϕ and χ, give

$$\ddot\Sigma + 3H\dot\Sigma + V_\Sigma = 0\,, \tag{227}$$

where

$$V_\Sigma = (\cos\theta)V_\phi + (\sin\theta)V_\chi\,. \tag{228}$$

$\delta\Sigma$ is the component of the two-field perturbation vector along the direction of the background fields' evolution. The parameter Σ acts as an affine parameter for the background solution. Conversely, fluctuations orthogonal to the background classical trajectory represent non-adiabatic perturbations, and we define the "entropy field", s, such that

$$ds = (\cos\theta)d\chi - (\sin\theta)d\phi\,. \tag{229}$$

From this definition, it follows that $\dot s = 0$ along the classical trajectory, and hence entropy perturbations are automatically gauge-invariant [27]. Perturbations in $\delta\Sigma$, with $\delta s = 0$, describe adiabatic field perturbations, and this is why we refer to Σ as the "adiabatic field".

The total momentum of the two-field system, given by Eq. (197), is then

$$\delta q_{,i} = -\dot\phi\delta\phi_{,i} - \dot\chi\delta\chi_{,i} = -\dot\Sigma\delta\Sigma_{,i}\,, \tag{230}$$

and the comoving curvature perturbation in Eq. (205) is given by

$$\begin{aligned}\mathcal{R} &= \psi + H\left(\frac{\dot\phi\delta\phi+\dot\chi\delta\chi}{\dot\phi^2+\dot\chi^2}\right)\,,\\ &= \psi + \frac{H}{\dot\Sigma}\delta\Sigma\,.\end{aligned} \tag{231}$$

These expressions, written in terms of the adiabatic field, Σ, are identical to those given in Eqs. (215) and (217) for a single field.

We can also write Eq. (231) as

$$\mathcal{R} = (\cos^2\theta)\mathcal{R}_\phi + (\sin^2\theta)\mathcal{R}_\chi\,, \tag{232}$$

where we define the comoving curvature perturbation for each of the original fields as

$$\mathcal{R}_I \equiv \psi + \frac{H}{\dot{\varphi}_I}\delta\varphi_I = \frac{H}{\dot{\varphi}_I}Q_I\,. \tag{233}$$

However, even fields with no explicit interaction will in general have non-zero intrinsic entropy perturbations on large scales in a multi-field system due to their gravitational interaction, so that $\mathcal{R}_I$ for each field is not conserved. Although the intrinsic entropy perturbation for each field is still of the form given by Eq. (222), it is no longer constrained by Eq. (202) to vanish as $k \to 0$. This is in contrast to the case of non-interacting perfect fluids, where it is possible to define a constant curvature perturbation for each fluid on large scales [17].

The comoving matter perturbation in Eq. (198) can be written as

$$\epsilon_m = \dot{\Sigma}\left(\delta\dot{\Sigma} - \dot{\Sigma}A\right) - \ddot{\Sigma}\delta\Sigma + 2V_s\delta s\,, \tag{234}$$

which acquires an additional term, compared with the single-field case, due to the dependence of the potential upon s, where

$$V_s = (\cos\theta)V_\chi - (\sin\theta)V_\phi\,. \tag{235}$$

The perturbed kinetic energy of s has no contribution to first-order as in the background solution $\dot{s} = 0$, by definition.

The total entropy perturbation, Eq. (211), for the two fields can be written as

$$\begin{aligned}\mathcal{S} = {} & \frac{2}{3\dot{\Sigma}^2(3H\dot{\Sigma} + 2V_\Sigma)} \times \\ & \times \left\{V_\Sigma\left[\dot{\Sigma}\left(\delta\dot{\Sigma} - \dot{\Sigma}A\right) - \ddot{\Sigma}\delta\Sigma\right] + 3H\dot{\Sigma}^2\dot{\theta}\delta s\right\}\,.\end{aligned} \tag{236}$$

Combining Eqs. (202), (234) and (236), we can write

$$\mathcal{S} = -\frac{V_\Sigma}{6\pi G\dot{\Sigma}^2[3H\dot{\Sigma} + 2V_\Sigma]}\left(\frac{k^2}{a^2}\Psi\right) - \frac{2V_s}{3\dot{\Sigma}^2}\delta s\,. \tag{237}$$

Comparing this with the single-field result given in Eq. (223), we see that the entropy perturbation on large scales is due solely to the relative entropy perturbation between the two fields, described by the entropy field δs.

The change in the comoving curvature perturbation given by Eq. (238) rate of change is given by [11, 25]

$$\dot{\mathcal{R}} = \frac{H}{\dot{H}}\frac{k^2}{a^2}\Psi + \frac{1}{2}H\left(\frac{\delta\phi}{\dot{\phi}} - \frac{\delta\chi}{\dot{\chi}}\right)\frac{d}{dt}\left(\frac{\dot{\phi}^2 - \dot{\chi}^2}{\dot{\phi}^2 + \dot{\chi}^2}\right)\,, \tag{238}$$

which can be expressed neatly in terms of the new variables:

$$\dot{\mathcal{R}} = \frac{H}{\dot{H}} \frac{k^2}{a^2} \Psi + \frac{2H}{\dot{\Sigma}} \dot{\theta} \delta s \,, \tag{239}$$

where

$$\dot{\theta} = -\frac{V_s}{\dot{\Sigma}} \,. \tag{240}$$

The new source term on the right-hand-side of this equation, compared with the single-field case, Eq. (224), is proportional to the relative entropy perturbation between the two fields, δs. Clearly, there can be significant changes to $\mathcal{R}$ on large scales if the entropy perturbation is not suppressed and if the background solution follows a curved trajectory, i.e., $\dot{\theta} \neq 0$, in field space [9]. This can then produce a change in the comoving curvature on arbitrarily large scales (i.e., even in the limit $k \to 0$) [11, 26].

Equations of motion for the adiabatic and entropy field perturbations can be derived from the perturbed scalar field equations (193), to give

$$\begin{aligned} & \delta\ddot{\Sigma} + 3H\delta\dot{\Sigma} + \left(\frac{k^2}{a^2} + V_{\Sigma\Sigma} - \dot{\theta}^2 \right) \delta\Sigma \\ & \quad = -2V_\Sigma A + \dot{\Sigma} \left[\dot{A} + 3\dot{\psi} + \frac{k^2}{a^2}(a^2\dot{E} - aB) \right] \\ & \qquad + 2(\dot{\theta}\delta s)^{\cdot} - 2\frac{V_\Sigma}{\dot{\Sigma}} \dot{\theta} \delta s \,, \end{aligned} \tag{241}$$

and

$$\begin{aligned} & \ddot{\delta s} + 3H\dot{\delta s} + \left(\frac{k^2}{a^2} + V_{ss} - \dot{\theta}^2 \right) \delta s \\ & \quad = -2\frac{\dot{\theta}}{\dot{\Sigma}} \left[\dot{\Sigma}(\delta\dot{\Sigma} - \dot{\Sigma}A) - \ddot{\Sigma}\delta\Sigma \right] \,, \end{aligned} \tag{242}$$

where

$$V_{\Sigma\Sigma} = (\sin^2\theta) V_{\chi\chi} + (\sin 2\theta) V_{\phi\chi} + (\cos^2\theta) V_{\phi\phi} \,, \tag{243}$$

$$V_{ss} = (\sin^2\theta) V_{\phi\phi} - (\sin 2\theta) V_{\phi\chi} + (\cos^2\theta) V_{\chi\chi} \,. \tag{244}$$

When $\dot{\theta} = 0$, the adiabatic and entropy perturbations decouple[8] . The equation of motion for $\delta\Sigma$ then reduces to that for a single scalar field in a perturbed FRW spacetime, as given in Eq. (213), while the equation for δs is that for a scalar field perturbation in an *unperturbed* FRW spacetime.

[8] If we employ the slow-roll approximation for the background fields, $\dot{\phi} \simeq -V_\phi/3H$ and $\dot{\chi} \simeq -V_\chi/3H$, we obtain $\dot{\theta} \simeq 0$. This reflects the fact that the rate of change of θ is slow – instantaneously it moves in an approximately straight line in field space. But the integrated change in θ cannot in general be neglected. Even working within the slow-roll approximation, fields do not in general follow a straight line trajectory in field space.

The only source term on the right-hand-side in Eq. (242) for the entropy perturbation comes from the intrinsic entropy perturbation in the Σ-field. From Eqs. (202) and (234) we have

$$\dot{\Sigma}(\dot{\delta\Sigma} - \dot{\Sigma}A) - \ddot{\Sigma}\delta\Sigma = 2\dot{\Sigma}\dot{\theta}\delta s - \frac{k^2}{4\pi G a^2}\Psi\,, \tag{245}$$

and hence we can rewrite the evolution equation for the entropy perturbation as

$$\ddot{\delta s} + 3H\dot{\delta s} + \left(\frac{k^2}{a^2} + V_{ss} + 3\dot{\theta}^2\right)\delta s = \frac{\dot{\theta}}{\dot{\Sigma}}\frac{k^2}{2\pi G a^2}\Psi\,. \tag{246}$$

Note that this evolution equation is automatically gauge-invariant and holds in any gauge. On large scales the inhomogeneous source term becomes negligible, and we have a homogeneous second-order equation of motion for the entropy perturbation, decoupled from the adiabatic field and metric perturbations. If the initial entropy perturbation is zero on large scales, it will remain so.

By contrast, we cannot neglect the metric back-reaction for the adiabatic field fluctuations, or the source terms due to the entropy perturbations. Working in the spatially flat gauge, defining

$$Q_\Sigma = \delta\Sigma_Q = \delta\Sigma + \frac{\dot{\Sigma}}{H}\psi\,, \tag{247}$$

and using

$$A_Q = 4\pi G\frac{\dot{\Sigma}}{H}Q_\Sigma\,, \tag{248}$$

we can rewrite the equation of motion for the adiabatic field perturbation as

$$\begin{aligned}\ddot{Q}_\Sigma + 3H\dot{Q}_\Sigma + \left[\frac{k^2}{a^2} + V_{\Sigma\Sigma} - \dot{\theta}^2 - \frac{8\pi G}{a^3}\left(\frac{a^3\dot{\Sigma}^2}{H}\right)^{\cdot}\right] Q_\Sigma \\ = 2(\dot{\theta}\delta s)^{\cdot} - 2\left(\frac{V_\Sigma}{\dot{\Sigma}} + \frac{\dot{H}}{H}\right)\dot{\theta}\delta s\,.\end{aligned} \tag{249}$$

When $\dot{\theta} = 0$, this reduces to the single-field equation (221), but for a curved trajectory in field space, the entropy perturbation acts as an additional source term in the equation of motion for the adiabatic field perturbation, even on large scales.

In order for small-scale quantum fluctuations to produce large-scale (super-Hubble) perturbations during inflation, a field must be "light" (i.e., overdamped). The effective mass for the entropy field in Eq. (246) is $\mu_s^2 = V_{ss} + 3\dot{\theta}^2$. For $\mu_s^2 > \frac{3}{2}H^2$, the fluctuations remain in the vacuum state and fluctuations on large scales are strongly suppressed. The existence of large-scale entropy perturbations therefore requires

$$\mu_s^2 \equiv V_{ss} + 3\dot{\theta}^2 < \frac{3}{2}H^2\,. \tag{250}$$

6.5.4 Entropy/adiabatic correlations from inflation

Equations (246) and (249) are the key equations which govern the evolution of the adiabatic and entropy perturbations in a two field system. Together with constraint equations (245) and (248) for the metric perturbations, they form a closed set of equations. They allow one to follow the effect on the adiabatic curvature perturbation due to the presence of entropy perturbations, absent in the single field model. This in turn will allow us to study the resulting correlations between the spectra of adiabatic and entropy perturbations produced on large-scales due to quantum fluctuations of the fields on small-scales during inflation.

A useful approximation commonly made when studying field perturbations during inflation, is to split the evolution of a given mode into a sub-Hubble regime ($k > aH$), in which the Hubble expansion is neglected, and a super-Hubble regime ($k < aH$), in which gradient terms are dropped.

If we assume that both fields ϕ and χ are light (i.e., overdamped) during inflation, then we can take the field fluctuations to be in their Minkowski vacuum state on sub-Hubble scales. This gives their amplitudes at Hubble crossing ($k = aH$) as

$$Q_I|_{k=aH} = \frac{H_k}{\sqrt{2k^3}}\, e_I(k)\,, \tag{251}$$

where $I = \phi, \chi$, H_k is the Hubble parameter when the mode crosses the Hubble radius (i.e., $H_k = k/a$), and e_ϕ and e_χ are independent Gaussian random variables satisfying

$$\langle e_I(k)\rangle = 0\,, \quad \langle e_I(k) e_J^*(k')\rangle = \delta_{IJ}\, \delta(k - k')\,, \tag{252}$$

with the angled brackets denoting ensemble averages. It follows from our definitions of the entropy and adiabatic perturbations in Eqs. (225) and (229) that their distributions at Hubble crossing have the same form:

$$Q_\Sigma|_{k=aH} = \frac{H_k}{\sqrt{2k^3}} e_\Sigma(k)\,, \quad \delta s|_{k=aH} = \frac{H_k}{\sqrt{2k^3}} e_s(k)\,, \tag{253}$$

where e_Σ and e_s are Gaussian random variables obeying the same relations given in Eq. (252), with $I, J = \Sigma, s$.

Super-Hubble modes are assumed to obey the equations of motion given in Eqs. (249) and (246), which we will write schematically as

$$\hat{O}^\Sigma(Q_\Sigma) = \hat{S}^\Sigma(\delta s)\,, \tag{254}$$

$$\hat{O}^s(\delta s) = 0\,, \tag{255}$$

where $\hat{O}^\Sigma(Q_\Sigma)$ and $\hat{O}^s(\delta s)$ are obtained by setting $k = 0$ on the left-hand side of Eqs. (249) and (246) respectively, and $\hat{S}^\Sigma(\delta s)$ is given by the right-hand side of Eq. (249). As remarked before, there is no source term for δs appearing on the right-hand side of Eq. (246) once we neglect gradient terms. The general super-Hubble solution can thus be written as

$$Q_\Sigma = A_+ f_+(t) + A_- f_-(t) + P(t)\,, \tag{256}$$

$$\delta s = B_+ g_+(t) + B_- g_-(t)\,, \tag{257}$$

where the real functions $f_\pm$ and $g_\pm$ are the growing/decaying modes of the homogeneous equations, $\hat{O}^\Sigma(f_\pm) = 0$ and $\hat{O}^s(g_\pm) = 0$, and $P(t)$ is a particular integral of the full inhomogeneous equation (254). During slow-roll evolution, the decaying modes can be neglected, and we take

$$\begin{aligned} Q_\Sigma &\simeq A_+ f_+(t) + P(t)\,, && (258)\\ \delta s &\simeq B_+ g_+(t)\,. && (259)\end{aligned}$$

Note that in the slow-roll approximation the homogeneous solution $f_+ \propto \dot{\Sigma}/H$. We can, without loss of generality, take $f_+ = 1 = g_+$ and $P = 0$ when $k = aH$, so that the amplitudes of the growing modes at Hubble-crossing are given by Eqs. (253) as

$$A_+(k) = \frac{H_k}{\sqrt{2k^3}} e_\Sigma(k)\,, \quad B_+(k) = \frac{H_k}{\sqrt{2k^3}} e_s(k)\,. \tag{260}$$

From Eq. (254), we see that the amplitude of the particular integral $P(t)$ at later times will be correlated with the amplitude of the entropy perturbation, B_+, and we can write $P(t) = B_+ \tilde{P}(t)$, where $\tilde{P}(t)$ is a real function independent of the random variables e_Σ, e_s.

In order to quantify the correlation, we define

$$\langle x(k) y^*(k') \rangle \equiv \frac{2\pi^2}{k^3}\, \mathcal{C}_{xy}\, \delta(k - k')\,. \tag{261}$$

The adiabatic and entropy power spectra are given by

$$\begin{aligned} \mathcal{P}_{Q_\Sigma} \equiv \mathcal{C}_{Q_\Sigma Q_\Sigma} &\simeq \left(\frac{H_k}{2\pi}\right)^2 \left[|f_+^2| + |\tilde{P}^2|\right]\,, && (262)\\ \mathcal{P}_{\delta s} \equiv \mathcal{C}_{\delta s \delta s} &\simeq \left(\frac{H_k}{2\pi}\right)^2 |g_+^2|\,, && (263)\end{aligned}$$

while the dimensionless cross-correlation is given by

$$\frac{\mathcal{C}_{Q_\Sigma \delta s}}{\sqrt{\mathcal{P}_{Q_\Sigma}}\sqrt{\mathcal{P}_{\delta s}}} \simeq \frac{g_+ \tilde{P}}{\sqrt{g_+^2}\sqrt{|f_+^2| + |\tilde{P}^2|}}\,. \tag{264}$$

Note that the adiabatic power spectrum at late times is always enhanced if it is coupled to entropy perturbations [i.e., $P(t) \neq 0$, in Eq. (258)], as the entropy field fluctuations at Hubble-crossing provide an uncorrelated extra source.

References

[1] J. M. Bardeen, Phys. Rev. D **22**, 1882 (1980).

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

[3] H. Kodama and M. Sasaki, Prog. Theor. Phys. Suppl. **78**, 1 (1984).

[4] V. F. Mukhanov, H. A. Feldman and R. H. Brandenberger, Phys. Rep. **215**, 203 (1992).

[5] M. Sasaki and E. D. Stewart, Prog. Theor. Phys. **95**, 71 (1996), astro-ph/9507001.

[6] M. Sasaki and T. Tanaka, Prog. Theor. Phys. **99**, 763 (1998), gr-qc/9801017.

[7] J. M. Bardeen, P. J. Steinhardt and M. S. Turner, Phys. Rev. D **28**, 679 (1983).

[8] J. M. Bardeen, in Cosmology and Particle Physics, eds. L. Fang and A. Zee, Gordon and Breach Science Publishers, New York (1988).

[9] D. H. Lyth and A. Riotto, Phys. Rep. **314**, 1 (1999), hep-ph/9807278.

[10] H. Kodama and M. Sasaki, Int. J. Mod. Phys. **A 2** 491 (1987).

[11] J. García-Bellido and D. Wands, Phys. Rev. D **53**, 5437 (1996), astro-ph/9511029.

[12] A. A. Starobinsky, JETP Lett. **42**, 51 (1985).

[13] D. S. Salopek, Phys. Rev. D **52**, 5563 (1995), astro-ph/9506146.

[14] T. Takahiro and E. D. Stewart, Phys. Lett. **B381**, 413 (1996), astro-ph/9604103.

[15] V. F. Mukhanov, JETP **68**, 1297 (1988).

[16] D. H. Lyth, Phys. Rev D **31**, 1792 (1985).

[17] D. Wands, K. A. Malik, D. H. Lyth and A. R. Liddle, Phys. Rev. D **62**, 043527 (2000) [arXiv:astro-ph/0003278].

[18] C. Gordon, D. Wands, B. A. Bassett and R. Maartens, Phys. Rev. D **63**, 023506 (2001) [arXiv:astro-ph/0009131].

[19] J. E. Lidsey, A. R. Liddle, E. W. Kolb, E. J. Copeland, T. Barreiro and M. Abney, Rev. Mod. Phys. **69**, 373 (1997) [arXiv:astro-ph/9508078].

[20] M. Sasaki, Prog. Theor. Phys. **76**, 1036 (1986); V. F. Mukhanov, Zh. Éksp. Teor. Fiz. **94**, 1 (1988) [Sov. Phys. JETP **68**, 1297 (1988)].

[21] A. Taruya and Y. Nambu, Phys. Lett. **B428**, 37 (1998).

[22] V. Lukash, Zh. Eksp. Teor. Fiz. **79**, 1601 (1980) [Sov. Phys. JETP **52**, 807 (1980)].

[23] J. Martin and D. J. Schwarz, Phys. Rev. **D57**, 3302 (1998) [gr-qc/9704049].

[24] A. R. Liddle and D. H. Lyth, Phys. Rept. **231**, 1 (1993) [astro-ph/9303019].

[25] F. Finelli and R. Brandenberger, hep-ph/0003172.

[26] B. A. Bassett, F. Tamburini, D. I. Kaiser, and R. Maartens, Nucl. Phys. **B561**, 188 (1999) [hep-ph/9901319].

[27] J.M. Stewart and M. Walker, Proc. R. Soc. Lond. A **341**, 49 (1974).

[28] J. Hwang, Astrophys. J. **427**, 542 (1994).

[29] M. Sasaki and E. D. Stewart, Prog. Theor. Phys. **95**, 71 (1996) [astro-ph/9507001].