Cosmology I

Hannu Kurki-Suonio

Fall 2018

Preface

These are the lecture notes for my Cosmology course at the University of Helsinki. I first lectured Cosmology at the Helsinki University of Technology in 1996 and then at University of Helsinki from 1997 to 2009. Syksy Räsänen taught the course from 2010 to 2015. I have lectured the course again since 2016. These notes are based on my notes from 2009, but I have adopted some improvements made by Syksy. As the course progresses I will keep updating the lecture notes.

A difficulty in teaching cosmology is that some very central aspects of modern cosmology rely on rather advanced physics, like quantum field theory in curved spacetime. On the other hand, the main applications of these aspects can be discussed in relatively simple terms, so requiring students to have such background would seem overkill, and would prevent many interested students in getting a taste of this exciting and important subject. Thus I am assuming just the standard bachelor level theoretical physics background (mechanics, special relativity, quantum mechanics, statistical physics). The more advanced theories that cosmology relies on, general relativity and quantum field theory, are reviewed as a part of this course to a sufficient extent, that we can go on. This represents a compromise which requires from the student an acceptance of some results without a proper derivation. Even a quantum mechanics or statistical physics background is not necessary, if the student is willing to accept some results taken from these fields (in the beginning of Chapter 4). As mathematical background, Cosmology I requires integral and differential calculus (as taught in Matemaattiset apuneuvot I, II). Cosmology II requires also Fourier analysis and spherical harmonic analysis (Fysiikan matemaattiset menetelmät I, II).

The course is divided into two parts. In Cosmology I, the universe is discussed in terms of the homogeneous and isotropic approximation (the Friedmann–Robertson–Walker model), which is good at the largest scales and in the early universe. In Cosmology II, deviations from this homogeneity and isotropy, i.e., the structure of the universe, are discussed. I thank Elina Keihänen, Jussi Väliviita, Ville Heikkilä, Reijo Keskitalo, and Elina Palmgren for preparing some of the figures and doing the calculations behind them.

– Hannu Kurki-Suonio, December 2017

1 Introduction

Cosmology is the study of the universe as a whole, its structure, its origin, and its evolution.

Cosmology is based on observations, mostly astronomical, and laws of physics. These lead naturally to the standard framework of modern cosmology, the *Hot Big Bang*.

As a science, cosmology has a severe restriction: there is only one universe.[1] We cannot make experiments in cosmology, and observations are restricted to a single object: the Universe. Thus we can make no comparative or statistical studies among many universes. Moreover, we are restricted to observations made from a single location, our solar system. It is quite possible that due to this special nature of cosmology, some important questions can never be answered.

Nevertheless, the last few decades have seen a remarkable progress in cosmology, as a significant body of relevant observational data has become available with modern astronomical instruments. We now have a good understanding of the overall history[2] and structure of the universe, but important open questions remain, e.g., the nature of *dark matter* and *dark energy*. Hopefully observations with more advanced instruments will resolve many of these questions in the coming decades.

The fundamental observation behind the big bang theory was the *redshift* of distant galaxies. Their spectra are shifted towards longer wavelengths. The further out they are, the larger is the shift. This implies that they are receding away from us; the distance between them and us is increasing. According to general relativity, we understand this as the expansion of the intergalactic space itself, not as actual motion of the galaxies. As the space expands, the wavelength of light traveling through space expands also.[3]

This expansion appears to be uniform over large scales: the whole universe expands at the same rate.[4] We describe this expansion by a time-dependent *scale factor*, $a(t)$. Starting from the observed present value of the expansion rate, $H \equiv(\mathrm{da/dt})/\mathrm{a}\equiv \dot{a}/a$, and knowledge of the energy content of the universe, we can use general relativity to calculate $a(t)$ as a function of time. The result is, using the standard model of particle physics for the energy content at high energies, that $a(t) \to 0$ about 14 billion years ago (I use the American convention, adopted now also by the British, where billion $\equiv 10^9$). At this *singularity*, the "beginning" of the big bang, which we choose as the origin of our time coordinate, $t = 0$, the density of the universe $\rho \to \infty$. In reality, we do not expect the standard model of particle physics to be applicable at extremely high energy densities. Thus there should be modifications to this picture at the very earliest times, probably just within the first nanosecond. A popular modification, discussed in Cosmology II, is cosmological *inflation*, which extends these earliest times, possibly, like in the "eternal inflation" model, infinitely (although usually inflation is thought to last only a small fraction of a second). At the least, when the density becomes comparable to the so called *Planck density*, $\rho_{\mathrm{Pl}} \sim 10^{96}\,\mathrm{kg/m^3}$, quantum gravitational effects should be large, so that general relativity itself

[1] There may, in principle, exist other universes, but they are not accessible to our observation. We spell Universe with a capital letter when we refer specifically to the universe we live in, wheres we spell it without a capital letter, when we refer to the more general or theoretical concept of the universe. In Finnish, 'maailmankaikkeus' is not capitalized.

[2] Except for the very beginning.

[3] These are not the most fundamental viewpoints. In general relativity the universe is understood as a four-dimensional curved spacetime, and its separation into space and time is a coordinate choice, based on convenience. The concepts of expansion of space and photon wavelength are based on such a coordinate choice. The most fundamental aspect is the curvature of spacetime. At large scales, the spacetime is curved in such a way that it is convenient to view this curvature as expansion of space, and in the related coordinate system the photon wavelength is expanding at the corresponding rate.

[4] This applies only at distance scales larger than the scale of galaxy clusters, about 10 Mpc. Bound systems, e.g., atoms, chairs, you and me, the Earth, the solar system, galaxies, or clusters of galaxies, do not expand. The expansion is related to the overall averaged gravitational effect of all matter in the universe. Within bound systems local gravitational effects are much stronger, so this overall effect is not relevant.

is no longer valid. To describe this *Planck era*, we would need a theory of *quantum gravity*, which we do not have.[5] Thus these earliest times, including $t = 0$, have to be excluded from the scientific big bang theory. Nevertheless, when discussing the universe after the Planck era and/or after inflation we customarily set the origin of the time coordinate $t = 0$, where the standard model solution would have the singularity.

Thus the proper way to understand the term "big bang", is not as some event by which the universe started or came into existence, but as a period in the early universe, when the universe was very hot,[6] very dense, and expanding rapidly.[7] Moreover, the universe was then filled with an almost homogeneous "primordial soup" of particles, which was in thermal equilibrium for a long time. Therefore we can describe the state of the early universe with a small number of thermodynamic variables, which makes the time evolution of the universe calculable.

1.1 Misconceptions

There are some popular misconceptions about the big bang, which we correct here:

The universe did not start from a point. The part of the universe which we can observe today was indeed very small at very early times, possibly smaller than 1 mm in diameter at the earliest times that can be sensibly discussed within the big bang framework. And if the inflation scenario is correct, even very much smaller than that before (or during earlier parts of) inflation, so in that sense the word "point" may be appropriate. But the universe extends beyond what can be observed today (beyond our "horizon"), and if the universe is infinite (we do not know whether the universe is finite or infinite), in current models it has always been infinite, from the very beginning.

As the universe expands it is not expanding into some space "around" the universe. The universe contains *all* space, and this space itself is "growing larger".[8]

1.2 Units and terminology

We shall use natural units where $c = \hbar = k_B = 1$.

1.2.1 $c = 1$

Relativity theory unifies space and time into a single concept, the 4-dimensional spacetime. It is thus natural to use the same units for measuring distance and time. Since the (vacuum) speed of light is $c = 299\ 792\ 458\,\mathrm{m/s}$, we set $1\ \mathrm{s} \equiv 299\ 792\ 458\,\mathrm{m}$, so that 1 second = 1 light second, 1 year = 1 light year, and $c = 1$. Velocity is thus a dimensionless quantity, and smaller than one[9] for massive objects. Energy and mass have now the same dimension, and Einstein's famous equivalence relation between mass and energy, $E = mc^2$, becomes $E = m$. This justifies a change in terminology; since mass and energy are the same thing, we do not waste two words on it. As is customary in particle physics we shall use the word "energy", E, for the above

[5] *String theory* is a candidate for the theory of quantum gravity. It is, however, very difficult to calculate definite predictions for the very early universe from string theory. This is a very active research area at present, but remains quite speculative.

[6] The realization that the early universe must have had a high temperature did not come immediately after the discovery of the expansion. The results of big bang nucleosynthesis and the discovery of the cosmic microwave background are convincing evidence that the Big Bang was Hot.

[7] There is no universal agreement among cosmologists about what time period the term "big bang" refers to. My convention is that it refers to the time from the end of inflation (or from whenever the standard hot big bang picture becomes valid) until recombination, so that it is actually a 380-000-year-long period, still short compared to the age of the universe.

[8] If the universe is infinite, we can of course not apply this statement to the volume of the entire universe, which is infinite, but it applies to finite parts of the universe.

[9] In the case of "physical" (as opposed to "coordinate") velocities.

quantity. By the word "mass", m, we mean the rest mass. Thus we do not write $E = m$, but $E = m\gamma$, where $\gamma = 1/\sqrt{1-v^2}$. The difference between energy and mass, $E - m$, is the kinetic energy of the object.[10]

1.2.2 $k_B = 1$

Temperature, T, is a parameter describing a thermal equilibrium distribution. The formula for the equilibrium occupation number of energy level E includes the exponential form $e^{\beta E}$, where the parameter $\beta = 1/k_B T$. The only function of the Boltzmann constant, $k_B = 1.3805 \times 10^{-23}\,\mathrm{J/K}$, is to convert temperature into energy units. Since we now decide to give temperatures directly in energy units, k_B becomes unnecessary. We define $1\,\mathrm{K} = 1.3806 \times 10^{-23}\,\mathrm{J}$, or

$$1\,\mathrm{eV} = 11600\,\mathrm{K} = 1.78 \times 10^{-36}\,\mathrm{kg} = 1.60 \times 10^{-19}\,\mathrm{J}. \tag{1}$$

Thus $k_B = 1$, and the exponential form is just $e^{E/T}$.

1.2.3 $\hbar = 1$

The third simplification in the natural system of units is to set the Planck constant $\hbar \equiv h/2\pi = 1$. This makes the dimension of mass and energy 1/time or 1/distance. This time and distance give the typical time and distance scales quantum mechanics associates with the particle energy. For example, the energy of a photon $E = \hbar\omega = \omega = 2\pi\nu$ is equal to its angular frequency. We have

$$1\,\mathrm{eV} = 5.07 \times 10^{6}\,\mathrm{m}^{-1} = 1.52 \times 10^{15}\,\mathrm{s}^{-1}. \tag{2}$$

A useful relation to remember is

$$\hbar = 197\,\mathrm{MeV\,fm} = 1, \tag{3}$$

where we have the energy scale $\sim 200\,\mathrm{MeV}$ and length scale $\sim 1\,\mathrm{fm}$ of strong interactions.

Equations become now simpler and the physical relations more transparent, since we do not have to include the above fundamental constants. This is not a completely free lunch, however; we often have to do conversions among the different units to give our answers in familiar units.

1.2.4 Astronomical units

A common unit of mass and energy is the solar mass, $\mathrm{M}_\odot = 1.99 \times 10^{30}\,\mathrm{kg}$, and a common unit of length is parsec, $1\,\mathrm{pc} = 3.26$ light years $= 3.09 \times 10^{16}\,\mathrm{m}$. One parsec is defined as the distance from which 1 astronomical unit (AU, the distance between the Earth and the Sun) forms an angle of one arcsecond, $1''$. More common in cosmology is $1\,\mathrm{Mpc} = 10^6\,\mathrm{pc}$, which is a typical distance between neighboring galaxies. For angles, 1 degree $(1^\circ) = 60$ arcminutes $(60') = 3600$ arcseconds $(3600'')$.

1.3 Brief History of the Early Universe

Because of the high temperature, particles had large energies in the early universe. To describe matter in that era, we need particle physics. The standard model of particle physics is called $\mathrm{SU}(3)_c \otimes \mathrm{SU}(2)_w \otimes \mathrm{U}(1)_y$, which describes the symmetries of the theory. From the viewpoint of the standard model, we live today in a low-energy universe, where many of the symmetries of the theory are broken. The "natural" energy scale of the theory is reached when the temperature of the universe exceeds $100\,\mathrm{GeV}$ (about $10^{15}\,\mathrm{K}$), which was the case when the universe was younger than $10^{-11}\,\mathrm{s}$. Then the primordial soup of particles consisted of free massless fermions (quarks

[10] The talk about "converting mass to energy" or *vice versa* can be understood to refer to conversion of rest mass into kinetic energy.

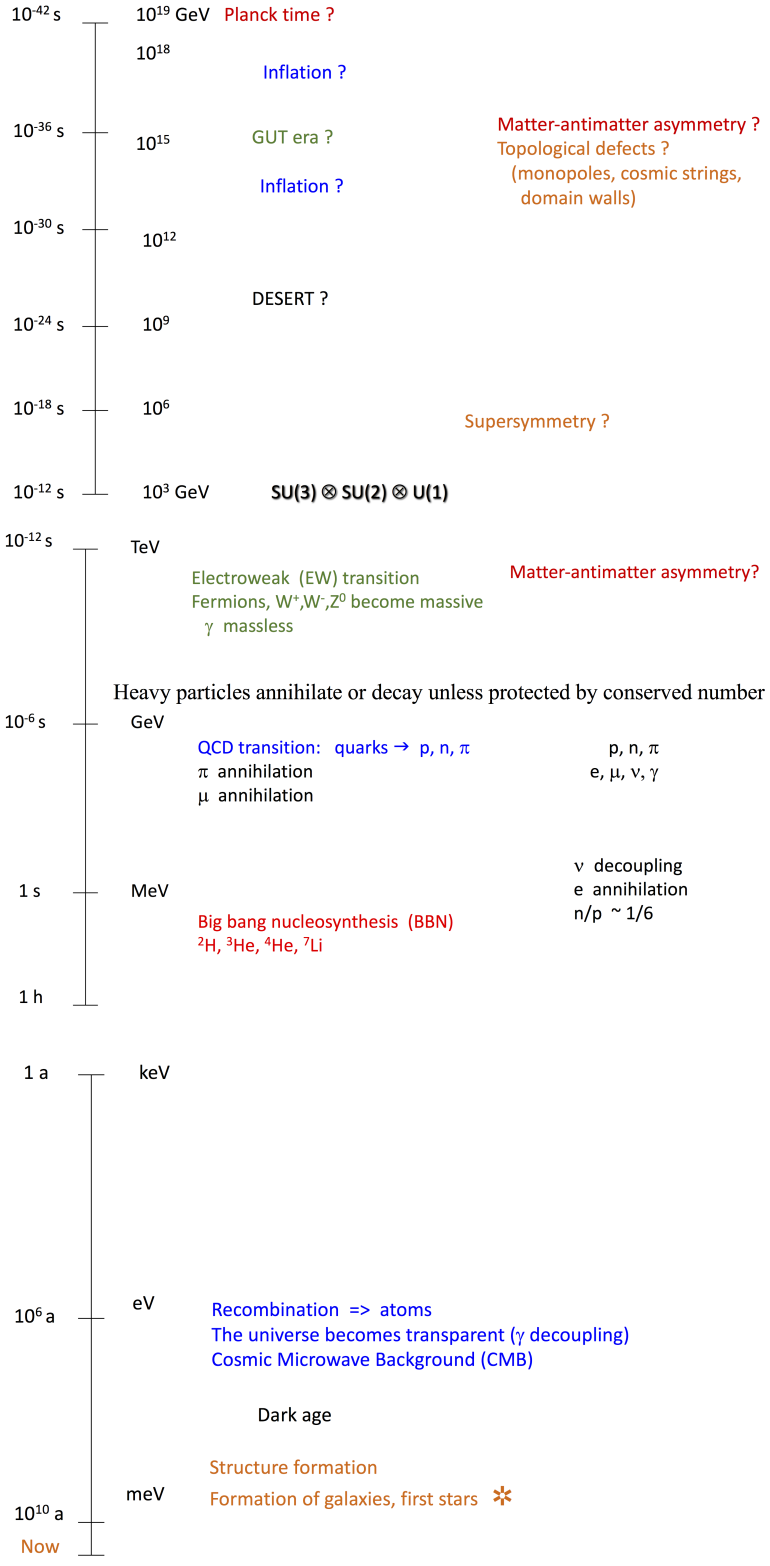


Figure 1: Short history of the universe.

and leptons) and massless gauge bosons mediating the interactions (color and electroweak) between these fermions. The standard model also includes a particle called the *Higgs boson*.

Higgs boson is responsible for the breaking of the electroweak (the $\mathrm{SU}(2)_w \otimes \mathrm{U}(1)_y$) symmetry. This is one of the *phase transitions*[11] in the early universe. In the electroweak (EW) phase transition the electroweak interaction becomes two separate interactions: 1) the weak interaction mediated by the massive gauge bosons $W^\pm$ and Z^0, and 2) the electromagnetic interaction mediated by the massless gauge boson γ, the photon. Fermions acquire their masses in the EW phase transition.[12] The mass is due to the interaction of the particle with the Higgs field. The EW phase transition took place when the universe cooled below the critical temperature $T_c \sim 100\,\mathrm{GeV}$ of the phase transition at $t \sim 10^{-11}\,\mathrm{s}$. See Fig. 1.

In addition to the standard model particles, the universe contains dark matter particles, whose exact nature is unknown. These will be discussed later, but we ignore them now for a while.

Another phase transition, the QCD (quantum chromodynamics) transition, or the quark–hadron transition, took place at $t \sim 10^{-5}\,\mathrm{s}$. The critical temperature of the QCD phase transition is $T_c \sim 150\,\mathrm{MeV}$. Quarks, which had been free until this time, formed hadrons: baryons, e.g., the nucleons n and p, and mesons, e.g., π, K. The matter filling the universe was converted from a quark–gluon plasma to a hadron gas.

To every type of particle there is a corresponding *antiparticle*, which has the same properties (e.g., mass and spin) as the particle, but its charges, like electric charge and color charge, have opposite sign. Particles which have no charges, like photons, are their own antiparticles. At high temperatures, $T \gg m$, where m is the mass of the particle, particles and antiparticles are constantly created and annihilated in various reactions, and there is roughly the same number of particles and antiparticles. But when $T \ll m$, particles and antiparticles may still annihilate each other (or decay, if they are unstable), put there is no more thermal production of particle-antiparticle pairs. As the universe cools, heavy particles and antiparticles therefore annihilate each other. These annihilation reactions produce additional lighter particles and antiparticles. If the universe had had an equal number of particles and antiparticles, only photons and neutrinos (of the known particles) would be left over today. The presence of matter today indicates that in the early universe there must have been slightly more nucleons and electrons than antinucleons and positrons, so that this excess was left over. The lightest known massive particle with strong or electroweak interactions is the electron,[13] so the last annihilation event was the electron-positron annihilation which took place when $T \sim m_\mathrm{e} \sim 0.5\,\mathrm{MeV}$ and $t \sim 1\,\mathrm{s}$. After this the only remaining antiparticles were the antineutrinos, and the primordial soup consisted of a large number of photons (who are their own antiparticles) and neutrinos (and antineutrinos) and a smaller number of "left-over" protons, neutrons, and electrons.

When the universe was a few minutes old, $T \sim 100\,\mathrm{keV}$, protons and neutrons formed nuclei of light elements. This event is known as Big Bang Nucleosynthesis (BBN), and it produced about 75% (of the total mass in ordinary matter) ^{1}H, 25% ^{4}He, 10^{-4} ^{2}H, 10^{-4} ^{3}He, and 10^{-9} ^{7}Li. (Other elements were formed much later, mainly in stars). At this time matter was completely ionized, all electrons were free. In this plasma the photons were constantly scattering from electrons, so that the mean free path of a photon between these scatterings was short. This means that the universe was opaque, not transparent to light.

The universe became transparent when it was 380 000 years old. At a temperature $T \sim$

[11] It may be that the EW and QCD phase transitions do not satisfy the technical definition of phase transition, but are instead just *cross-overs*, which means that they don't have a sharp critical temperature, but rather correspond to a temperature interval. The exact nature of these transitions is an open research problem.

[12] Except possibly neutrinos, the origin of whose masses in uncertain.

[13] According to observational evidence from *neutrino oscillations*, neutrinos also have small masses. However, at temperatures less than the neutrino mass, the neutrino interactions are so weak that the neutrinos and antineutrinos cannot annihilate each other.

3000 K ($\sim$ 0.25 eV), the electrons and nuclei formed neutral atoms, and the photon mean free path became longer than the radius of the observable universe. This event is called *recombination* (although it actually was the first combination of electrons with nuclei, not a *re*combination). Since recombination the primordial photons have been traveling through space mostly without scattering. We can observe them today as the *cosmic microwave background* (CMB). It is light from the early universe. We can thus "see" the big bang.

After recombination, the universe was filled with hydrogen and helium gas (with traces of lithium). The first stars formed from this gas when the universe was a few hundred million years old; but most of this gas was left as interstellar gas. The radiation from stars *reionized* the interstellar gas when the universe was 700 million years old.

1.4 Cosmological Principle

The ancients thought that the Earth is at the center of the Universe. This is an example of misconceptions that may result from having observations only from a single location (in this case, from the Earth). In the sixteenth century Nicolaus Copernicus proposed the heliocentric model of the universe, where Earth and the other planets orbited the Sun. This was the first step in moving "us" away from the center of the Universe. Later it was realized that neither the Sun, nor our galaxy, lies at the center of the Universe. This lesson has led to the *Copernican principle: We do not occupy a privileged position in the universe.* This is closely related to the *Cosmological principle: The universe is homogeneous and isotropic.*

Homogeneous means that all locations are equal, so that the universe appears the same no matter where you are. *Isotropic* means that all directions are equal, so that the universe appears the same no matter which direction you look at. Isotropy refers to isotropy with respect to some particular location, but 1) from isotropy with respect to one location and homogeneity follows isotropy with respect to every location, and 2) from isotropy with respect to all locations follows homogeneity.

There are two variants of the cosmological principle when applied to the real universe. As phrased above, it clearly does not apply at small scales: planets, stars, galaxies, and galaxy clusters are obvious inhomogeneities. In the first variant the principle is taken to mean that a homogeneous and isotropic model of the universe is a good approximation to the real universe at large scales (larger than the scale of galaxy clusters). In the second variant we add to this that the small-scale deviations from this model are *statistically homogeneous and isotropic.* This means that if we calculate the statistical properties of these inhomogeneities and anisotropies over a sufficiently large region, these statistical measures are the same for different such regions.

The Copernican principle is a philosophical viewpoint. Once you adopt it, observations lead to the first variant of the cosmological principle. CMB is highly isotropic and so is the distribution of distant galaxies, so we have solid observational support for isotropy with respect to our location. Direct evidence for homogeneity is weaker, but adopting the Copernican principle, we expect isotropy to hold also for other locations in the Universe, so that then the Universe should also be homogeneous. Thus we adopt the cosmological principle for the simplest model of the universe, which is an approximation to the true universe. This should be a good approximation at large scales, and in the early universe also for smaller scales.

The second variant of the cosmological principle cannot be deduced the same way from observations and the Copernican principle, but it follows naturally from the inflation scenario discussed in Cosmology II.

1.5 Structure Formation

CMB tells us that the early universe was very homogeneous, unlike the present universe, where matter has accumulated into stars and galaxies. The early universe had, however, very small

density variations, at the 10^{-5} to 10^{-3} level, which we see as small intensity variations of the CMB (the CMB *anisotropy*). Due to gravity, these slight overdensities have grown in time, and eventually became galaxies. This is called *structure formation* in the universe. The galaxies are not evenly distributed in space but form various structures, galaxy groups, clusters (large gravitationally bound groups), "filaments", and "walls", separated by large, relatively empty "voids". This present *large scale structure* of the universe forms a significant body of observational data in cosmology, which we can explain fairly well by cosmological theory.

There are two parts to structure formation:

1. The origin of the primordial density fluctuations, the "seeds of galaxies". These are believed to be due to some particle physics phenomenon in the very early universe, probably well before the EW transition. The particle physics theories applicable to this period are rather speculative. The currently favored explanation for the origin of primordial fluctuations is known as *inflation*. Inflation, discussed in Cosmology II, is not a specific theory, but it is a certain kind of behavior of the universe that could result from many different fundamental theories. Until the 1990s the main competitor to inflation was *topological defects*. Such defects (e.g., *cosmic strings*) may form in some phase transitions. The CMB data has ruled out topological defects at least as the main cause of structure formation.

2. The growth of these fluctuations as we approach the present time. The growth is due to gravity, but depends on the composition and total amount (average density) of matter and energy in the universe.

1.6 Dark Matter and Dark Energy

One of the main problems in cosmology today is that most of the matter and energy content of the universe appears to be in some unknown forms, called *dark matter* and *dark energy*. The dark matter problem dates back to 1930s, whereas the dark energy problem arose in late 1990s.

From the motions of galaxies we can deduce that the matter we can directly observe as stars and other "luminous matter" is just a small fraction of the total mass which affects the galaxy motions through gravity. The rest is dark matter, something which we observe only due to its *gravitational effect*. We do not know what most of this dark matter is. A smaller part of it is just ordinary, "baryonic", matter, which consists of atoms (or ions and electrons) just like stars, but does not shine enough for us to notice it. Possibilities include planet-like bodies in interstellar space, "failed" stars (too small, $m < 0.07\,\mathrm{M}_\odot$, to ignite thermonuclear fusion) called *brown dwarfs*, old white dwarf stars, and tenuous intergalactic gas. In fact, in large clusters of galaxies the intergalactic gas[14] is so hot that we can observe its radiation. Thus its mass can be estimated and it turns out to be several times larger than the total mass of the stars in the galaxies. We can infer that other parts of the universe, where this gas would be too cold to be observable from here, also contain significant amounts of thin gas; which thus is apparently the main component of *baryonic dark matter* (BDM). However, there is not nearly enough of it to explain the dark matter problem.

Beyond these mass estimates, there are more fundamental reasons (BBN, structure formation) why baryonic dark matter cannot be the main component of dark matter. Most of the dark matter must be non-baryonic, meaning that it is not made out of protons and neutrons[15]. The only non-baryonic particles in the standard model of particle physics that could act as dark matter, are neutrinos. If neutrinos had a suitable mass, ~ 1 eV, the neutrinos left from the early universe would have a sufficient total mass to be a significant dark matter component.

[14] This gas is ionized, so it should more properly be called *plasma*. Astronomers, however, often use the word "gas" also when it is ionized.

[15] And electrons. Although technically electrons are not baryons (they are leptons), cosmologists refer to matter made out of protons, electrons, and neutrons as "baryonic". The electrons are anyway so light, that most of the mass comes from the true baryons, protons and neutrons.

However, structure formation in the universe requires most of the dark matter to have different properties than neutrinos have. Technically, most of the dark matter must be "cold", instead of "hot". These are terms that just refer to the dynamics of the particles making up the matter, and do not further specify the nature of these particles. The difference between *hot dark matter* (HDM) and *cold dark matter* (CDM) is that HDM is made of particles whose velocities were large compared to escape velocities from the gravity of overdensities, when structure formation began, but CDM particles had small velocities. Neutrinos with $m \sim 1$ eV, would be HDM. An intermediate case is called *warm dark matter* (WDM). Structure formation requires that most of the dark matter is CDM, or possibly WDM, but the standard model of particle physics contains no suitable particles. Thus it appears that most of the matter in the universe is made out of some unknown particles.

Fortunately, particle physicists have independently come to the conclusion that the standard model is not the final word in particle physics, but needs to be "extended". The proposed extensions to the standard model contain many suitable CDM particle candidates (e.g., neutralinos, axions). Their interactions with standard model particles would have to be rather weak to explain why they have not been detected so far. Since these extensions were not invented to explain dark matter, but were strongly motivated by particle physics reasons, the cosmological evidence for dark matter is good, rather than bad, news from a particle physics viewpoint.

In these days the term "dark matter" usually refers to the nonbaryonic dark matter, and often excludes also neutrinos, so that it refers only to the unknown particles that are not part of the standard model of particle physics.

Since all the cosmological evidence for CDM comes from its gravitational effects, it has been suggested by some that it does not exist, and that these gravitational effects might instead be explained by suitably modifying the law of gravity at large distances. However, the suggested modifications do not appear very convincing, and the evidence is in favor of the CDM hypothesis. The gravitational effect of CDM has a role at many different levels in the history and structure of the universe, so it is difficult for a competing theory to explain all of them. Most cosmologists consider the existence of CDM as an established fact, and are just waiting for the eventual discovery of the CDM particle in the laboratory (perhaps produced with the Large Hadronic Collider (LHC) at CERN).[16]

The situation with the so-called *dark energy* is different. While dark matter fits well into theoretical expectations, the status of dark energy is much more obscure. The accumulation of astronomical data relevant to cosmology has made it possible to determine the geometry and expansion history of the universe accurately. It looks like yet another component to the energy density of the universe is required to make everything fit, in particular to explain the observed acceleration of the expansion. This component is called "dark energy". Unlike dark matter, which is clustered, the dark energy should be relatively uniform in the observable universe. And while dark matter has negligible pressure, dark energy should have large, but *negative* pressure. The simplest possibility for dark energy is a *cosmological constant* or *vacuum energy*. Unlike dark matter, dark energy was not anticipated by high-energy-physics theory, and it appears difficult to incorporate it in a natural way. Again, another possible explanation is a modification of the law of gravity at large distances. In the dark energy case, this possibility is still being seriously considered. The difference from dark matter is that there is more theoretical freedom, since there are fewer relevant observed facts to explain, and that the various proposed models for dark energy do not appear very natural. A nonzero vacuum energy by itself would be natural from quantum field theory considerations, but the observed energy scale is unnaturally low.

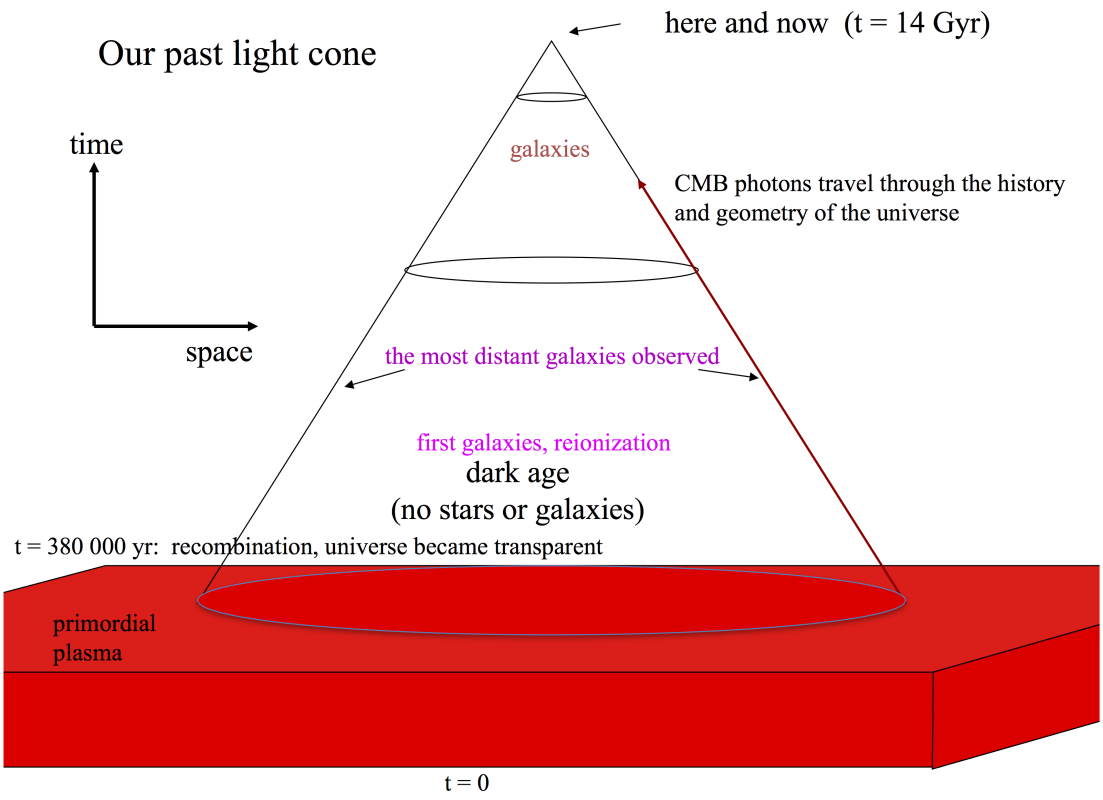


Figure 2: Our past light cone.

1.7 Observable Universe

The observations relevant to cosmology are mainly astronomical. The speed of light is finite, and therefore, when we look far away, we also look back in time. The universe has been transparent since recombination, so more than 99.99% of the history of the universe is out there for us to see. (See Fig. 2.)

The most important channel of observation is the electromagnetic radiation (light, radio waves, X-rays, etc.) coming from space. We also observe particles, *cosmic rays* (protons, electrons, nuclei) and neutrinos coming from space. A new channel, opened in 2015 by the first observation by LIGO (Laser Interferometer Gravitational-wave Observatory), are gravitational waves from space. In addition, the composition of matter in the solar system has cosmological significance.

1.7.1 Big bang and the steady-state theory

In the 1950s observational data on cosmology was rather sparse. It consisted mainly of the redshifts of galaxies, which were understood to be due to the expansion of space. At that time there was still room for different basic theories of cosmology. The main competitors were the *steady-state* theory and the *Big Bang* theory.

The steady-state theory is also known as the theory of continuous creation, since it postulates that matter is constantly being created out of nothing, so that the average density of the universe stays the same despite the expansion. According to the steady-state theory the universe has always existed and will always exist and will always look essentially the same, so that there is no overall evolution.

[16] By 2017, it is already a disappointment that LHC has not yet found a dark matter particle.

According to the Big Bang theory, the universe had a beginning at a finite time ago in the past; the universe started at very high density, and as the universe expands its density goes down. In the Big Bang theory the universe evolves; it was different in the past, and it keeps changing in the future. The name "Big Bang" was given to this theory by Fred Hoyle, one of the advocates of the steady-state theory, to ridicule it. Hoyle preferred the steady-state theory on philosophical grounds; to him, an eternal universe with no evolution was preferable to an evolving one with a mysterious beginning.

Both theories treated the observed expansion of the universe according to Einstein's theory of *General Relativity*. The steady-state theory added to it a continuous creation of matter, whereas the Big Bang theory "had all the creation in the beginning".[17]

The accumulation of further observational data led to the abandonment of the steady-state theory. These observations were: 1) the cosmic microwave background (predicted by the Big Bang theory, problematic for steady-state), 2) the evolution of cosmic radio sources (they were more powerful in the past, or there were more of them), and 3) the abundances of light elements and their isotopes (predicted correctly by the Big Bang theory).

By today the evidence has become so compelling that it appears extremely unlikely that the Big Bang theory could be wrong in any essential way, and the Big Bang theory has become the accepted basic framework, or "paradigm" of cosmology. Thus it has become arcane to talk about "Big Bang theory", when we are just referring to modern cosmology. The term "Big Bang" should be understood as originating from this historical context. Thus it refers to the present universe evolving from a completely different early stage: hot, dense, rapidly expanding and cooling, instead of being eternal and unchanging. There are still, of course, many open questions on the details, and the very beginning is still completely unknown.

1.7.2 Electromagnetic channel

Although the interstellar space is transparent (except for radio waves longer than 100 m, absorbed by interstellar ionized gas, and short-wavelength ultraviolet radiation, absorbed by neutral gas), Earth's atmosphere is opaque except for two wavelength ranges, the *optical window* ($\lambda = 300$–$800\,\mathrm{nm}$), which includes visible light, and the *radio window* ($\lambda = 1\,\mathrm{mm}$–$20\,\mathrm{m}$). The atmosphere is partially transparent to infrared radiation, which is absorbed by water molecules in the air; high altitude and dry air favors infrared astronomy. Accordingly, the traditional branches of astronomy are optical astronomy and radio astronomy. Observations at other wavelengths have become possible only during the past few decades, from space (satellites) or at very high altitude in the atmosphere (planes, rockets, balloons).

From optical astronomy we know that there are *stars* in space. The stars are grouped into *galaxies*. There are different kinds of galaxies: 1) irregular, 2) elliptical, and 3) flat disks or spirals. Our own galaxy (the Galaxy, or Milky Way galaxy) is a disk. The plane of the disk can be seen (at a dark night) as a faint band – the milky way – across the sky.

Notable nearby galaxies are the Andromeda galaxy (M31) and the Magellanic clouds (LMC, Large Magellanic Cloud, and SMC, Small Magellanic Cloud). These are the only other galaxies that are visible to the naked eye. The Magellanic clouds (as well as the center of the Milky Way) lie too far south, however, to be seen from Finland. The number of galaxies that can be seen with powerful telescopes is many billions.

[17] Thus the steady-state theory postulates a modification to known laws of physics, this continuous creation of matter out of nothing. The Big Bang theory, on the other hand, is based only on known laws of physics, but it leads to an evolution which, when extended backwards in time, leads eventually to extreme conditions where the known laws of physics can not be expected to hold any more. Whether there was "creation" or something else there, is beyond the realm of the Big Bang theory. Thus the Big Bang theory can be said to be "incomplete" in this sense, in contrast to the steady-state theory being complete in covering all of the history of the universe.

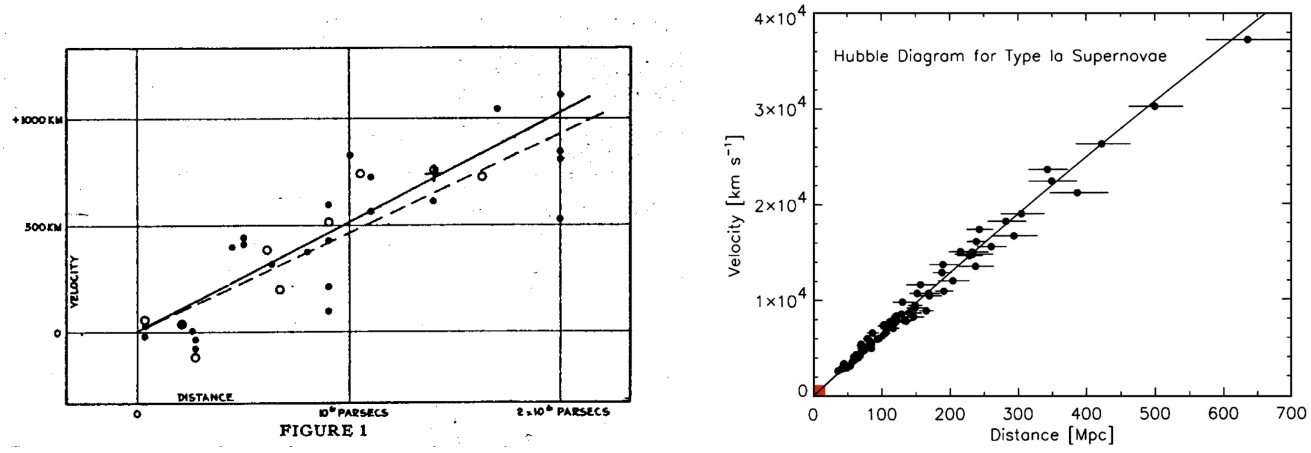


Figure 3: Left: The original Hubble diagram by Hubble. Right: A modern Hubble diagram (R.P. Kishner, PNAS 101, 8 (2004)).

Other observable objects include dust clouds, which hide the stars behind them, and gas clouds. Gas clouds absorb starlight at certain frequencies, which excite the gas atoms to higher energy states. As the atoms return to lower energy states they then emit photons at the corresponding wavelength. Thus we can determine from the spectrum of light what elements the gas cloud is made of. In the same way the composition of stellar surfaces can be determined.

The earliest "cosmological observation" was that the night sky is dark. If the universe were eternal and infinitely large, unchanging, static (not expanding, unlike in the steady state theory), and similar everywhere, our eye would eventually meet the surface of a star in every direction. Thus the entire night sky would be as bright as the Sun. This is called the *Olbers' paradox*. The Olbers' paradox is explained away by the finite age of the universe: we can not see stars further out than the distance light has travelled since the first stars were formed.[18]

1.7.3 Redshift and the Hubble law

Modern cosmology originated from the observation by Edwin Hubble[19] (in about 1929) that the redshifts of galaxies were proportional to their distance. See Fig. 3. The light from distant galaxies is redder (has longer wavelength) when it arrives here. This *redshift* can be determined with high accuracy from the spectral lines of the galaxy. These lines are caused by transitions between different energy states of atoms, and thus their original wavelengths λ_0 are known. The redshift z is defined as

$$z = \frac{\lambda - \lambda_0}{\lambda_0} \qquad \text{or} \qquad 1 + z = \frac{\lambda}{\lambda_0} \tag{4}$$

where λ is the observed wavelength. The redshift is observed to be independent of wavelength. The proportionality relation

$$z = H_0 d \tag{5}$$

is called the *Hubble law*, and the proportionality constant H_0 the *Hubble constant*. Here d is the distance to the galaxy and z its redshift.

For small redshifts ($z \ll 1$) the redshift can be interpreted as the Doppler effect due to the relative motion of the source and the observer. The distant galaxies are thus receding from us

[18] The expansion of the universe also contributes: the redshift makes distant stars fainter, and the different spacetime geometry also has an effect. Thus also the steady-state theory resolved Olbers' paradox.

[19] This proportionality was actually discovered by Lemaître before Hubble, but he published in a relatively unknown journal, so his discovery went unnoticed at the time.

with the velocity

$$v = z. \tag{6}$$

The further out they are, the faster they are receding. Astronomers often report the redshift in velocity units (i.e., km/s). Note that 1 km/s = 1/299792.458 = 0.000003356. Since the distances to galaxies are convenient to give in units of Mpc, the Hubble constant is customarily given in units of km/s/Mpc, although clearly its dimension is just 1/time or 1/distance.

This is, however, not the proper way to understand the redshift. The galaxies are not actually moving, but the distances between the galaxies are increasing because the intergalactic space between the galaxies is expanding, in the manner described by general relativity. We shall later derive the redshift from general relativity. It turns out that equations (5) and (6) hold only at the limit $z \ll 1$, and the general result, $d(z)$, relating distance d and redshift z is more complicated (discussed in Chapter 3). In particular, the redshift increases much faster than distance for large z, reaching infinity at finite d. However, redshift is directly related to the expansion. The easiest way to understand the cosmological redshift is that the wavelength of traveling light expands with the universe. (We derive this result in Chapter 3.) Thus the universe has expanded by a factor $1 + z$ during the time light traveled from an object with redshift z to us.

While the redshift can be determined with high accuracy, it is difficult to determine the distance d. See Fig. 3, right panel. The distance determinations are usually based on the *cosmic distance ladder*. This means a series of relative distance determinations between more nearby and faraway objects. The first step of the ladder is made of nearby stars, whose absolute distance can be determined from their *parallax*, their apparent motion on the sky due to our motion around the Sun. The other steps require "standard candles", classes of objects with the same absolute luminosity (radiated power), so that their relative distances are inversely related to the square roots of their "brightness" or apparent luminosity (received flux density). Several steps are needed, since objects that can be found close by are too faint to be observed from very far away.

An important standard candle is a class of variable stars called *Cepheids*. They are so bright that they can be observed (with the Hubble Space Telescope) in other galaxies as far away as the Virgo cluster of galaxies, more than 10 Mpc away. There are many Cepheids in the LMC, and the distance to the LMC (about 50 kpc) is an important step in the distance ladder. For larger distances *supernovae* (a particular type of supernovae, called Type Ia) are used as standard candles.

Errors (inaccuracies) accumulate from step to step, so that cosmological distances, and thus the value of the Hubble constant, are not known accurately. This uncertainty of distance scale is reflected in many cosmological quantities. It is customary to give these quantities multiplied by the appropriate power of h, defined by

$$H_0 = h \cdot 100\,\mathrm{km/s/Mpc}. \tag{7}$$

Still in the 1980s different observers reported values ranging from 50 to 100 km/s/Mpc ($h = 0.5$ to 1).[20]

It was a stated goal of the Hubble Space Telescope (HST) to determine the Hubble constant with 10% accuracy. As a result of some 10 years of observations the Hubble Space Telescope Key Project to Measure the Hubble Constant gave as their result in 2001 as [2]

$$H_0 = 72 \pm 8\,\mathrm{km/s/Mpc}\,. \tag{8}$$

[20] In fact, there were two "camps" of observers, one reporting values close to 50, the other close to 100, both claiming error estimates much smaller than the difference.

Modern observations have narrowed down the range and a recent value is [3]

$$H_0 = 72.5 \pm 2.5\,\mathrm{km/s/Mpc} \tag{9}$$

($h = 0.725 \pm 0.025$). Here the uncertainty (± 2.5) represents a 68% confidence range, i.e., it is estimated 68% probable that the true value lies in this range. (Unless otherwise noted, we give uncertainties as 68% confidence ranges. If the probability distribution is the so-called normal (Gaussian) distribution, this corresponds to the standard deviation (σ) of the distribution, i.e., a 1σ error estimate.) Results from different observers are not all entirely consistent with this result, so that the contribution of systematic effects to the probable error may have been underestimated.[21] To single-digit precision, we can use $h = 0.7$.

The largest observed redshifts of galaxies and quasars are about $z \sim 9$. Thus the universe has expanded by a factor of ten while the observed light has been on its way. When the light left such a galaxy, the age of the universe was only about 500 million years. At that time the first galaxies were just being formed. This upper limit in the observations is, however, not due to there being no earlier galaxies; such galaxies are just too faint due to both the large distance and the large redshift. There may well be galaxies with a redshift greater than 10. NASA is building a new space telescope, the James Webb Space Telescope[22] (JWST), which would be able to observe these.

The Hubble constant is called a "constant", since it is constant as a function of position. It is, however, a function of time, $H(t)$, in the cosmological time scale. $H(t)$ is called the Hubble parameter, and its present value is called the Hubble constant, H_0. In cosmology, it is customary to denote the present values of quantities with the subscript 0. Thus $H_0 = H(t_0)$.

The galaxies are not exactly at rest in the expanding space. Each galaxy has its own *peculiar motion* $\mathbf{v}_{\mathrm{gal}}$, caused by the gravity of nearby mass concentrations (other galaxies). Neighboring galaxies fall towards each other, orbit each other etc. Thus the redshift of an individual galaxy is the sum of the cosmic and the peculiar redshift.

$$z = H_0 d + \hat{\mathbf{n}} \cdot \mathbf{v}_{\mathrm{gal}} \qquad (\text{when } z \ll 1). \tag{10}$$

(Here $\hat{\mathbf{n}}$ is the "line-of-sight" unit vector giving the direction from the observer towards the galaxy.) Usually only the redshift is known precisely. Typically v_{gal} is of the order 500 km/s. (In large galaxy clusters, where galaxies orbit each other, it can be several thousand km/s; but then one can take the average redshift of the cluster.) For faraway galaxies, $H_0 r \gg v_{\mathrm{gal}}$, and the redshift can be used as a measure of distance. It is also related to the age of the universe at the observed time. Objects with a large z are seen in a younger universe (as the light takes a longer time to travel from this more distant object).

1.7.4 Horizon

Because of the finite speed of light and the finite age of the universe, only a finite part of the universe is observable. Our *horizon* is at that distance from which light has just had time to reach us during the entire age of the universe. Were it not for the expansion of the universe, the distance to this horizon d_{hor} would equal the age of the universe, 14 billion light years (4300 Mpc). The expansion complicates the situation; we shall calculate the horizon distance later. For large distances the redshift grows faster than (5). At the horizon $z \to \infty$, i.e., $d_{\mathrm{hor}} = d(z = \infty)$. The universe has been *transparent* only for $z < 1090$ (after recombination), so the "practical horizon", i.e., the limit to what we can see, lies already at $z \sim 1090$. The distances $d(z = 1090)$

[21] We discuss in Chapter 9 how CMB observations[6] lead to a, model-dependent, smaller value, $H_0 = 67.4 \pm 0.5\,\mathrm{km/s/Mpc}$.

[22] www.jwst.nasa.gov

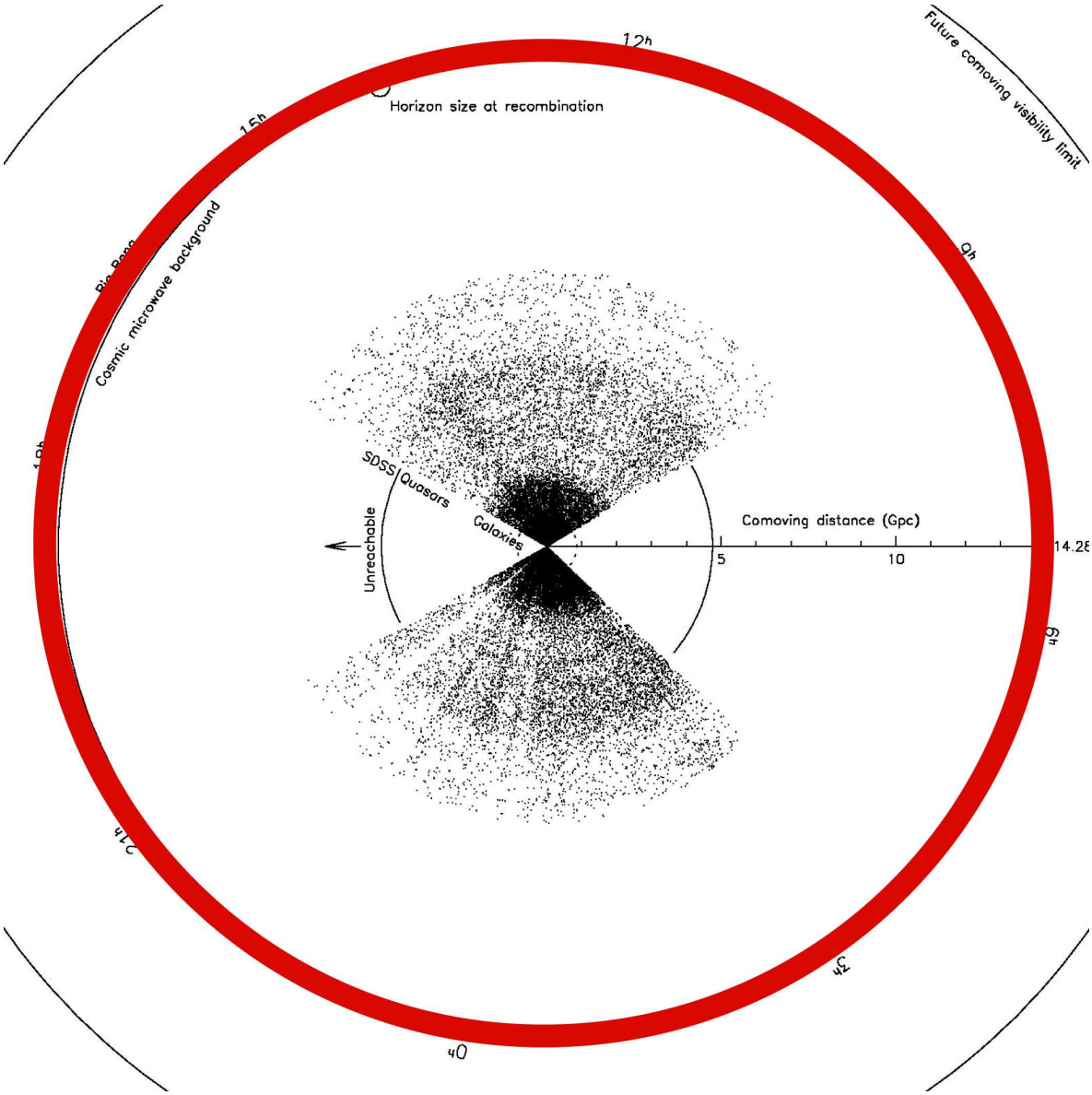


Figure 4: The distribution of galaxies from the Sloan Digital Sky Survey (SDSS) and the horizon. We are at the center of this diagram. Each dot represents an observed galaxy. The empty sectors are regions not surveyed. The figure shows fewer galaxies further out, since only the brightest galaxies can be seen at large distances. The red color represents the primordial plasma through which we cannot see. This figure can be thought of as our past light cone seen from the "top" (compare to Fig. 2). We see the inner surface of this sphere as the cosmic microwave background (see Fig. 6). As time goes on the horizon recedes and we can see further out. The "Future comoving visibility limit" is how far one can eventually see in the very distant future, assuming the "Concordance Model" for the universe (Sec. 3.3). Because of the accelerated expansion of the universe it is not possible to reach the most distant galaxies we see (beyond the circle marked "Unreachable"), even if traveling at (arbitrarily close to) the speed of light. Fig. 5 zooms in to the center region marked with the dotted circle. Figure from Gott et al: "Map of the Universe" (2005) [1].

and $d(z = \infty)$ are close to each other; $z = 4$ lies about halfway from here to horizon. The expansion of the universe complicates the concept of distance; the statements above refer to the comoving distance, defined later.

Thus the question of whether the universe is finite or infinite in space is somewhat meaningless. In any case we can only observe a finite region, enclosed in the sphere with radius $d_{\rm hor}$. Sometimes the word "universe" is used to denote just this observable part of the "whole" universe. Then we can say that the universe contains some 10^{11} or 10^{12} galaxies and about 10^{23} stars. Over cosmological time scales the horizon of course recedes and parts of the universe which are beyond our present horizon become observable. However, if the expansion keeps accelerating, as the observations indicate it has been doing already for several billion years, the observable region is already close to its maximum extent, and in the distant future galaxies which are now observable will disappear from our sight due to their increasing redshift.

1.7.5 Optical astronomy and the large scale structure

There is a large body of data relevant to cosmology from optical astronomy. Counting the number of stars and galaxies we can estimate the matter density they contribute to the universe. Counting the number density of galaxies as a function of their distance, we can try to determine whether the geometry of space deviates from Euclidean (as it might, according to general relativity). Evolution effects complicate the latter, and this approach never led to conclusive results.

From the different redshifts of galaxies within the same galaxy cluster we obtain their relative motions, which reflect the gravitating mass within the system. The mass estimates for galaxy clusters obtained this way are much larger than those obtained by counting the visible stars and galaxies in the cluster, pointing to the existence of *dark matter*.

From the spectral lines of stars and gas clouds we can determine the relative amounts of different elements and their isotopes in the universe.

The distribution of galaxies in space and their relative velocities tell us about the *large scale structure* of the universe. The galaxies are not distributed uniformly. There are galaxy groups and clusters. Our own galaxy belongs to a small group of galaxies called the Local Group. The Local Group consists of three large spiral galaxies: M31 (the Andromeda galaxy), M33 (the Triangulum galaxy[23] ; both M31 and M33 are named after the constellations they are located in), and the Milky Way, and about 30 smaller (*dwarf*) galaxies. The nearest large cluster is the Virgo Cluster. The grouping of galaxies into clusters is not as strong as the grouping of stars into galaxies. Rather the distribution of galaxies is just uneven; with denser and more sparse regions. The dense regions can be flat structures ("walls") which enclose regions with a much lower galaxy density ("voids"). See Fig. 5. The densest concentrations are called galaxy clusters, but most galaxies are not part of any well defined cluster, where the galaxies orbit the center of the cluster.

1.7.6 Radio astronomy

The sky looks very different to radio astronomy. There are many strong radio sources very far away. These are galaxies which are optically barely observable. They are distributed isotropically, i.e., there are equal numbers of them in every direction, but there is a higher density of them far away (at $z > 1$) than close by ($z < 1$). The isotropy is evidence of the homogeneity of the universe at the largest scales – there is structure only at smaller scales. The dependence on distance is a time evolution effect. It shows that the universe is not static or stationary, but

[23] Sometimes it is called the Pinwheel galaxy, but this name is also being used for M83, M99, and M101.

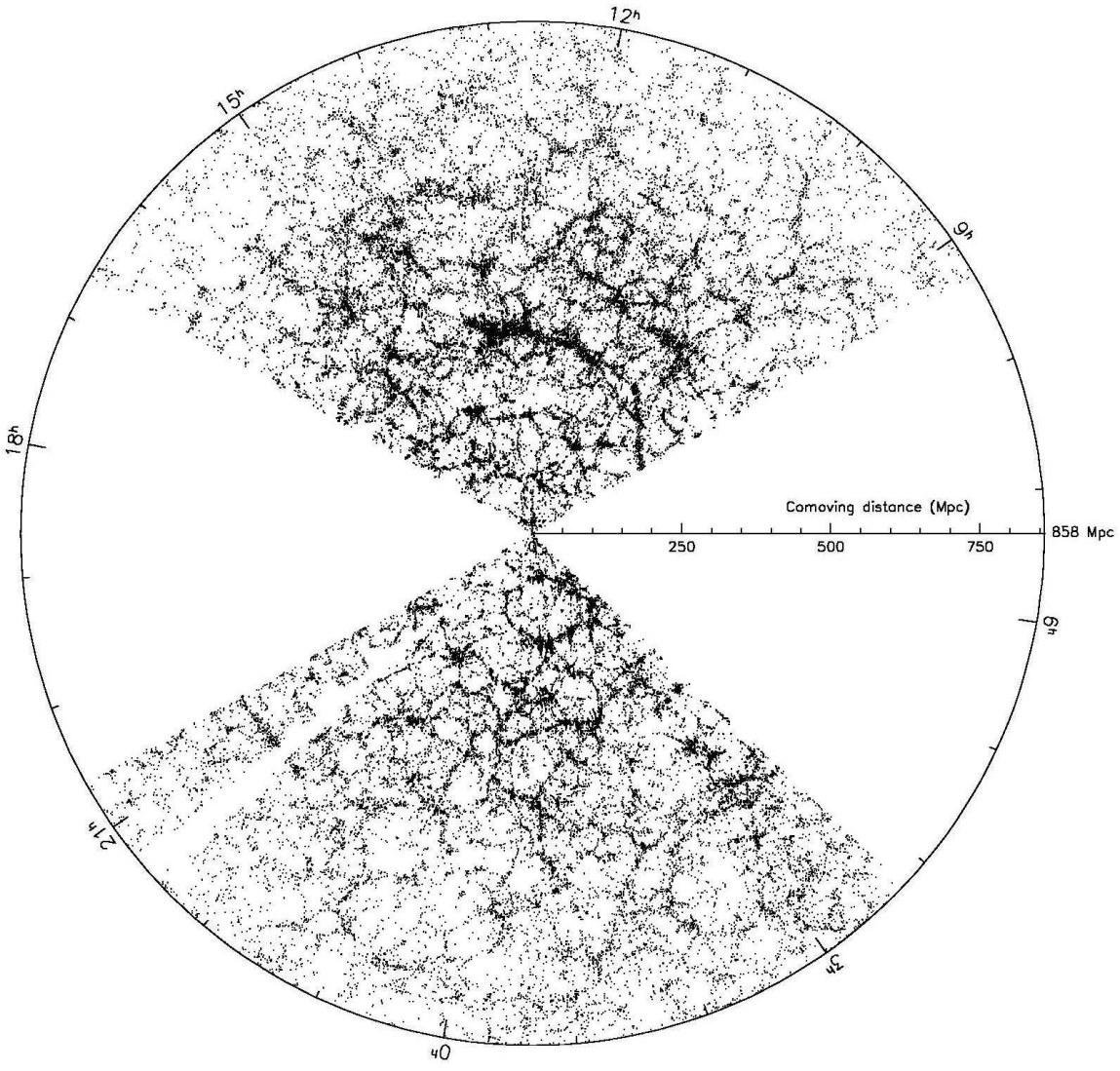


Figure 5: The distribution of galaxies from SDSS. This figure shows observed galaxies that are within 2° of the equator and closer than 858 Mpc. The empty sectors are regions not surveyed. Figure from Gott et al: "Map of the Universe" (2005) [1].

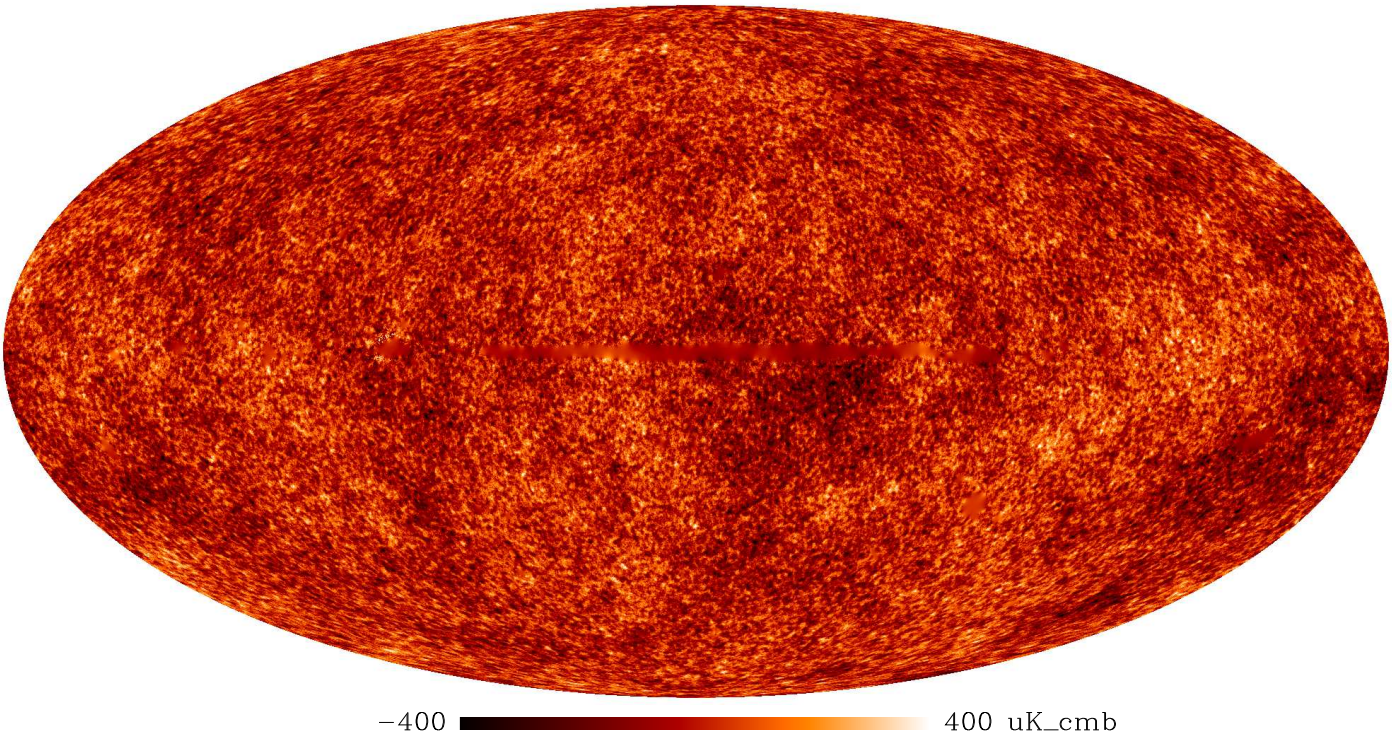


Figure 6: The cosmic microwave background. This figure shows the CMB temperature variations over the entire sky. The color scale shows deviations of -400μK to $+400\mu$K from the average temperature of 2.725 K. The plane of the milky way is horizontally in the middle. The fuzzy regions are those where the CMB is obscured by our galaxy, or nearby galaxies (can you find the LMC?).

evolves with time. Some galaxies are strong radio sources when they are young, but become weaker with age by a factor of more than 1000.

Cold gas clouds can be mapped using the 21 cm spectral line of hydrogen. The ground state ($n = 1$) of hydrogen is split into two very close energy levels depending on whether the proton and electron spins are parallel or antiparallel (the hyperfine structure). The separation of these energy levels, the hyperfine structure constant, is 5.9 μeV, corresponding to a photon wavelength of 21 cm, i.e., radio waves. The redshift of this spectral line shows that redshift is independent of wavelength (the same for radio waves and visible light), as it should be according to standard theory.

1.7.7 Cosmic microwave background

At microwave frequencies the sky is dominated by the *cosmic microwave background* (CMB), which is highly isotropic, i.e., the microwave sky appears glowing uniformly without any features, unless our detectors are extremely sensitive to small contrasts. The electromagnetic spectrum of the CMB is the black body spectrum with a temperature of $T_0 = 2.725 \pm 0.001$ K (COBE 1999 [4]). In fact, it follows the theoretical black body spectrum better than anything else we can observe or produce. There is no other plausible explanation for its origin than that it was produced in the Big Bang. It shows that the universe was homogeneous and in thermal

equilibrium at the time ($z = 1090$) when this radiation originated. The redshift of the photons causes the temperature of the CMB to fall as $(1 + z)^{-1}$, so that its original temperature was about $T = 3000\,\mathrm{K}$.

The state of a system in thermal equilibrium is determined by just a small number of thermodynamic variables, in this case the temperature and density (or densities, when there are several conserved particle numbers). The observed temperature of the CMB and the observed density of the present universe allows us to fix the evolution of the temperature and the density of the universe, which then allows us to calculate the sequence of events during the Big Bang. That the early universe was hot and in thermal equilibrium is a central part of the Big Bang paradigm, and it is often called the Hot Big Bang to spell this out.

With sensitive instruments a small anisotropy can be observed in the microwave sky. This is dominated by the *dipole anisotropy* (one side of the sky is slightly hotter and the other side colder), with an amplitude of $3362.1 \pm 1.0\,\mu\mathrm{K}$, or $\Delta T/T_0 = 0.001234$. This is a Doppler effect due to the motion of the observer, i.e., the motion of the Solar System with respect to the radiating matter at our horizon. The velocity of this motion is $v = \Delta T/T_0 = 369.8 \pm 0.1\,\mathrm{km/s}$ and it is directed towards the constellation of Leo (galactic coordinates $l = 264.02°$, $b = 48.25°$; equatorial coordinates RA $11^\mathrm{h}11^\mathrm{m}46^\mathrm{s}$, Dec $-6°57'$), near the autumnal equinox (where the ecliptic and the equator cross on the sky) [5]. It is due to two components, the motion of the Sun around the center of the Galaxy, and the peculiar motion of the Galaxy due to the gravitational pull of nearby galaxy clusters[24].

When we subtract the effect of this motion from the observations (and look away from the plane of the Galaxy – the Milky Way also emits microwave radiation, but with a nonthermal spectrum) the true anisotropy of the CMB remains, with an amplitude of about 3×10^{-5}, or 80 microkelvins.[25] See Fig. 6. This anisotropy gives a picture of the small density variations in the early universe, the "seeds" of galaxies. Theories of structure formation have to match the small inhomogeneity of the order 10^{-4} at $z \sim 1090$ and the structure observed today ($z = 0$).

1.8 Distance, luminosity, and magnitude

In astronomy, the radiated power L of an object, e.g., a star or a galaxy, is called its *absolute luminosity*. The flux density l (power per unit area) of its radiation here where we observe it, is called its *apparent luminosity*. Assuming Euclidean geometry, and that the object radiates isotropically, these are related as

$$l = \frac{L}{4\pi d^2}\,, \tag{11}$$

where d is our distance to the object. For example, the Sun has

$$L_\odot = 3.9 \times 10^{26}\,\mathrm{W} \qquad d_\odot = 1.496 \times 10^{11}\,\mathrm{m} \qquad l_\odot = 1370\,\mathrm{W/m^2}\,.$$

[24] Sometimes it is asked whether there is a contradiction with special relativity here – doesn't CMB provide an absolute reference frame? There is no contradiction. The relativity principle just says that the *laws of physics* are the same in the different reference frames. It does not say that *systems* cannot have reference frames which are particularly natural for that system, e.g., the center-of-mass frame or the laboratory frame. For road transportation, the surface of the Earth is a natural reference frame. In cosmology, CMB gives us a good "natural" reference frame – it is closely related to the center-of-mass frame of the observable part of the universe, or rather, a part of it which is close to the horizon (the *last scattering surface*). There is nothing absolute here; the different parts of the plasma from which the CMB originates are moving with different velocities (part of the 3×10^{-5} anisotropy is due to these velocity variations); we just take the average of what we see. If there is something surprising here, it is that these relative velocities are so small, of the order of just a few km/s; reflecting the astonishing homogeneity of the early universe over large scales. We shall return to the question, whether these are natural initial conditions, later, when we discuss *inflation*.

[25] The numbers refer to the standard deviation of the CMB temperature on the sky. The hottest and coldest spots deviate some 4 or 5 times this amount from the average temperature.

The ancients classified the stars visible to the naked eye into six classes according to their brightness. The concept of *magnitude* in modern astronomy is defined so that it roughly matches this ancient classification, but it is a real number, not an integer. The magnitude scale is a logarithmic scale, so that a difference of 5 magnitudes corresponds to a factor of 100 in luminosity. Thus a difference of 1 magnitude corresponds to a factor $100^{1/5} = 2.512$. The *absolute magnitude* M and the apparent magnitude m of an object are defined as

$$\begin{aligned} M &\equiv -2.5 \lg \frac{L}{L_0} \\ m &\equiv -2.5 \lg \frac{l}{l_0}\,, \end{aligned} \tag{12}$$

where L_0 and l_0 are reference luminosities ($\lg \equiv \log_{10}$). There are actually different magnitude scales corresponding to different regions of the electromagnetic spectrum, with different reference luminosities. The *bolometric* magnitude and luminosity refer to the power or flux integrated over all frequencies, whereas the *visual* magnitude and luminosity refer only to the visible light. In the bolometric magnitude scale $L_0 = 3.0 \times 10^{28}$ W. The reference luminosity l_0 for the apparent scale is chosen so in relation to the absolute scale that a star whose distance is $d = 10$ pc has $m = M$ (**exercise:** find the value of l_0). From this, (11), and (12) follows that the difference between the apparent and absolute magnitudes are related to distance as

$$m - M = -5 + 5 \lg d(\mathrm{pc}) \tag{13}$$

This difference is called the *distance modulus*, and often astronomers just quote the distance modulus, when they have determined the distance to an object. If two objects are known to have the same absolute magnitude, but the apparent magnitudes differ by 5, we can conclude that the fainter one is 10 times farther away (assuming Euclidean geometry).

For the Sun we have

$$\begin{aligned} M &= 4.79 \qquad \text{(visual)} \\ M &= 4.72 \qquad \text{(bolometric)} \\ &\text{and} \\ m &= -26.78 \qquad \text{(visual)}\,, \end{aligned} \tag{14}$$

where the apparent magnitude is as seen from Earth. Note that the smaller the magnitude, the brighter the object.

Exercises

The first three exercises are not based on these lecture notes. They should be doable with your previous physics background.

Nuclear cosmochronometers. The uranium isotopes 235 and 238 have half-lives $t_{1/2}(235) = 0.704 \times 10^9$ a ja $t_{1/2}(238) = 4.47 \times 10^9$ a. The ratio of their abundances on Earth is $^{235}\mathrm{U}/^{238}\mathrm{U} = 0.00725$. When were they equal in abundance? The heavy elements were created in supernova explosions and mixed with the interstellar gas and dust, from which the earth was formed. According to supernova calculations the uranium isotopes are produced in ratio $^{235}\mathrm{U}/^{238}\mathrm{U} = 1.3 \pm 0.2$. What does this tell us about the age of the Earth and the age of the Universe?

Olbers' paradox.

1. Assume the universe is infinite, eternal, and unchanging (and has Euclidean geometry). For simplicity, assume also that all stars are the same size as the sun, and distributed evenly in space. Show that the line of sight meets the surface of a star in every direction, sooner or later. Use Euclidean geometry.

2. Let's put in some numbers: The luminosity density of the universe is $10^8\, L_\odot/\mathrm{Mpc}^3$ (within a factor of 2). With the above assumption we have then a number density of stars $n_* = 10^8\,\mathrm{Mpc}^{-3}$. The radius of the sun is $r_\odot = 7 \times 10^8\,\mathrm{m}$. Define $r_{1/2}$ so that stars closer than $r_{1/2}$ cover 50 % of the sky. Calculate $r_{1/2}$.

3. Let's assume instead that stars have finite ages: they all appeared $t_\odot = 4.6 \times 10^9\,\mathrm{a}$ ago. What fraction f of the sky do theycover? What is the energy density of starlight in the universe, in $\mathrm{kg/m}^3$? (The luminosity, or radiated power, of the sun is $L_\odot = 3.85 \times 10^{26}\,\mathrm{W}$).

4. Calculate $r_{1/2}$ and f for galaxies, using $n_G = 3 \times 10^{-3}\,\mathrm{Mpc}^{-3}$, $r_G = 10\,\mathrm{kpc}$, and $t_G = 10^{10}\,\mathrm{a}$.

Newtonian cosmology. Use Euclidean geometry and Newtonian gravity, so that we interpret the expansion of the universe as an actual motion of galaxies instead of an expansion of space itself. Consider thus a spherical group of galaxies in otherwise empty space. At a sufficiently large scale you can treat this as a homogeneous cloud (the galaxies are the cloud particles). Let the mass density of the cloud be $\rho(t)$. Assume that each galaxy moves according to Hubble's law $\mathbf{v}(t, \mathbf{r}) = H(t)\mathbf{r}$. The expansion of the cloud slows down due to its own gravity. What is the acceleration as a function of ρ and $r \equiv |\mathbf{r}|$? Express this as an equation for $\dot{H}(t)$ (here the overdot denotes time derivative). Choose some reference time $t = t_0$ and define $a(t) \equiv r(t)/r(t_0)$. Show that $a(t)$ is the same function for each galaxy, regardless of the value of $r(t_0)$. Note that $\rho(t) = \rho(t_0)a(t)^{-3}$. Rewrite your differential equation for $H(t)$ as a differential equation for $a(t)$. You can solve $H(t)$ also using energy conservation. Denote the total energy (kinetic + potential) of a galaxy per unit mass by κ. Show that $K \equiv -2\kappa/r(t_0)^2$ has the same value for each galaxy, regardless of the value of $r(t_0)$. Relate $H(t)$ to $\rho(t_0)$, K, and $a(t)$. Whether the expansion continues forever, or stops and turns into a collapse, depends on how large H is in relation to ρ. Find out the critical value for H (corresponding to the escape velocity for the galaxies) separating these two possibilities. Turn the relation around to give the *critical density* corresponding to a given "Hubble constant" H. What is this critical density (in $\mathrm{kg/m}^3$) for $H = 70\,\mathrm{km/s/Mpc}$?

Practice with natural units.

1. The Planck mass is defined as $M_{\mathrm{Pl}} \equiv \frac{1}{\sqrt{8\pi G}}$, where G is Newton's gravitational constant. Give Planck mass in units of kg, J, eV, K, m^{-1}, and s^{-1}.

2. The energy density of the cosmic microwave background is $\rho_\gamma = \frac{\pi^2}{15}T^4$ and its photon density is $n_\gamma = \frac{2}{\pi^2}\zeta(3)T^3$, where ζ is Riemann's zeta function and $\zeta(3) = 1.20206$. What is this energy density in units of $\mathrm{kg/m}^3$ and the photon density in units of m^{-3}, i) today, when $T = 2.725\,\mathrm{K}$, ii) when the temperature was $T = 1\,\mathrm{MeV}$? What was the average photon energy, and what was the wavelength and frequency of such an average photon?

3. Suppose the mass of an average galaxy is $m_G = 10^{11}\,\mathrm{M}_\odot$ and the galaxy density in the universe is $n_G = 3 \times 10^{-3}\,\mathrm{Mpc}^{-3}$. What is the galactic contribution to the average mass density of the universe, in $\mathrm{kg/m}^3$?

4. The critical density for the universe is $\rho_{\mathrm{cr0}} \equiv \frac{3}{8\pi G}H_0^2$, where H_0 is the Hubble constant, whose value we take to be $70\,\mathrm{km/s/Mpc}$. How much is the critical density in units of $\mathrm{kg/m}^3$ and in MeV^4? What fraction of the critical density is contributed by the microwave background (today), by starlight (see earlier exercise above), and by galaxies?

Reference luminosity. Find the value of l_0 for the bolometric scale.

References

[1] J. Richard Gott III et al., *A Map of the Universe*, Astrophys. J. **624**, 463 (2005), astro-ph/0310571

[2] W.L. Freedman et al., *Final Results from the Hubble Space Telescope Key Project to Measure the Hubble Constant*, Astrophys. J. 553, 47 (2001)

[3] G. Efstathiou, H_0 *revisited*, Mon. Not. Roy. Astron. Soc. 440 (2014) 1138 [arXiv:1311.3461]

[4] J.C. Mather et al. (COBE), *Calibrator Design for the COBE Far Infrared Absolute Spectrophotometer (FIRAS)*, Astrophys. J. 512, 511 (1999); D.J. Fixsen and J.C. Mather, *The Spectral Results of the Far-Infrared Absolute Spectrophotometer Instrument on COBE*, Astrophys. J. 581, 817 (1999)

[5] Planck Collaboration, *Planck 2018 results. I. Overview, and the cosmological legacy of Planck*, arXiv:1807.06205v1

[6] Planck Collaboration, *Planck 2018 results. VI. Cosmological parameters*, arXiv:1807.06209v1

2 General Relativity

The general theory of relativity (Einstein 1915) is the theory of gravity. General relativity ("Einstein's theory") replaced the previous theory of gravity, Newton's theory. The fundamental idea in (both special and general) relativity is that space and time form together a 4-dimensional spacetime. The fundamental idea in general relativity is that gravity is manifested as *curvature* of this spacetime. While in Newton's theory gravity acts directly as a force between two bodies, in Einstein's theory the gravitational interaction is mediated by the spacetime. A massive body curves the surrounding spacetime. This curvature then affects the motion of other bodies. "Matter tells spacetime how to curve, spacetime tells matter how to move" [1]. From the viewpoint of general relativity, gravity is not a force at all; if there are no (other) forces (than gravity) acting on a body, we say the body is in *free fall*. A freely falling body is moving as straight as possible in the curved spacetime, along a *geodesic line*. If there are (other) forces, they cause the body to deviate from the geodesic line. It is important to remember that the viewpoint is that of *spacetime*, not just space. For example, the orbit of Earth around the Sun is curved in space, but as straight as possible in spacetime.

If a spacetime is not curved, we say it is *flat*, which just means that it has the geometry of Minkowski space (note the possibly confusing terminology: it is conventional to say "Minkowski space", although it is a spacetime). In the case of 2- or 3-dimensional (2D or 3D) space, "flat" means that the geometry is Euclidean.

2.1 Curved 2D and 3D space

If you are familiar with the concept of curved space and how its geometry is given by the metric, you can skip the following discussion of 2- and 3-dimensional spaces and jump to Sec. 2.3.

Ordinary human brains cannot visualize a curved 3-dimensional space, let alone a curved 4-dimensional spacetime. However, we can visualize *some* curved 2-dimensional spaces by considering them embedded in flat 3-dimensional space.[1] So let us consider first a 2D space. Imagine there are 2D beings living in this 2D space. They have no access to a third dimension. How can they determine whether the space they live in is curved? By examining whether the laws of Euclidean geometry hold. If the space is flat, then the sum of the angles of any triangle is 180°, and the circumference of any circle with radius r is $2\pi r$. If by measurement they find that this does not hold for some triangles or circles, then they can conclude that the space is curved.

A simple example of a curved 2D space is the sphere. The sum of angles of any triangle on a sphere is greater than 180°, and the circumference of any circle is less than $2\pi r$. Straight, i.e., geodesic, lines, e.g., sides of a triangle, on the sphere are sections of *great circles*, which divide the sphere into two equal hemispheres. The radius of a circle is measured along the sphere surface. See Fig. 1.

Note that the surface of a cylinder has Euclidean geometry, i.e., there is no way that 2D beings living on it could conclude that it differs from a flat surface, and thus by our definition it is a flat 2D space. (Except that by traveling around the cylinder they could conclude that their space has a strange *topology*).

In a similar manner we could try to determine whether the 3D space around us is curved, by measuring whether the sum of angles of a triangle is 180° or whether a sphere with radius r has surface area $4\pi r^2$. In fact, the space around Earth is curved due to Earth's gravity, but the

[1] This embedding is only an aid in visualization. A curved 2D space is defined completely in terms of its 2 independent coordinates, without any reference to a higher dimension, the geometry being given by the metric (a part of the definition of the 2D space), an expression in terms of these coordinates. Some such curved 2D spaces have the same geometry as some 2D surface in flat 3D space. We then say that the 2D space can be embedded in flat 3D space. But other curved 2D spaces have no such corresponding surface, i.e., they can not be embedded in flat 3D space.

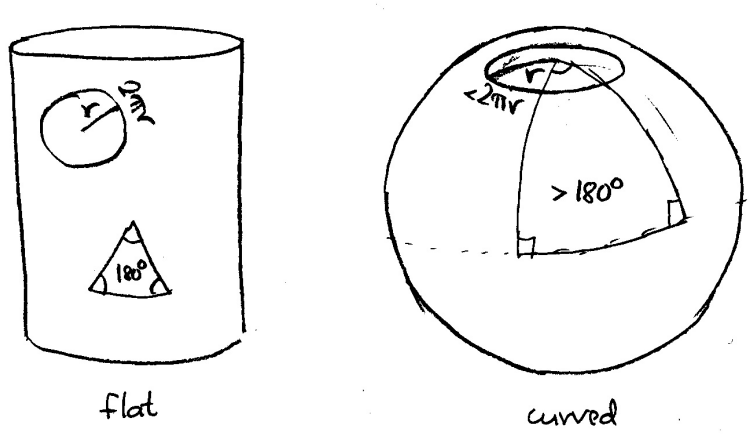


Figure 1: Cylinder and sphere.

curvature is so small that more sophisticated measurements than the ones described above are needed to detect it.

2.2 The metric of 2D and 3D space

The tool to describe the geometry of space is the *metric*. The metric is given in terms of a set of coordinates. The coordinate system can be an arbitrary curved coordinate system. The coordinates are numbers which identify locations, but do not, by themselves, yet say anything about physical distances. The distance information is in the metric.

To introduce the concept of a metric, let us first consider Euclidean 2-dimensional space with Cartesian coordinates x,y. A parameterized curve $x(\eta)$, $y(\eta)$, begins at η_1 and ends at η_2. See Fig. 2. The length of the curve is given by

$$s = \int ds = \int \sqrt{dx^2 + dy^2} = \int_{\eta_1}^{\eta_2} \sqrt{x'^2 + y'^2} d\eta \,, \tag{1}$$

where $x' \equiv dx/d\eta$, $y' \equiv dy/d\eta$. Here $ds = \sqrt{dx^2 + dy^2}$ is the *line element*. The square of the line element, the *metric*, is

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

The line element has the dimension of distance. If our coordinates are dimensionless, we need to include the distance scale in the metric. If the separation of neighboring coordinate lines, e.g., $x = 1$ and $x = 2$ is a (say, $a = 1$cm), then we have

$$ds^2 = a^2 \left(dx^2 + dy^2\right) \tag{3}$$

where a could be called the *scale factor*. As a working definition for the *metric*, we can use that *the metric is an expression which gives the square of the line element in terms of the coordinate differentials.*

We could use another coordinate system on the same 2-dimensional Euclidean space, e.g., polar coordinates. Then the metric is

$$ds^2 = a^2 \left(dr^2 + r^2 d\varphi^2\right) \,, \tag{4}$$

giving the length of a curve as

$$s = \int ds = \int a\sqrt{dr^2 + r^2 d\varphi^2} = \int_{\eta_1}^{\eta_2} a\sqrt{r'^2 + r^2 \varphi'^2} d\eta \,. \tag{5}$$

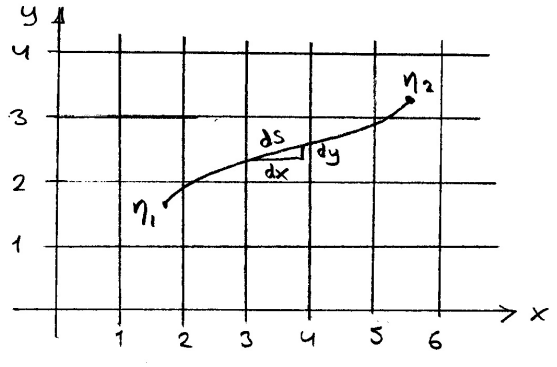


Figure 2: A parameterized curve in Euclidean 2D space with Cartesian coordinates.

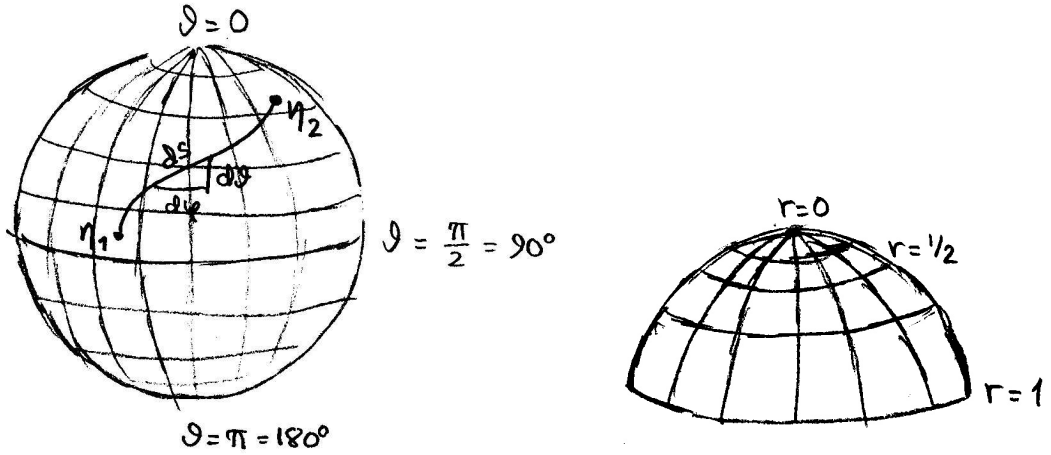


Figure 3: Left: A parameterized curve on a 2D sphere with spherical coordinates. Right: The part of the sphere covered by the coordinates in Eq. (10).

In a similar manner, in 3-dimensional Euclidean space, the metric is

$$ds^2 = dx^2 + dy^2 + dz^2 \tag{6}$$

in (dimensionful) Cartesian coordinates, and

$$ds^2 = dr^2 + r^2 d\vartheta^2 + r^2 \sin^2 \vartheta d\varphi^2 \tag{7}$$

in spherical coordinates (where the r coordinate has the dimension of distance, but the angular coordinates ϑ and φ are of course dimensionless).

Now we can go to our first example of a curved (2-dimensional) space, the sphere (the 2-sphere). Let the radius of the sphere be a. For the two coordinates on this 2D space we can take the angles ϑ and φ. We get the metric from the Euclidean 3D metric in spherical coordinates by setting $r \equiv a$,

$$ds^2 = a^2 \left(d\vartheta^2 + \sin^2 \vartheta d\varphi^2 \right) . \tag{8}$$

The length of a curve $\vartheta(\eta), \varphi(\eta)$ on this sphere (see Fig. 3) is given by

$$s = \int ds = \int_{\eta_1}^{\eta_2} a \sqrt{\vartheta'^2 + \sin^2 \vartheta \varphi'^2} d\eta \, . \tag{9}$$

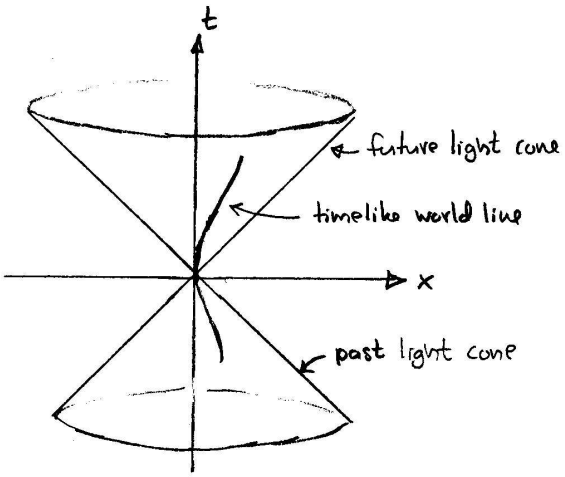


Figure 4: The light cone.

For later application in cosmology, it is instructive to now consider a coordinate transformation $r = \sin\vartheta$ (this new coordinate r has nothing to do with the earlier r of 3D space, it is a coordinate on the sphere growing in the same direction as ϑ, starting at $r = 0$ from the North Pole ($\vartheta = 0$)). Since now $dr = \cos\vartheta d\vartheta = \sqrt{1 - r^2}d\vartheta$, the metric becomes

$$ds^2 = a^2\left(\frac{dr^2}{1 - r^2} + r^2 d\varphi^2\right). \tag{10}$$

For $r \ll 1$ (in the vicinity of the North Pole), this metric is approximately the same as Eq. (4), i.e., it becomes polar coordinates on the "Arctic plain", with scale factor a. Only as r gets larger we begin to notice the deviation from flat geometry. Note that we run into a problem when $r = 1$. This corresponds to $\vartheta = \pi/2 = 90°$, i.e. the "equator". After this $r = \sin\vartheta$ begins to decrease again, repeating the same values. Also, at $r = 1$, the $1/(1 - r^2)$ factor in the metric becomes infinite. We say we have a *coordinate singularity* at the equator. There is nothing wrong with the space itself, but our chosen coordinate system applies only for a part of this space, the region "north" of the equator.

2.3 4D flat spacetime

Let us now return to the 4-dimensional spacetime. The coordinates of the 4-dimensional spacetime are (x^0, x^1, x^2, x^3), where x^0 is a time coordinate. Some examples are "Cartesian" (t, x, y, z) and spherical $(t, r, \vartheta, \varphi)$ coordinates. The coordinate system can be an arbitrary curved coordinate system. The coordinates do not, by themselves, yet say anything about physical distances. The distance information is in the metric. A *Greek index* is used to denote an arbitrary spacetime coordinate, x^μ, where it is understood that μ can have any of the values 0, 1, 2, 3. *Latin* indices are used to denote space coordinates, x^i, where it is understood that i can have any of the values 1, 2, 3.

The metric of the Minkowski space of *special relativity* is

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2, \tag{11}$$

in Cartesian coordinates. In spherical coordinates it is

$$ds^2 = -dt^2 + dr^2 + r^2 d\vartheta^2 + r^2\sin^2\vartheta\, d\varphi^2, \tag{12}$$

The fact that time appears in the metric with a different sign, is responsible for the special geometric features of Minkowski space. (I am assuming you already have some familiarity with special relativity.) There are three kinds of directions,

- timelike, $ds^2 < 0$
- lightlike, $ds^2 = 0$
- spacelike, $ds^2 > 0$.

The lightlike directions form the observer's future and past *light cones*.[2] Light moves along the light cone, so that everything we see lies on our past light cone. To see us as we are now, the observer has to lie on our future light cone. As we move in time along our world line, we drag our light cones with us so that they sweep over the spacetime. The motion of any massive body is always timelike.

2.4 Curved spacetime

These features of the Minkowski space are inherited by the spacetime of general relativity. However, spacetime is now *curved*, whereas in Minkowski space it is *flat* (i.e., not curved). The (proper) length of a spacelike curve is $\Delta s \equiv \int ds$. Light moves along lightlike world lines, $ds^2 = 0$, massive objects along timelike world lines $ds^2 < 0$. The time measured by a clock carried by the object, the *proper time*, is $\Delta\tau = \int d\tau$, where $d\tau \equiv \sqrt{-ds^2}$, so that $d\tau^2 = -ds^2 > 0$. The proper time τ is a natural parameter for the world line, $x^\mu(\tau)$. The *four-velocity* of an object is defined as

$$u^\mu = \frac{dx^\mu}{d\tau}. \tag{13}$$

The zeroth component of the 4-velocity, $u^0 = dx^0/d\tau = dt/d\tau$ relates the proper time τ to the *coordinate time* t, and the other components of the 4-velocity, $u^i = dx^i/d\tau$, to *coordinate velocity* $v^i \equiv dx^i/dt = u^i/u^0$. To convert this coordinate velocity into a "physical" velocity (with respect to the coordinate system), we still need to use the metric, see below.

In an *orthogonal* coordinate system the coordinate lines are everywhere orthogonal to each other. The metric is then diagonal, of the form

$$ds^2 \;=\; -a^2dt^2 + b^2dx^2 + c^2dy^2 + e^2dz^2 \tag{14}$$

(where a, b, c, and d are, in general, functions of t, x, y, and z), meaning that it contains no cross terms like $dxdy$. We shall only use orthogonal coordinate systems in this course. The physical distance travelled in the x direction is then bdx, and the time measured by an observer at rest in the coordinate system is adt, so that the physical velocity (in the x direction and with respect to the coordinate system) is $v_{\rm phys} = bdx/adt$.

The three-dimensional subspace ("hypersurface") $t = const$ of spacetime is called the space (or the *universe*) at time t, or a *time slice* of the spacetime. It is possible to slice the same spacetime in many different ways, i.e., to use coordinate systems with different $t = const$ hypersurfaces. See Fig. 5. The volume of a 3D region within this space given by some range in the space coordinates is given by

$$V = \int dV\,, \qquad \text{where } dV = bce\,dxdydz\,, \tag{15}$$

if the metric is (14).

[2]The light cone refers both to this set of directions and to the 3D surface in spacetime covered by light rays to/from these directions from/to the reference point (observer).

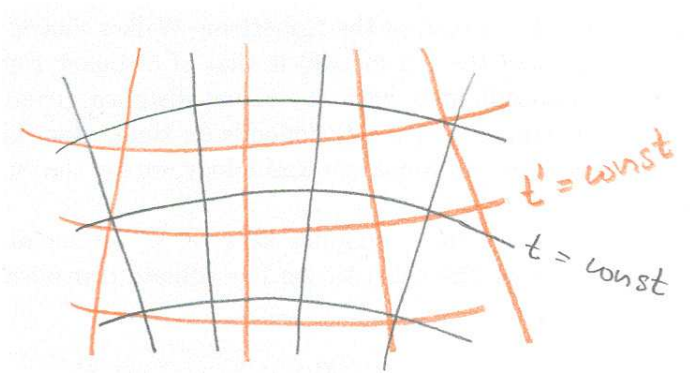


Figure 5: Two coordinate systems with different time slicings.

2.5 Einstein equation

The idea that gravitation is curvature of spacetime and the geometry of spacetime can be expressed with a metric, is only one part of general relativity (GR). To complete the theory one must give the law that determines this geometry. In GR this is given by the Einstein equation, which relates the curvature to the distribution of energy. The Einstein equation is discussed in Appendix A. For a proper introduction to the Einstein equation one should take the General Relativity course. One can also invent other metric theories of gravity where the Einstein equation is replaced with something more complicated. General relativity is the simplest possibility and is supported by observations, except for the dark energy problem. In Cosmology I we will need only the special case of the Einstein equation called the Friedmann equations, introduced in the next chapter.

References

[1] C.W. Misner, K.S. Thorne, J.A. Wheeler, Gravitation (Freeman 1973)

3 Friedmann–Robertson–Walker Universe

3.1 Kinematics

3.1.1 Robertson–Walker metric

We adopt now the cosmological principle, and discuss the homogeneous and isotropic model for the universe. This is called the Friedmann–Robertson–Walker (FRW) or the Friedmann–Lemaître–Robertson–Walker universe. Here the homogeneity refers to *spatial homogeneity* only, so that the universe will still be different at different times.

Spatial homogeneity means that there exists a coordinate system whose $t = const$ hypersurfaces are homogeneous. This time coordinate is called the *cosmic time*. Thus the spatial homogeneity property selects a preferred slicing of the spacetime, reintroducing a separation between space and time.

There is good evidence, that the Universe is indeed rather homogeneous (all places look the same) and isotropic (all directions look the same) at sufficiently large scales (i.e., ignoring smaller scale features), larger than 100 Mpc. (Recall the discussion of the cosmological principle in Chapter 1.)

Homogeneity and isotropy mean that the curvature of spacetime must be the same everywhere and into every space direction, but it may change in time.[1] It can be shown that the metric can then be given (by a suitable choice of the coordinates) in the form

$$\boxed{ds^2 = -dt^2 + a^2(t)\left[\frac{dr^2}{1-Kr^2} + r^2 d\vartheta^2 + r^2\sin^2\vartheta\, d\varphi^2\right],} \tag{1}$$

the *Robertson–Walker* (RW) metric in spherical coordinates. Doing a coordinate transformation we can also write it in Cartesian coordinates (**exercise**) :

$$ds^2 = -dt^2 + a^2(t)\frac{dx^2+dy^2+dz^2}{\left[1+\frac{1}{4}K\left(x^2+y^2+z^2\right)\right]^2}, \tag{2}$$

where $x = \tilde{r}\sin\vartheta\cos\varphi$, $y = \tilde{r}\sin\vartheta\sin\varphi$, $z = \tilde{r}\cos\vartheta$, and $\tilde{r} = (1-\sqrt{1-Kr^2})/(\frac{1}{2}Kr)$. Usually the spherical coordinates are more useful.

This is thus the metric of our universe, to first approximation, and we shall work with this metric for a large part of this course.[2] The time coordinate t is the *cosmic time*. Here K is a constant, related to curvature of *space*, and $a(t)$ is a function of time, related to expansion (or possible contraction) of the universe. We call

$$R_{\rm curv} \equiv a(t)/\sqrt{|K|} \tag{3}$$

the *curvature radius* of space (at time t).

The time-dependent factor $a(t)$ is called the *scale factor*. When the Einstein equation is applied to the RW metric, we will get the Friedmann equations, from which we can solve $a(t)$.

[1] If we drop the condition of isotropy, there are several different possible cosmological models. These spatially homogeneous but anisotropic models are called Bianchi models, after the Bianchi classification. There are nine classes, Bianchi I–IX, some of them with subclasses. The simplest is Bianchi I, where the geometry of the 3D universe is flat, but it expands at different rates in different directions. There is no evidence to favor any Bianchi model over the FRW. The FRW models are special cases of the Bianchi models, the limit where their anisotropy goes to zero; so cosmological observations can be applied to the Bianchi models to put upper limits to the anisotropy.

[2] That is, for the whole of Cosmology I. In Cosmology II we shall consider deviations from this homogeneity.

This will be done in Sec. 3.2.[3] For now, $a(t)$ is an arbitrary function of the time coordinate t. However, for much of the following discussion, we will assume that $a(t)$ grows with time; and when we refer to the age of the universe, we assume that $a(t)$ becomes zero at some finite past, which we take as the origin of the time coordinate, $t = 0$.

We use the dot, $\dot{} \equiv d/dt$, to denote derivatives with respect to cosmic time t and define

$$H \equiv \dot{a}/a\,. \tag{4}$$

This quantity $H = H(t)$ gives the expansion rate of the universe, and it is called the *Hubble parameter*. Its present value H_0 is the *Hubble constant*. (In cosmology it is customary to denote the present values of quantities with the subscript $_0$.) The dimension of H is 1/time (or velocity/distance). In time dt a distance gets stretched by a factor $1 + H dt$ (a distance L grows with velocity HL).

Note that although the metric describes a homogeneous universe, the metric itself is not explicitly homogeneous, because it depends also on the coordinate system in addition to the geometry. (This is a common situation, just like the spherical coordinates of a sphere do not form a homogeneous coordinate system, although the sphere itself is homogeneous.) However, any physical quantities that we calculate from the metric are homogeneous and isotropic.

We notice immediately that the 2-dimensional surfaces $t = r = const$ have the metric of a sphere with radius ar. Since the universe is homogeneous, the location of the origin ($r = 0$) in space can be chosen freely. We naturally tend to put ourselves at the origin, but for calculations this freedom may be useful.

We have the freedom to rescale the radial coordinate r. For example, we can multiply all values of r by a factor of 2, if we also divide a by a factor of 2 and K by a factor of 4. The geometry of the spacetime stays the same, just the meaning of the coordinate r has changed: the point that had a given value of r has now twice that value in the rescaled coordinate system. There are two common ways to rescale:

1. If $K \neq 0$, we can rescale r to make K equal to ± 1. In this case K is usually denoted k, and it has three possible values, $k = -1, 0, +1$. In this case r is dimensionless, and $a(t)$ has the dimension of distance. For $k = \pm 1$, $a(t)$ becomes equal to $R_{\rm curv}$ and is often denoted $R(t)$. Equations in this convention will be written in blue.

2. The other way is to rescale a to be one at present[4], $a(t_0) \equiv a_0 = 1$. In this case $a(t)$ is dimensionless, whereas r and $K^{-1/2}$ have the dimension of distance. We will adopt this convention from Sec. 3.1.4 on.

Choosing one of these two scalings will simplify some of our equations. One must be careful about the possible confusion resulting from comparing equations using different scaling conventions.

If $K = 0$, the space part ($t = const$) of the Robertson–Walker metric is flat. The 3-metric (the space part of the full metric) is that of ordinary Euclidean space written in spherical coordinates, with the radial distance given by ar. The *spacetime*, however, is curved, since $a(t)$ depends on time, describing the expansion or contraction of space. In common terminology, we say the "universe is flat" in this case.

If $K > 0$, the coordinate system is singular at $r = 1/\sqrt{K}$. (Remember our discussion of the 2-sphere!) With the substitution (coordinate transformation) $r = K^{-1/2} \sin(K^{1/2}\chi)$ the metric

[3] I have adopted from Syksy the separation of this chapter into Kinematics (Sec. 3.1: RW metric only) and Dynamics (Sec. 3.2: RW metric + Friedmann equations). This has the advantage that this Kinematics section applies also to other metric theories of gravity than general relativity, which one may want to consider at some point.

[4] In some discussions of the early universe, it may also be convenient to rescale a to be one at some particular early time.

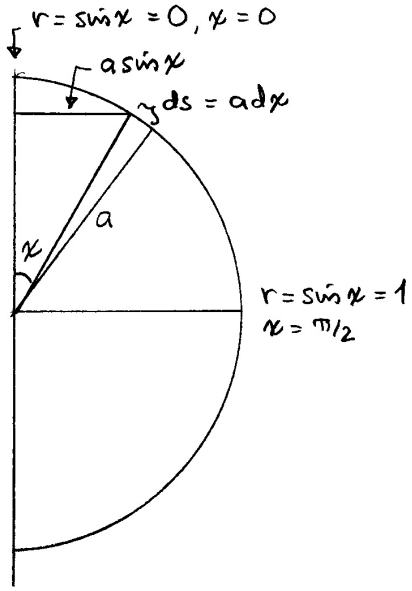


Figure 1: The hypersphere. This figure is for $K = k = 1$. Consider the semicircle in the figure. It corresponds to χ ranging from 0 to π. You get the (2-dimensional) sphere by rotating this semicircle off the paper around the vertical axis by an angle $\Delta\varphi = 2\pi$. You get the (3-dimensional) hypersphere by rotating it twice, in two extra dimensions, by $\Delta\vartheta = \pi$ and by $\Delta\varphi = 2\pi$, so that each point makes a sphere. Thus each point in the semicircle corresponds to a full sphere with coordinates ϑ and φ, and radius $(a/\sqrt{K})\sin\chi$.

becomes

$$ds^2 = -dt^2 + a^2(t)\left[d\chi^2 + K^{-1}\sin^2(K^{1/2}\chi)\,d\vartheta^2 + K^{-1}\sin^2(K^{1/2}\chi)\sin^2\vartheta\,d\varphi^2\right]. \tag{5}$$

With the scaling choice $K = k = 1$ this simplifies to

$$ds^2 = -dt^2 + a^2(t)\left[d\chi^2 + \sin^2\chi\,d\vartheta^2 + \sin^2\chi\sin^2\vartheta\,d\varphi^2\right]. \tag{6}$$

The space part has the metric of a *hypersphere* (a 3-sphere), a sphere with one extra dimension. $\sqrt{K}\chi$ is a new angular coordinate, whose values range over 0–π, just like ϑ. The singularity at $r = 1/\sqrt{K}$ disappears in this coordinate transformation, showing that it was just a coordinate singularity, not a singularity of the spacetime. The original coordinates covered only half of the hypersphere, as the coordinate singularity $r = 1/\sqrt{K}$ divides the hypersphere into two halves. The case $K > 0$ corresponds to a *closed* universe, whose (spatial) curvature is *positive*.[5] This is a finite universe, with circumference $2\pi a/\sqrt{K} = 2\pi R_{\rm curv}$ and volume $2\pi^2 K^{-3/2}a^3 = 2\pi^2 R^3_{\rm curv}$, and we can think of $R_{\rm curv}$ as the radius of the hypersphere.

If $K < 0$, we do not have a coordinate singularity, and r can range from 0 to ∞. The substitution $r = |K|^{-1/2}\sinh(|K|^{1/2}\chi)$ is, however, often useful in calculations. The case $K < 0$ corresponds to an *open* universe, whose (spatial) curvature is *negative*. The metric is then

$$ds^2 = -dt^2 + a^2(t)\left[d\chi^2 + |K|^{-1}\sinh^2(|K|^{1/2}\chi)\,\left(d\vartheta^2 + \sin^2\vartheta\,d\varphi^2\right)\right]. \tag{7}$$

This universe is infinite, just like the case $K = 0$.

[5]Positive (negative) curvature means that the sum of angles of any triangle is greater than (less than) 180° and that the area of a sphere with radius χ is less than (greater than) $4\pi\chi^2$.

To handle all three curvature cases simultaneously, we define

$$f_K(\chi) \equiv \begin{cases} K^{-1/2}\sin(K^{1/2}\chi)\,, & (K>0) \\ \chi\,, & (K=0) \\ |K|^{-1/2}\,\sinh(|K|^{1/2}\chi)\,, & (K<0) \end{cases} \tag{8}$$

which allows us to write the RW metric as

$$\boxed{ds^2 = -dt^2 + a^2(t)\Big[d\chi^2 + f_K^2(\chi)\big(d\vartheta^2 + \sin^2\vartheta\, d\varphi^2\big)\Big],} \tag{9}$$

The RW metric (at a given time) has two associated length scales. The first is the curvature radius, $R_{\rm curv} \equiv a|K|^{-1/2}$. The second is given by the time scale of the expansion, the *Hubble time* or *Hubble length* $t_H \equiv l_H \equiv H^{-1}$, whose present value is

$$H_0^{-1} = 9.7781h^{-1}\,\mathrm{Gyr} = 2997.92458h^{-1}\,\mathrm{Mpc}\,. \tag{10}$$

(Note that due to the definition of h, the digits 2997.92458 is just the speed of light in units of 100 km/s, which makes this value of l_H easy to remember.) In the case $K = 0$ the universe is flat, so the only length scale is the Hubble length.

The coordinates $(t, r, \vartheta, \varphi)$ or (t, x, y, z) of the RW metric are called *comoving* coordinates. This means that the coordinate system follows the expansion of space, so that the space coordinates of objects which *do not move* remain the same. The homogeneity of the universe fixes a special frame of reference, the *cosmic rest frame* given by the above coordinate system, so that, unlike in special relativity, the concept "does not move" has a specific meaning. The coordinate distance between two such objects stays the same, but the physical, or *proper* distance between them grows with time as space expands. The time coordinate t, the *cosmic time*, gives the time measured by such an observer at rest, at $(r, \vartheta, \varphi) = const.$

It can be shown that the expansion causes the motion of an object in free fall to slow down with respect to the comoving coordinate system. For nonrelativistic physical velocities v,

$$v(t_2) = \frac{a(t_1)}{a(t_2)}\, v(t_1). \tag{11}$$

The *peculiar velocity* of a galaxy is its velocity with respect to the comoving coordinate system.

3.1.2 Redshift

Let us now ignore the peculiar velocities of galaxies (i.e., we assume they are = 0), so that they will stay at fixed coordinate values (r, ϑ, φ), and find how their observed redshift z arises. We set the origin of our coordinate system at galaxy O (observer). Let the r-coordinate of galaxy A be r_A. Since we assumed the peculiar velocity of galaxy A to be 0, the coordinate r_A stays constant with time.

Light leaves the galaxy at time t_1 with wavelength λ_1 and arrives at galaxy O at time t_2 with wavelength λ_2. It takes a time $\delta t_1 = \lambda_1/c = 1/\nu_1$ to send one wavelength and a time $\delta t_2 = \lambda_2/c = 1/\nu_2$ to receive one wavelength (ν_1 and ν_2 are the frequencies, sent and received waves per time). Follow now the two light rays sent at times t_1 and $t_1 + \delta t_1$ (see figure). Along the light rays t and r change, ϑ and φ stay constant (this is clear from the symmetry of the problem). Light obeys the *lightlike* condition

$$ds^2 = 0\,. \tag{12}$$

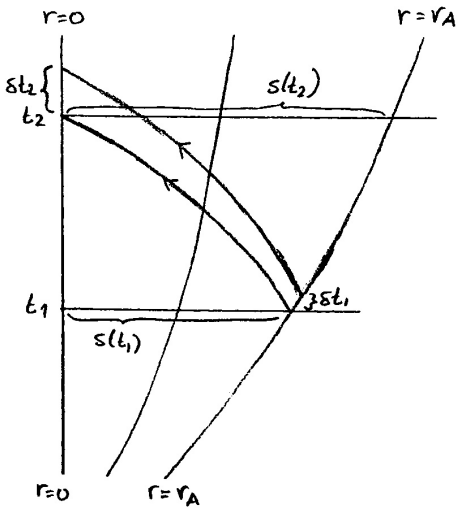


Figure 2: The two light rays to establish the redshift.

We have thus

$$ds^2 \;=\; -dt^2 + a^2(t)\frac{dr^2}{1-Kr^2} \;=\; -dt^2 + a^2(t)d\chi^2 \;=\; 0 \tag{13}$$

$$\Rightarrow \quad \frac{dt}{a(t)} \;=\; \frac{-dr}{\sqrt{1-Kr^2}} \;=\; -d\chi\,. \tag{14}$$

Integrating this, we get for the first light ray,

$$\int_{t_1}^{t_2}\frac{dt}{a(t)} \;=\; \int_0^{r_A}\frac{dr}{\sqrt{1-Kr^2}} \;=\; \int_0^{\chi_A} d\chi = \chi_A\,, \tag{15}$$

and for the second,

$$\int_{t_1+\delta t_1}^{t_2+\delta t_2}\frac{dt}{a(t)} \;=\; \int_0^{r_A}\frac{dr}{\sqrt{1-Kr^2}} \;=\; \int_0^{\chi_A} d\chi = \chi_A\,. \tag{16}$$

The right hand sides of the two equations are the same, since the sender and the receiver have not moved (they have stayed at $r = r_A$ and $r = 0$). Thus

$$0 \;=\; \int_{t_1+\delta t_1}^{t_2+\delta t_2}\frac{dt}{a(t)} - \int_{t_1}^{t_2}\frac{dt}{a(t)} \;=\; \int_{t_2}^{t_2+\delta t_2}\frac{dt}{a(t)} - \int_{t_1}^{t_1+\delta t_1}\frac{dt}{a(t)} \;=\; \frac{\delta t_2}{a(t_2)} - \frac{\delta t_1}{a(t_1)}, \tag{17}$$

and the time to receive one wavelength is

$$\delta t_2 = \frac{a(t_2)}{a(t_1)}\,\delta t_1. \tag{18}$$

As is clear from the derivation, this *cosmological time dilation* effect applies to observing any event taking place in galaxy A. As we observe galaxy A, we see everything happening in "slow motion", slowed down by the factor $a(t_2)/a(t_1)$, which is the factor by which the universe has expanded since the light (or any electromagnetic signal) left the galaxy. This effect can be observed, e.g., in the light curves of supernovae (their luminosity as a function of time).

For the redshift we get

$$1 + z \equiv \frac{\lambda_2}{\lambda_1} \;=\; \frac{\delta t_2}{\delta t_1} \;=\; \frac{a(t_2)}{a(t_1)}. \tag{19}$$

The redshift of a galaxy directly tells us how much smaller the universe was when the light left the galaxy. The result is easy to remember; the wavelength expands with the universe.

Thus the redshift z is related to the value of $a(t)$ and thus to the time t, the age of the universe, when the light left the galaxy. We can thus use a or z as alternative time coordinates. Their relation is

$$1+z \;=\; \frac{a_0}{a} \qquad \text{or} \qquad a \;=\; \frac{a_0}{1+z} \quad \Rightarrow \quad \frac{da}{a} \;=\; -\frac{dz}{1+z} \quad \Rightarrow \quad da = -\frac{a_0 dz}{(1+z)^2}\,. \tag{20}$$

Note that while a grows with time, z decreases with time: $z=\infty$ at $a=t=0$ and $z=0$ at $t=t_0$.

3.1.3 Age-redshift relation

If the observed redshift of a galaxy is z, what was the age of the universe when the light left the galaxy? Without knowing the function $a(t)$ we cannot answer this and other similar questions, but it is useful to derive general expressions in terms of $a(t)$, z, and $H(t)$. We get

$$H \;=\; \frac{1}{a}\frac{da}{dt} \quad \Rightarrow \quad dt \;=\; \frac{da}{aH} \;=\; -\frac{dz}{(1+z)H}\,, \tag{21}$$

so that the age of the universe at redshift z is

$$t(z) \;=\; \int_0^t dt' \;=\; \int_z^\infty \frac{dz'}{(1+z')H}\,. \tag{22}$$

and the present age of the universe is

$$t_0 \;\equiv\; t(z=0) \;=\; \int_0^\infty \frac{dz'}{(1+z')H}\,. \tag{23}$$

The difference gives the light travel time, i.e., how far in the past we see the galaxy,

$$t_0 - t(z) \;=\; \int_0^z \frac{dz'}{(1+z')H}\,. \tag{24}$$

3.1.4 Distance

In cosmology, the typical velocities of observers (with respect to the comoving coordinates) are small, $v < 1000\,\mathrm{km/s}$, so that we do not have to worry about Lorentz contraction (or about the velocity-related time dilation) and in the FRW model we can use the cosmic rest frame. The expansion of the universe brings, however, other complications to the concept of distance. Do we mean by the distance to a galaxy how far it is now (longer), how far it was when the observed light left the galaxy (shorter), or the distance the light has traveled (intermediate)?

The *proper distance* (or "physical distance") $d^p(t)$ between two objects[6] is defined as their distance measured along the hypersurface of constant cosmic time t. By *comoving distance* we mean the proper distance scaled to the present value of the scale factor (or sometimes to some other special time we choose as the reference time). If the objects have no peculiar velocity their comoving distance *at any time* is the same as their proper distance today.

[6] or more generally between two points (r,ϑ,φ) on the $t = const$ hypersurface. In relativity, *proper length* of an object refers to the length of an object in its rest frame, so the use of the word 'proper' in 'proper distance' is perhaps proper only when the objects are at rest in the FRW coordinate system. Nevertheless, we define it now this way.

To calculate the proper distance $d^p(t)$ between galaxies (one at $r = 0$, another at $r = r_A$) at time t, we need the metric, since $d^p(t) = \int_0^{r_A} ds$. We integrate along the path $t, \vartheta, \varphi = const$, or $dt = d\vartheta = d\varphi = 0$, so $ds^2 = a^2(t)d\chi^2 = a^2(t)\,\frac{dr^2}{1-Kr^2}$, and get

$$\begin{aligned} d^p(t) &= a(t)\int_0^{r_A}\frac{dr}{\sqrt{1-Kr^2}} = a(t)\int_0^{\chi_A} d\chi \\ &= \begin{cases} K^{-1/2}a(t)\arcsin(K^{1/2}r_A) & (K>0) \\ a(t)\,r_A & (K=0) \\ |K|^{-1/2}a(t)\,\mathrm{arsinh}(|K|^{1/2}r_A) & (K<0) \end{cases} \\ &\equiv a(t)f_K^{-1}(r_A) = a(t)\chi_A \end{aligned} \tag{25}$$

The functions $f_K(\chi)$ and

$$f_K^{-1}(r) \equiv \int_0^r \frac{dr}{\sqrt{1-Kr^2}} = \begin{cases} K^{-1/2}\arcsin(K^{1/2}r)\,, & (K>0) \\ r\,, & (K=0) \\ |K|^{-1/2}\mathrm{arsinh}(|K|^{1/2}r)\,. & (K<0) \end{cases} \tag{26}$$

convert between the two natural "unscaled" (i.e., you still need to multiply this distance by the scale factor a) radial distance definitions for the RW metric:

$$\chi = f_K^{-1}(r) = \frac{d^p}{a}\,, \tag{27}$$

the proper distance measured along the radial line, and

$$r = f_K(\chi) = f_K\left(d^p/a\right) \tag{28}$$

which is related to the length of the circle and the area of the sphere at this distance with the familiar $2\pi ar$ and $4\pi(ar)^2$.

As the universe expands, the proper distance grows,

$$d^p(t) = a(t)\chi = \frac{a_0}{1+z}\chi \equiv \frac{d^c}{1+z}\,, \tag{29}$$

where $d^c \equiv a_0\chi = d^p(t_0)$ is the present proper distance to r, or the *comoving distance* to r.

We adopt now the scaling convention $a_0 = 1$, so that the coordinate χ becomes equivalent to the comoving distance from the origin. The comoving distance between two different objects, A and B, *lying along the same line of sight*, i.e., having the same ϑ and φ coordinates, is simply $\chi_B - \chi_A$. Both f_K and f_K^{-1} have the dimension of distance.

Neither the proper distance d^p, nor the coordinate r of a galaxy are directly observable. Observable quantities are, e.g., the redshift z, location on the sky (ϑ, φ) when the observer is at $r = 0$, the angular diameter, and the apparent luminosity. We want to use the RW metric to relate these observable quantities to the coordinates and actual distances.

Let us first derive the *distance-redshift relation*. See Fig. 3. We see a galaxy with redshift z; how far is it? (We assume z is entirely due to the Hubble expansion, $1 + z = 1/a$, i.e., we ignore the contribution from the peculiar velocity of the galaxy or the observer).

Since for light,

$$ds^2 = -dt^2 + a^2(t)\frac{dr^2}{1-Kr^2} = -dt^2 + a^2(t)d\chi^2 = 0, \tag{30}$$

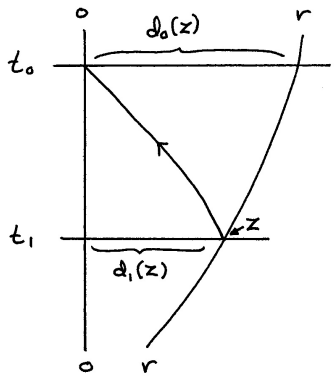


Figure 3: Calculation of the distance-redshift relation.

we have

$$dt \;=\; -a(t)\frac{dr}{\sqrt{1-Kr^2}} \;=\; -a(t)d\chi \quad\Rightarrow\quad \int_{t_1}^{t_0}\frac{dt}{a(t)} \;=\; \int_0^r \frac{dr}{\sqrt{1-Kr^2}} \;=\; \chi \;=\; d^c\,. \tag{31}$$

The comoving distance to redshift z is thus

$$d^c(z) \;=\; \chi(z) = \int_{t_1}^{t_0}\frac{dt}{a(t)} \;=\; \int_{\frac{1}{1+z}}^{1}\frac{da}{a}\frac{1}{da/dt} = \int_0^z \frac{dz'}{H(z')}\,. \tag{32}$$

The proper distance at the time the light left the galaxy is

$$d^p(z) \;=\; \frac{1}{1+z}\int_0^z \frac{dz'}{H(z')}\,. \tag{33}$$

The "distance light has traveled" (i.e., adding up the infinitesimal distances measured by a sequence of observers at rest along the light path) is equal to the light travel time, Eq. (24). In a monotonously expanding (or contracting) universe it is intermediate between $d^p(z)$ and $d^c(z)$.

We encounter the beginning of time, $t = 0$, at $a = 0$ or $z = \infty$. Thus the comoving distance light has traveled during the entire age of the universe is

$$d^c_{\rm hor} \;=\; \chi_{\rm hor} = \int_0^\infty \frac{dz'}{H(z')}\,. \tag{34}$$

This distance (or the sphere with radius $d^c_{\rm hor}$, centered on the observer) is called the *horizon*, since it represent the maximum distance we can see, or receive any information from.

There are actually several different concepts in cosmology called the horizon. To be exact, the one defined above is the *particle horizon*. Another horizon concept is the *event horizon*, which is related to how far light can travel in the future. The *Hubble distance* H^{-1} is also often referred to as the horizon (especially when one talks about *subhorizon* and *superhorizon* distance scales).

3.1.5 Volume

The objects we observe lie on our past light cone, and the observed quantities are z, ϑ, φ, so these are the observer's coordinates for the light cone. What is the volume of space corresponding to a range $\Delta z\Delta\vartheta\Delta\varphi$? Note that the light cone is a lightlike surface, so its "volume" is zero. Here we mean instead the volume that we get when we project a section of it onto the $t =$

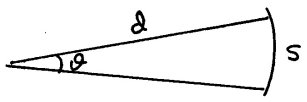


Figure 4: Defining the angular diameter distance.

const hypersurface crossing this section at a particular z (which is unique when $\Delta z = dz$ is infinitesimal).

From Eq. (33), the proper distance corresponding to dz is $dz/[(1+z)H(z)]$. Directly from the RW metric, the area corresponding to $d\vartheta d\varphi$ is $ard\vartheta \times ar\sin\vartheta d\varphi$, so that the proper volume element becomes

$$dV^p \;=\; \frac{a^2r^2\sin\vartheta}{(1+z)H(z)}dzd\vartheta d\varphi \;=\; \frac{a^2r^2}{(1+z)H(z)}dzd\Omega \tag{35}$$

and the comoving volume is

$$dV^c \;=\; (1+z)^3dV \;=\; \frac{r^2}{H(z)}dzd\Omega\,. \tag{36}$$

These are the volume elements for counting the number density or comoving number density of galaxies from observations. If the number of galaxies is conserved, in a homogeneous universe their comoving number density should be independent of z. Thus, in principle, from such observations one should be able to determine $H(z)$. In practice this is made difficult by evolution of galaxies with time, mergers of galaxies, and the fact that it is more difficult to observe galaxies at larger z.

3.1.6 Angular diameter distance

The distance-redshift relation (32) obtained above would be nice if we already knew the function $a(t)$. We can turn the situation around and use an *observed* distance-redshift relation, to obtain information about $a(t)$, or equivalently, about $H(z)$. But for that we need a different distance-redshift relation, one where the "distance" is replaced by some directly observable quantity.

Astronomers employ various such auxiliary distance concepts, like the *angular diameter distance* or the *luminosity distance*. These would be equal to the true distance in Euclidean non-expanding space.

To answer the question: "what is the physical size s of an object, whom we see at redshift z subtending an angle ϑ on the sky?" we need the concept of *angular diameter distance* d_A.

In Euclidean geometry (see Fig. 4),

$$s = \vartheta d \qquad \text{or} \qquad d = \frac{s}{\vartheta}. \tag{37}$$

Accordingly, we *define*

$$d_A \equiv \frac{s^p}{\vartheta}, \tag{38}$$

where s^p was the proper diameter of the object when the light we see left it, and ϑ is the observed angle. For large-scale structures, which expand with the universe, we use the *comoving* angular diameter distance $d_A^c \equiv s^c/\vartheta$, where $s^c = (1+z)s^p$ is the comoving diameter of the structure and z is its redshift. Thus $d_A^c = (1+z)d_A$.

From the RW metric, the physical length s^p corresponding to an angle ϑ is, from $ds^2 = a^2(t)r^2 d\vartheta^2 \quad \Rightarrow \quad s^p = a(t)r\vartheta$. Thus

$$d_A(z) \;=\; a(t)r \;=\; \frac{r}{1+z} \;=\; \frac{f_K(\chi)}{1+z} \;=\; \frac{1}{1+z} f_K\left(\int_{\frac{1}{1+z}}^{1} \frac{da}{a}\frac{1}{da/dt}\right)$$
$$= \frac{1}{1+z} f_K\left(\int_0^z \frac{dz'}{H(z')}\right) \tag{39}$$

The comoving angular diameter distance is then

$$d_A^c \;=\; r \;=\; f_K\left(\int_{\frac{1}{1+z}}^{1} \frac{da}{a}\frac{1}{da/dt}\right) \;=\; f_K\left(\int_0^z \frac{dz'}{H(z')}\right). \tag{40}$$

For the flat ($K = 0$) FRW model $r = f_K(\chi) = \chi$, so that the angular diameter distance is equal to the proper distance when the light left the object and the comoving angular diameter distance is equal to the comoving distance.

For large distances (redshifts) the angular diameter distance may have the counterintuitive property that after some z it begins to decrease as a function of z. Thus objects with are behind other objects as seen from here will nevertheless have a smaller angular diameter distance. There are two reasons for such behavior:

In a closed ($K > 0$) universe objects which are on the "other side" of the universe (the 3-sphere), i.e., with $K^{1/2}\chi > \pi/2$, will cover a larger angle as seen from here because of the spherical geometry (if we can see this far). This effect comes from the f_K in Eq. (39). An object at exactly opposite end ($K^{1/2}\chi = \pi$) would cover the entire sky as light from it would reach us from every direction after traveling half-way around the 3-sphere. In our universe these situations do not occur in practice, because lower limits to the size of the 3-sphere[7] are much larger than the distance light has traveled in the age of the Universe.

The second reason, which does apply to the observed universe, and applies only to d_A, not to d_A^c, is the expansion of the universe. An object, which does not expand with the universe, occupied a much larger comoving volume in the smaller universe of the past. This effect is the $1/(1+z)$ factor in Eq. (39), which for large z decreases faster than the other part grows. In other words, the physical size of the 2-sphere corresponding to a given redshift z has a maximum at some finite redshift (of the order $z \sim 1$), and for larger redshifts it is again smaller. (The same behavior applies to the proper distance $d^p(z)$.)

Suppose we have a set of *standard rulers*, objects that we know are all the same size s^p, observed at different redshifts. Their observed angular sizes $\vartheta(z)$ then give us the *observed* angular diameter distance as $d_A(z) = s^p/\vartheta(z)$. This observed function can be used to determine the expansion history $a(t)$, or $H(z)$.

3.1.7 Luminosity distance

In transparent Euclidean space, an object whose distance is d and whose absolute luminosity (radiated power) is L would have an apparent luminosity $l = L/4\pi d^2$. Thus we define the *luminosity distance* of an object as

$$d_L \equiv \sqrt{\frac{L}{4\pi l}}\,. \tag{41}$$

Consider the situation in the RW metric. The absolute luminosity can be expressed as:

$$L = \frac{\text{number of photons emitted}}{\text{time}} \times \text{their average energy} = \frac{N_\gamma E_{\rm em}}{t_{\rm em}}\,. \tag{42}$$

[7]Observations tell us that the curvature of the Universe is very small, so that we have not been able to determine which of the three geometries applies to it.

If the observer (at present time, $a_0 = 1$) is at a coordinate distance r from the source (note how we now put the origin of the coordinate system at the source), the photons have at that distance spread over an area

$$A = 4\pi r^2 \,. \tag{43}$$

The apparent luminosity can be expressed as:

$$l = \frac{\text{number of photons observed}}{\text{time} \cdot \text{area}} \times \text{their average energy} = \frac{N_\gamma E_{\rm obs}}{t_{\rm obs} A} \,. \tag{44}$$

The number of photons N_γ is conserved, but their energy is redshifted, $E_{\rm obs} = E_{\rm em}/(1+z)$. Also, if the source is at redshift z, it takes a factor $1+z$ longer to receive the photons $\Rightarrow t_{\rm obs} = (1+z)t_{\rm em}$. Thus,

$$l \;=\; \frac{N_\gamma E_{\rm obs}}{t_{\rm obs} A} \;=\; \frac{N_\gamma E_{\rm em}}{t_{\rm em}} \frac{1}{(1+z)^2} \frac{1}{4\pi r^2} \,. \tag{45}$$

From Eq. (41),

$$d_L \;\equiv\; \sqrt{\frac{L}{4\pi l}} \;=\; (1+z) r \;=\; (1+z) d_A^c(z) \;=\; (1+z)^2 d_A(z) \,. \tag{46}$$

Compared to the comoving angular diameter distance, $d_A^c(z)$, we have a factor $(1+z)$, which causes d_L to increase faster with z than $d_A^c(z)$. There is one factor of $(1+z)^{1/2}$ from photon redshift and another factor of $(1+z)^{1/2}$ from cosmological time dilation, both contributing to making large-redshift objects dimmer. When ompared to $d_A(z)$, there is another factor of $(1+z)$ from the expansion of the universe, which we discussed in Sec. 3.1.6, which causes distant objects to appear larger on the sky, but does not contribute to their apparent luminosity. Thus the *surface brightness* (flux density per solid angle) of objects decreases with redshift as

$$d_A^2/d_L^2 \;=\; (1+z)^{-4} \tag{47}$$

(flux density $l \propto d_L^{-2}$, solid angle $\Omega \propto d_A^{-2}$.)[8]

Suppose that we have a set of *standard candles*, objects that we know all have the same L. From their observed redshifts and apparent luminosities we get an observed luminosity-distance-redshift relation $d_L(z) = \sqrt{L/4\pi l}$, which can be used to determine $a(t)$, or $H(z)$.

3.1.8 Hubble law

In Sec. 1 we introduced the Hubble law

$$z = H_0 d \quad \Rightarrow \quad d = H_0^{-1} z \,, \tag{48}$$

which was based on observations (at small redshifts). Now that we have introduced the different distance concepts, d^p, d^c, d_A, d_A^c, d_L, in an expanding universe, and derived exact formulae (33, 32, 39, 40, 46) for them in the RW metric, we can see that for $z \ll 1$ (when we can approximate $H(z) = H_0$) all of them give the Hubble law as an approximation, but all of them deviate from it, in a different manner, for $z \sim 1$ and larger.

[8] In practical observations there is the additional issue that observations are made in some frequency band (wavelenth range), and different redshifts bring different parts of the spectrum of the object within this band.

3.1.9 Conformal time

In the comoving coordinates of Eqs. (1), (6), and (7), the space part of the coordinate system is expanding with the expansion of the universe. It is often practical to make a corresponding change in the time coordinate, so that the "unit of time" (i.e., separation of time coordinate surfaces) also expands with the universe. The *conformal time* η is defined by

$$d\eta \equiv \frac{dt}{a(t)}, \qquad \text{or} \qquad \eta = \int_0^t \frac{dt'}{a(t')}. \tag{49}$$

The RW metric acquires the form

$$ds^2 = a^2(\eta)\left[-d\eta^2 + \frac{dr^2}{1-Kr^2} + r^2\big(d\vartheta^2 + \sin^2\vartheta\, d\varphi^2\big)\right], \tag{50}$$

or with the other choice of the radial coordinate, χ,

$$ds^2 = a^2(\eta)\left[-d\eta^2 + d\chi^2 + f_K^2(\chi)\big(d\vartheta^2 + \sin^2\vartheta\, d\varphi^2\big)\right], \tag{51}$$

The form (51) is especially nice for studying radial ($d\vartheta = d\varphi = 0$) light propagation, because the lightlike condition $ds^2 = 0$ becomes $d\eta = \pm d\chi$. In the end of the calculation one may need to convert conformal time back to cosmic time to express the answer in terms of the latter.

3.2 Dynamics

3.2.1 Friedmann equations

The fundamental equation of general relativity is the Einstein equation, which relates the curvature of spacetime to the distribution of matter and energy. When applied to the homogeneous and isotropic case, i.e., the Robertson-Walker metric, it leads[9] to the *Friedmann equations*[10]

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{K}{a^2} = \frac{8\pi G}{3}\rho \tag{54}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p)\,. \tag{55}$$

("Friedmann equation" in singular refers to Eq. (54).) On the left, we have the curvature of spacetime, which in the RW metric appears as expansion of space given by $H \equiv \dot{a}/a$ and curvature of space given by K/a^2. On the right, we have the energy density ρ and pressure p of matter/energy. G is the gravitational constant, the same as in Newton's theory of gravity. Homogeneity implies the same density and pressure everywhere, so that they depend on time alone,

$$\rho = \rho(t)\,, \qquad p = p(t)\,. \tag{56}$$

Using the Hubble parameter

$$H \equiv \dot{a}/a \quad \Rightarrow \quad \dot{H} = \frac{\ddot{a}}{a} - \frac{\dot{a}^2}{a^2} \quad \Rightarrow \quad \frac{\ddot{a}}{a} = \dot{H} + H^2 \tag{57}$$

we can write the Friedmann equations also as

$$H^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2} \tag{58}$$

$$\dot{H} = -4\pi G(\rho + p) + \frac{K}{a^2}\,. \tag{59}$$

In general relativity, we do not have, in general, conservation of energy or momentum. The theoretical physics viewpoint is that conservation laws result from symmetries; energy conservation follows from time-translation symmetry and momentum conservation from space-translation symmetry. Unless the geometry of spacetime has such symmetries we do not have these conservation laws. In particular, expansion of the universe breaks time-translation symmetry and therefore energy is not conserved. The homogeneity of the RW metric leads to a form of momentum conservation, $a\mathbf{p} = const$, for particles moving in this metric.

However, the equivalence principle of general relativity requires that locally (at small scales where we do not notice the curvature of spacetime), energy and momentum are conserved. From this follows a law, called energy-momentum continuity, that applies at all scales. It can be derived from the Einstein equation. In the present case this becomes the *energy continuity equation*

$$\dot{\rho} = -3(\rho + p)\frac{\dot{a}}{a}. \tag{60}$$

[9] The Einstein equation and the derivation of the Friedmann equations from it are discussed in Appendix A.

[10] Including the cosmological constant Λ (the simplest possible modification of the Einstein equation, discussed in Appendix A), these equations take the form

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{K}{a^2} - \frac{\Lambda}{3} = \frac{8\pi G}{3}\rho \tag{52}$$

$$\frac{\ddot{a}}{a} - \frac{\Lambda}{3} = -\frac{4\pi G}{3}(\rho + 3p)\,. \tag{53}$$

We shall not include Λ in these equations. Instead, we allow for the presence of vacuum energy $\rho_{\rm vac}$, which has the same effect.

(**Exercise**: Derive this from the Friedmann equations!) Since the fluid is at rest, there is no equation for its momentum.

The Friedmann equation (58) connects the three quantities, the density ρ, the space curvature K/a^2, and the expansion rate H of the universe,

$$\rho \;=\; \frac{3}{8\pi G}\left(H^2 + \frac{K}{a^2}\right) \;\equiv\; \rho_{\rm cr} + \frac{3K}{8\pi G a^2}\,. \tag{61}$$

(Note that the curvature quantity K/a^2 is invariant under the r coordinate scaling we discussed earlier.) We defined the *critical density*

$$\rho_{\rm cr} \equiv \frac{3H^2}{8\pi G}\,, \tag{62}$$

corresponding to a given value of the Hubble parameter.[11] The critical density changes in time as the Hubble parameter evolves. The present value of the critical density is given by the Hubble constant as

$$\begin{aligned}\rho_{\rm cr0} \;&\equiv\; \rho_{\rm cr}(t_0) \;\equiv\; \frac{3H_0^2}{8\pi G} \;=\; 1.87881\times 10^{-26} h^2\,{\rm kg/m^3} \\ &=\; 10.54 h^2\,{\rm GeV/m^3} \;=\; 2.77\times 10^{11} h^2\,{\rm M_\odot/Mpc^3}\,.\end{aligned} \tag{63}$$

The nature of the curvature then depends on the density ρ:

$$\rho < \rho_{\rm cr} \qquad \Rightarrow \quad K < 0 \tag{64}$$

$$\rho = \rho_{\rm cr} \qquad \Rightarrow \quad K = 0 \tag{65}$$

$$\rho > \rho_{\rm cr} \qquad \Rightarrow \quad K > 0\,. \tag{66}$$

The *density parameter* Ω is defined

$$\Omega \equiv \frac{\rho}{\rho_{\rm cr}} \tag{67}$$

(where all three quantities are functions of time). Thus $\Omega = 1$ implies a flat universe, $\Omega < 1$ an open universe, and $\Omega > 1$ a closed universe. The Friedmann equation can now be written as

$$\boxed{\Omega \;=\; 1 + \frac{K}{H^2 a^2}} \quad \Rightarrow \quad \Omega_k(t) \;\equiv\; 1 - \Omega(t) \;=\; -\frac{K}{H(t)^2 a(t)^2}\,, \tag{68}$$

a very useful relation. Here K is a constant, and the other quantities are functions of time $\Omega(t)$, $H(t)$, and $a(t)$. The two length scales are thus related by

$$R_{\rm curv} \;=\; \frac{H^{-1}}{\sqrt{|\Omega_k(t)|}}\,. \tag{69}$$

Note that if $\Omega < 1$ (or > 1), it will stay that way. And if $\Omega = 1$, it will stay constant, $\Omega = \Omega_0 = 1$. Observations suggest that the density of the universe today is close to critical, $\Omega_0 \approx 1$, so that $R_{\rm curv0} \gg H_0^{-1}$ unless $K = 0$ (so that $R_{\rm curv} = \infty$). Writing in the present values, (68) gives

$$\Omega_k \;\equiv\; 1 - \Omega_0 \;=\; -\frac{K}{H_0^2}\,. \tag{70}$$

[11] We could also define likewise a critical Hubble parameter H_c corresponding to a given density ρ, but since, of the above three quantities, the Hubble constant has usually been the best determined observationally, it has been better to refer other quantities to it.

We defined a new notation Ω_k to represent the deviation of Ω from 1, due to curvature. Note that we write just Ω_k for its present value (instead of Ω_{k0}); if we mean the time-dependent value $1-\Omega$, we always write $\Omega_k(t)$. We adopt this common custom since we will mostly refer to the present value, and don't like to have multiple subscripts there. Note that a positive Ω_k corresponds to negative curvature and vice versa. (This sign convention is so that we have a pleasing symmetry in the Friedmann equation, see Sec. 3.2.2, Eqs. 108, 109, 110.)

Newtonian cosmology. Newtonian gravity is known to be a good approximation to general relativity for many situations, so we should be able to get something like the Friedmann equations from it, too. Consider therefore a large spherically symmetric expanding homogeneous group of galaxies in otherwise empty Euclidean space. Spherical symmetry implies that all motion is radial. Let $r(t)$ be the radial coordinate (distance from origin) of some galaxy. The velocity of the galaxy is then $\dot{r}$. Denote the total mass of all galaxies within r by

$$M(r) \equiv \frac{4\pi}{3} r^3 \rho \,, \tag{71}$$

where ρ is the mass density, assumed homogeneous, due to the galaxies. We know that in Newtonian gravity the gravitational force at r due to a spherically symmetric mass distribution is equal to that of a point mass $M(r)$ located at $r=0$ (the force due to the outer masses, beyond r, cancels). Therefore the acceleration of the galaxy is

$$\ddot{r} = -G\frac{M(r)}{r^2} = -\frac{4\pi G}{3}\rho r^2 \quad \Rightarrow \quad \frac{\ddot{r}}{r} = -\frac{4\pi G}{3}\rho \,. \tag{72}$$

Defining $H \equiv \dot{r}/r$ gives

$$\dot{H} + H^2 = \frac{\ddot{r}}{r} = -\frac{4\pi G}{3}\rho \,. \tag{73}$$

Choose now a reference time t_0, denote $r(t_0) \equiv r_0$, and define

$$a(t) \equiv \frac{r(t)}{r_0} \quad \Rightarrow \quad \frac{\dot{a}}{a} = H(t) \tag{74}$$

and

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\rho(t) \,. \tag{75}$$

So far we considered the motion of some individual galaxy. Assume now as an initial condition that $H(t_0) = H_0$ is the same for all galaxies (Hubble law). Since the differential equation (75) and the initial conditions $a(t_0) = 1$ and $\dot{a}(t_0) = H_0$ are the same for all galaxies, the solution $a(t)$ will also be the same for all galaxies. Therefore no galaxy will move past another and the "mass inside" for any particular galaxy will stay constant

$$M(r) = \frac{4\pi}{3}\rho a^3 r_0^3 = const \quad \Rightarrow \quad \rho(t) \propto a^{-3} \,. \tag{76}$$

Thus the mass density decreases homogeneously. (There is an apparent circular argument here, since we already assumed that ρ in (75) will stay homogenous. But we have now shown that this assumption leads to a consistent solution; and since the solution must be unique, the assumption must be correct.)

The solution depends on the initial conditions H_0 (related to kinetic energy) and ρ_0 (related to gravitational potential energy). Consider the situation from energy conservation. The total energy of an individual galaxy is

$$E = \tfrac{1}{2}m\dot{r}^2 - m\frac{GM(r)}{r} = \tfrac{1}{2}m\dot{r}^2 - \frac{4\pi G}{3}m\rho r^2 \equiv m\kappa = const \,, \tag{77}$$

where m is the mass of the galaxy and

$$\kappa \equiv \tfrac{1}{2}\dot{r}^2 - \frac{4\pi G}{3}\rho r^2 = \tfrac{1}{2}\dot{a}^2 r_0^2 - \frac{4\pi G}{3}\rho a^2 r_0^2 \tag{78}$$

is energy/mass. We get that

$$K \equiv -\frac{2\kappa}{r_0^2} = -\dot{a}^2 + \frac{8\pi G}{3}\rho a^2 \tag{79}$$

is the same for all galaxies. Dividing by a^2 this energy conservation law becomes the Friedmann equation

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{K}{a^2}\,. \tag{80}$$

Comparing (80) and (75) to (54) and (55) we note that the only apparent difference is that there is no pressure p in (75). This is a difference between Newtonian gravity and general relativity: in general relativity also pressure is a source of gravity. Besides this apparent difference there are fundamental conceptual differences: in the Newtonian description K referred to total (kinetic+potential) energy of a galaxy and the space was Euclidian; in the relativistic description K gives the curvature of space and the concept of gravitational potential energy does not exist. In the Newtonian description the galaxies are moving; in the relativistic description the space is expanding. The Newtonian description requires that the group of galaxies has an outer boundary, which has to be spherical. If this boundary, beyond which there should be no galaxies, is far away, at $r \geq H^{-1}$, galaxies there would be moving faster than the speed of light (with respect to the galaxies at the center). Even in the case where pressure is negligible, so that the "Friedmann" equations are the same, these conceptual differences lead to different physical results (e.g., redshift) due to the different spacetime geometry.

To solve the Friedmann equations, we need the *equation of state* that relates p and ρ. In general, the pressure p of matter may depend also on other thermodynamic variables than the energy density ρ. The equation of state is called *barotropic* is p is uniquely determimed by ρ, i.e., $p = p(\rho)$. Regardless of the nature of matter, in a homogeneous universe we have $p = p(\rho)$ in practice, if the energy density decreases monotonously with time, since $p = p(t)$, $\rho = \rho(t)$ and we can invert the latter to get $t(\rho)$, so that we can write $p = p(t) = p(t(\rho)) \equiv p(\rho)$.

We define the *equation-of-state parameter*

$$w \equiv \frac{p}{\rho} \tag{81}$$

so that we can formally write the equation of state as

$$p = w\rho\,, \tag{82}$$

and the energy continuity equation as

$$\frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{a}}{a} \quad \Rightarrow \quad d\ln\rho = -3(1+w)d\ln a\,, \tag{83}$$

where, in general, $w = w(t)$. Equation (83) can be formally integrated to

$$\frac{\rho}{\rho_0} = \exp\left\{\int_a^1 3[1+w(a')]\,\frac{da'}{a'}\right\} = \exp\left\{\int_0^z dz'\,\frac{3[1+w(z')]}{1+z'}\right\}. \tag{84}$$

The simplest case is the one where $p \propto \rho$, so that

$$p = w\rho\,, \qquad w = const\,, \tag{85}$$

in which case the solution of (83) is

$$\rho = \rho_0 a^{-3(1+w)}\,. \tag{86}$$

There are three such cases:

- **"Matter"** ($w = 0$) (called "matter" in cosmology, but "dust" in general relativity), meaning nonrelativistic matter (particle velocities $v \ll 1$), for which $p \ll \rho$, so that we can forget the pressure, and approximate $p = 0$. From Eq. (60), $d(\rho a^3)/dt = 0$, or $\rho \propto a^{-3}$.

- **"Radiation"** ($w = 1/3$), meaning ultrarelativistic matter (where particle energies are $\gg$ their rest masses, which is always true for massless particles like photons), for which $p = \rho/3$. From Eq. (60), $d(\rho a^4)/dt = 0$, or $\rho \propto a^{-4}$.
- **Vacuum energy** ($w = -1$) (or the cosmological constant), for which $\rho = const$ (property of the vacuum, a fundamental constant). From Eq. (60) follows the equation of state for vacuum energy: $p = -\rho$. Thus a positive vacuum energy[12] corresponds to a negative vacuum pressure. You may be used to pressure being positive, but there is nothing unphysical about negative pressure. In other contexts it is often called (positive) "tension" instead of (negative) "pressure".[13]

We know that the Universe contains ordinary, nonrelativistic matter. We also know that there is radiation, especially the cosmic microwave background. In Chapter 4 we shall discuss how the different known particle species behave as radiation in the early universe when it is very hot, but as the universe cools, the massive particles change from ultrarelativistic (radiation) to nonrelativistic (matter). During the transition period the pressure due to that particle species falls from $p \approx \rho/3$ to $p \approx 0$. We shall discuss these transition periods in Chapter 4. In this chapter we focus on the later evolution of the universe (after big bang nucleosynthesis, BBN). Then the known forms of matter and energy in the universe can be divided into these two classes: matter ($p \approx 0$) and radiation ($p \approx \rho/3$).[14]

We already revealed in Chapter 1 that the present observational data cannot be explained in terms of known forms of particles and energy using known laws of physics, and therefore we believe that there are other, unknown forms of energy in the universe, called "dark matter" and "dark energy". Dark matter has by definition negligible pressure, so that we can ignore its pressure in the Friedmann equations. However, to explain the observed expansion history of the universe, an energy component with negative pressure is needed. This we call dark energy. We do not know its equation of state. The simplest possibility for dark energy is just the cosmological constant (vacuum energy), which fits current data perfectly. Therefore we shall carry on our discussion assuming three energy components: matter, radiation, and vacuum energy. We shall later (at end of Sec. 3.2.4 and in Cosmology II) comment on how much current observations actually constrain the equation of state for dark energy.

If the universe contains several energy components

$$\rho = \sum_i \rho_i \qquad \text{with} \qquad p_i = w_i \rho_i \tag{87}$$

without significant energy transfer between them, then each component satisfies the energy continuity equation separately,

$$\frac{\dot{\rho}_i}{\rho_i} = -3(1 + w_i)\frac{\dot{a}}{a}. \tag{88}$$

[12] In the quantum field theory view, "vacuum" is the minimum energy density state of the system. Therefore any other contribution to energy density is necessarily positive, but whether the vacuum energy density itself needs to be nonnegative is less clear. Other physics except general relativity is sensitive only to energy differencies and thus does not care about the value of vacuum energy density. In general relativity it is a source of gravity, but cannot be distinguished from a cosmological constant, which is a modification of the law of gravity by an arbitrary constant that could be negative just as well as positive. For simplicity we will here include the possible cosmological constant in the concept of vacuum energy, and thus we should allow for negative vacuum energy density also.

[13] In Chapter 4 we derive formulae for the pressures of different particle species in thermal equilibrium. These always give a positive pressure. The point is that there we ignore interparticle forces. To make the pressure from particles negative would require attractive forces between particles. But the vacuum pressure is not from particles, it's from the vacuum. If the dark energy is not just vacuum energy, it is usually thought to be some kind of field. For fields, a negative pressure comes out more naturally than for particles.

[14] Except that we do not know the small masses of neutrinos. Depending on the values of these masses, neutrinos may make this radiation-to-matter transition sometime during this "later evolution".

and, if $w_i = const$,

$$\rho_i \propto a^{-3(1+w_i)} \quad \Rightarrow \quad \rho_i = \rho_{i0}\left(\frac{a}{a_0}\right)^{-3(1+w_i)} . \tag{89}$$

In the early universe there were times where such energy transfer was important, but after BBN it was negligible, so then we have the above case with

$$\rho = \rho_r + \rho_m + \rho_{\rm vac} \qquad \text{with} \qquad w_r = 1/3\,,\; w_m = 0\,,\; w_{\rm vac} = -1\,,. \tag{90}$$

We can then arrange Eq. (54) into the form

$$\left(\frac{\dot a}{a}\right)^2 = \alpha^2 a^{-4} + \beta^2 a^{-3} - K a^{-2} + \frac{1}{3}\Lambda, \tag{91}$$

where α, β, K, and $\Lambda = 8\pi G\rho_{\rm vac}$ are constants (α and β are temporary notation, which we replace with standard cosmological quantities in Eq. 108). The four terms on the right are due to radiation, matter, curvature, and vacuum energy, in that order. As the universe expands (a grows), different components on the right become important at different times. Early on, when a was very small, the Universe was radiation-dominated. If the Universe keeps expanding without limit, eventually the vacuum energy will become dominant (already it appears to be the largest term). In the middle we may have matter-dominated and curvature-dominated eras. In practice it seems the curvature of the Uuniverse is quite small and therefore there never was a curvature-dominated era, but there was a long matter-dominated era.

We know that the radiation component is insignificant at present, and we can ignore it in Eq. (91), if we exclude the first few million years of the Universe from discussion. Conversely, during those first few million years we can ignore the curvature and vacuum energy.

In the "inflationary scenario", there was something resembling a very large vacuum energy density in the very early universe (during a small fraction of the first second), which then disappeared. So there may have been a very early "vacuum-dominated" era (inflation), discussed in Cosmology II.

Let us now solve the Friedmann equation for the case where one of the four terms dominates. The equation has the form

$$\left(\frac{\dot a}{a}\right)^2 = \alpha^2 a^{-n} \qquad \text{or} \qquad a^{\frac{n}{2}-1}\,da = \alpha\,dt. \tag{92}$$

Integration gives

$$\frac{2}{n}a^{\frac{n}{2}} = \alpha t, \tag{93}$$

where we chose the integration constant so that $a(t=0) = 0$. We get the three cases:

$n = 4$	radiation dominated	$a \propto t^{1/2}$
$n = 3$	matter dominated	$a \propto t^{2/3}$
$n = 2$	curvature dominated ($K < 0$)	$a \propto t$

The cases $K > 0$ and vacuum energy have to be treated differently (**exercise**).

Example: The Einstein–de Sitter universe. Consider the simplest case, $\Omega = 1$ ($K = 0$) and $\Lambda = 0$. The first couple of million years when radiation can not be ignored, makes an insignificant contribution to the present age of the universe, so we ignore radiation also. We have now the matter-dominated case. For the density we have

$$\rho \;=\; \rho_0 a^{-3} \;=\; \Omega_0\rho_{\rm cr0}a^{-3} \;=\; \rho_{\rm cr0}a^{-3}\,. \tag{94}$$

The Friedmann equation is now

$$\left(\frac{\dot{a}}{a}\right)^2 = \underbrace{\frac{8\pi G}{3}\rho_{\rm cr0}}_{H_0^2}\, a^{-3} \qquad \Rightarrow \qquad a^{1/2}da = H_0 dt$$

$$\Rightarrow \int_{a_1}^{a_2} a^{1/2}da = H_0 \int_{t_1}^{t_2} dt \qquad \Rightarrow \qquad \frac{2}{3}(a_2^{3/2} - a_1^{3/2}) = H_0(t_2 - t_1)\,.$$

Thus we get

$$t_2 - t_1 \;=\; \frac{2}{3}H_0^{-1}\left(a_2^{3/2} - a_1^{3/2}\right) \;=\; \frac{2}{3}H_0^{-1}\left[\frac{1}{(1+z_2)^{3/2}} - \frac{1}{(1+z_1)^{3/2}}\right] \tag{95}$$

where z is the redshift.

- Let $t_2 = t_0$ be the present time ($z = 0$). The time elapsed since $t = t_1$ corresponding to redshift z is

$$t_0 - t \;=\; \frac{2}{3}H_0^{-1}\left(1 - a_1^{3/2}\right) \;=\; \tfrac{2}{3}H_0^{-1}\left[1 - \frac{1}{(1+z)^{3/2}}\right]. \tag{96}$$

- Let $t_1 = 0$ and $t_2 = t(z)$ be the time corresponding to redshift z. The age of the universe corresponding to z is

$$t \;=\; \frac{2}{3}H_0^{-1}a_2^{3/2} \;=\; \frac{2}{3}H_0^{-1}\frac{1}{(1+z)^{3/2}} \;=\; t(z). \tag{97}$$

  This is the *age-redshift relation*. For the present ($z = 0$) age of the universe we get

$$t_0 = \frac{2}{3}H_0^{-1}. \tag{98}$$

The Hubble constant is $H_0 \equiv h{\cdot}100$ km/s/Mpc $= h/(9.78 \times 10^9\,{\rm yr})$, or $H_0^{-1} = h^{-1}{\cdot}\,9.78 \times 10^9$ yr. Thus

$$t_0 = h^{-1} \cdot 6.52 \times 10^9 {\rm yr} = \begin{cases} 9.3 \times 10^9\,{\rm yr} & h = 0.7 \\ 13.0 \times 10^9\,{\rm yr} & h = 0.5 \end{cases} \tag{99}$$

The ages of the oldest stars appear to be at least about 12×10^9 years. Considering the HST value for the Hubble constant ($h = 0.72 \pm 0.08$), this model has an *age problem*.

Example: The closed Friedmann model. The FRW models with $\rho = \rho_m$, so that $\rho = \rho_0 a^{-3}$, are called Friedmann models[1, 2]. The Friedmann equation becomes

$$\left(\frac{\dot{a}}{a}\right) \;=\; \frac{8\pi G}{3}\rho_0 a^{-3} - \frac{K}{a^2} \quad \Rightarrow \quad \frac{da}{dt} = \sqrt{\Omega_0 H_0^2 a^{-1} - K}. \tag{100}$$

There are three Friedmann models: open ($K < 0$), flat ($K = 0$), and closed ($K > 0$). (The flat case is the Einstein–de Sitter universe.) Consider the closed one. The Friedmann equation is then

$$\frac{da}{dt} \;=\; \sqrt{K}\sqrt{\frac{C-a}{a}}\,, \tag{101}$$

where $C \equiv \Omega_0 H_0^2/K = \Omega_0/|\Omega_k|$. The solution of (101) as $a(t)$ is not an elementary function, but we can obtain it in parametrized form $a(\psi)$, $t(\psi)$ by doing the substitution

$$a(\psi) \;=\; C\sin^2\tfrac{1}{2}\psi \;=\; \tfrac{1}{2}C(1 - \cos\psi)\,. \tag{102}$$

Sticking this in both sides of (101) gives

$$C\sin\tfrac{1}{2}\psi\cos\tfrac{1}{2}\psi\frac{d\psi}{dt} \;=\; \sqrt{K}\frac{\cos\frac{1}{2}\psi}{\sin\frac{1}{2}\psi} \quad \Rightarrow \quad \frac{dt}{d\psi} \;=\; \frac{C}{\sqrt{K}}\sin^2\tfrac{1}{2}\psi \;=\; \frac{C}{2\sqrt{K}}(1 - \cos\psi)\,, \tag{103}$$

which is easy to integrate to

$$t(\psi) \;=\; \frac{C}{2\sqrt{K}}(\psi - \sin\psi)\,. \tag{104}$$

The parameter ψ is called *development angle*. The resulting curve $a(t)$ has the form of a *cycloid*, the path made by a point at the rim of a wheel, ψ being the rotation angle of the wheel. Note that since $dt/d\psi = a/\sqrt{K}$, ψ is proportional to the conformal time, $\psi = \sqrt{K}\eta$.

Exercise: The open Friedmann model. Find the corresponding results for $K < 0$.

Example: The Einstein universe and the Eddington universe. To be added some day here.

3.2.2 Cosmological parameters

We divide the density into its matter, radiation, and vacuum components $\rho = \rho_m + \rho_r + \rho_{\rm vac}$, and likewise for the density parameter, $\Omega = \Omega_m(t) + \Omega_r(t) + \Omega_\Lambda(t)$, where $\Omega_m(t) \equiv \rho_m/\rho_{\rm cr}$, $\Omega_r(t) \equiv \rho_r/\rho_{\rm cr}$, and $\Omega_\Lambda(t) \equiv \rho_{\rm vac}/\rho_{\rm cr} \equiv \Lambda/3H^2$. $\Omega_m(t)$, $\Omega_r(t)$, and $\Omega_\Lambda(t)$ are functions of time (although $\rho_{\rm vac}$ is constant, $\rho_{\rm cr}(t)$ is not). We follow the common practice where Ω_m, Ω_r, and Ω_Λ denote the present values of these density parameters, and we write $\Omega_m(t)$, $\Omega_r(t)$, and $\Omega_\Lambda(t)$, if we want to refer to their values at other times. Thus we write

$$\Omega_0 \equiv \Omega_m + \Omega_r + \Omega_\Lambda \,. \tag{105}$$

We have both

$$\Omega_m + \Omega_r + \Omega_\Lambda + \Omega_k = 1 \qquad \text{and} \qquad \Omega_m(t) + \Omega_r(t) + \Omega_\Lambda(t) + \Omega_k(t) = 1 \tag{106}$$

The present radiation density is relatively small, $\Omega_r \sim 10^{-4}$ (we shall calculate it in Chapter 4), so that we usually write just

$$\Omega_0 = \Omega_m + \Omega_\Lambda. \tag{107}$$

The radiation density is also known very accurately from the temperature of the cosmic microwave background, and therefore Ω_r is not usually considered a cosmological parameter (in the sense of an inaccurately known number that we try to fit with observations). The FRW cosmological model is thus defined by giving the present values of the three cosmological parameters, H_0, Ω_m, and Ω_Λ.

We can now write the Friedmann equation as

$$\begin{aligned} H^2 \;\equiv\; \left(\frac{\dot a}{a}\right)^2 \;&=\; \underbrace{\frac{8\pi G}{3}\Omega_r\rho_{\rm cr0}}_{\Omega_r H_0^2}\, a^{-4} + \underbrace{\frac{8\pi G}{3}\Omega_m\rho_{\rm cr0}}_{\Omega_m H_0^2}\, a^{-3} + \Omega_\Lambda H_0^2 \underbrace{-K}_{+\Omega_k H_0^2}\, a^{-2} \\ &=\; H_0^2\left(\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda\right) . \end{aligned} \tag{108}$$

or

$$\boxed{H(z) \;=\; H_0\sqrt{\Omega_r(1+z)^4 + \Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}\,.} \tag{109}$$

Observations favor the values $h \sim 0.7$, $\Omega_m \sim 0.3$, and $\Omega_\Lambda \sim 0.7$. (We discussed the observational determination of H_0 in Chapter 1. We shall discuss the observational determination of Ω_m and Ω_Λ both in this chapter and later.)

Since the critical density is $\propto h^2$, it is often useful to use instead the "physical" or "reduced" density parameters, $\omega_m \equiv \Omega_m h^2$, $\omega_r \equiv \Omega_r h^2$, which are directly proportional to the actual densities in kg/m^3. (An ω_Λ turns out not be so useful and is not used.)

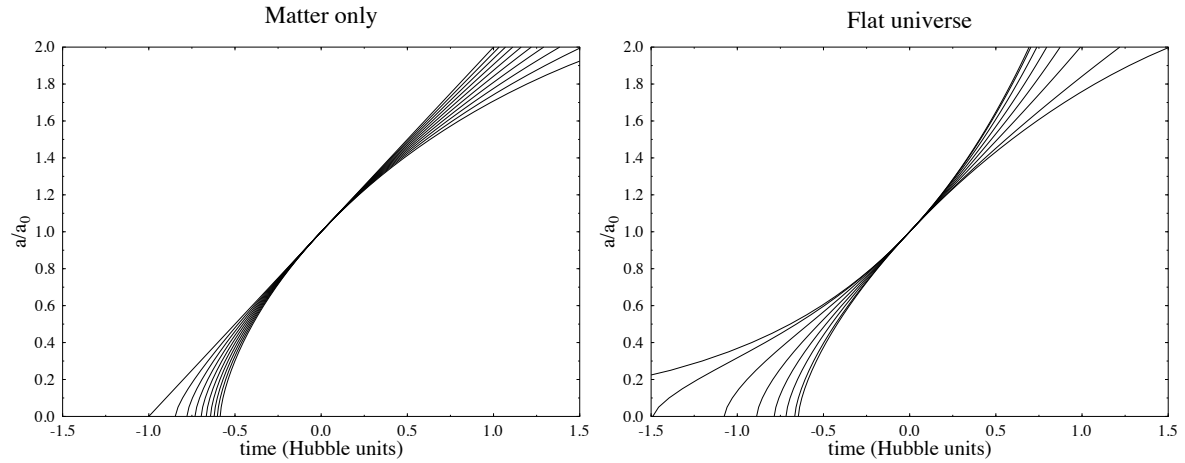


Figure 5: The expansion of the universe $a(t)$ for a) the matter-only universe $\Omega_\Lambda = 0$, $\Omega_m = 0, 0.2, \ldots, 1.8$ (from top to bottom) b) the flat universe $\Omega_0 = 1$ ($\Omega_\Lambda = 1 - \Omega_m$), $\Omega_m = 0$, 0.05, 0.2, 0.4, 0.6, 0.8, 1.0, 1.05 (from top to bottom). The time axis gives $H_0(t - t_0)$, i.e, 0.0 corresponds to the present time.

3.2.3 Age of the FRW universe

From (108) we get

$$\boxed{\frac{da}{dt} \;=\; H_0\sqrt{\Omega_r a^{-2} + \Omega_m a^{-1} + \Omega_k + \Omega_\Lambda a^2}\,.} \tag{110}$$

We shall later have much use for this convenient form of the Friedmann equation. Integrate from it the time it takes for the universe to expand from a_1 to a_2, or from redshift z_1 to z_2,

$$\begin{aligned}\int_{t_1}^{t_2} dt \;&=\; \int_{a_1}^{a_2} \frac{da}{da/dt} \;=\; H_0^{-1} \int_{\frac{1}{1+z_1}}^{\frac{1}{1+z_2}} \frac{da}{\sqrt{\Omega_r a^{-2} + \Omega_m a^{-1} + \Omega_k + \Omega_\Lambda a^2}} \\ &=\; \int_{z_2}^{z_1} \frac{dz}{(1+z)H(z)} \;=\; H_0^{-1} \int_{z_2}^{z_1} \frac{dz}{\sqrt{\Omega_r(1+z)^6 + \Omega_m(1+z)^5 + \Omega_k(1+z)^4 + \Omega_\Lambda(1+z)^2}}\,.\end{aligned} \tag{111}$$

This is integrable to an elementary function if two of the four terms under the root sign are absent.

From this we get the *age-redshift relation*

$$t(z) = \int_0^t dt = H_0^{-1} \int_0^{\frac{1}{1+z}} \frac{da}{\sqrt{\Omega_r a^{-2} + \Omega_m a^{-1} + \Omega_k + \Omega_\Lambda a^2}}. \tag{112}$$

(This gives $t(z)$, that is, $t(a)$. Inverting this function gives us $a(t)$, the scale factor as a function of time. Now $a(t)$ is not necessarily an elementary function, even if $t(a)$ is. Sometimes one can get a parametric representation $a(\psi)$, $t(\psi)$ in terms of elementary functions.)

In Fig. 5 we have integrated Eq. (110) from the initial conditions $a = 1$, $\dot{a} = H_0$, both backwards and forwards from the present time $t = t_0$ to find $a(t)$ as a function of time.

For the present *age of the universe* we get

$$t_0 = \int_0^{t_0} dt = H_0^{-1} \int_0^1 \frac{da}{\sqrt{\Omega_r a^{-2} + \Omega_m a^{-1} + \Omega_k + \Omega_\Lambda a^2}}\,. \tag{113}$$

The simplest cases, where only one of the terms under the square root is nonzero, give:

radiation dominated	$(\Omega_r = \Omega_0 = 1)$:	$t_0 = \frac{1}{2}H_0^{-1}$
matter dominated	$(\Omega_m = \Omega_0 = 1)$:	$t_0 = \frac{2}{3}H_0^{-1}$
curvature dominated	$(\Omega_0 = 0)$:	$t_0 = H_0^{-1}$
vacuum dominated	$(\Omega_\Lambda = \Omega_0 = 1)$:	$t_0 = \infty$.

These results can be applied also at other times (by considering some other time to be the "present time"), e.g., during the radiation-dominated epoch the age of the universe was related to the Hubble parameter by $t = \frac{1}{2}H^{-1}$ and during the matter-dominated epoch by $t = \frac{2}{3}H^{-1}$ (assuming that we can ignore the effect of the earlier epochs on the age). Returning to the present time, we know that Ω_r is so small that ignoring that term causes negligible error.

Example: Age of the open universe. Consider now the case of the open universe ($K < 0$ or $\Omega_0 < 1$), but without vacuum energy ($\Omega_\Lambda = 0$), and approximating $\Omega_r \approx 0$. Integrating Eq. (113) (e.g., with substitution $x = \frac{\Omega_m}{1-\Omega_m}\sinh^2\frac{\psi}{2}$) gives for the age of the open universe

$$\begin{aligned} t_0 &= H_0^{-1}\int_0^1 \frac{da}{\sqrt{1-\Omega_m+\Omega_m a^{-1}}} \\ &= H_0^{-1}\left[\frac{1}{1-\Omega_m} - \frac{\Omega_m}{2(1-\Omega_m)^{3/2}}\operatorname{arcosh}\left(\frac{2}{\Omega_m}-1\right)\right]. \end{aligned} \tag{114}$$

A special case of the open universe is the empty, or curvature-dominated, universe ($\Omega_m = 0$ and $\Omega_\Lambda = 0$). Now the Friedmann equation says $dx/dt = H_0$, or $a = H_0 t$, and $t_0 = H_0^{-1}$.

From the cases considered so far we get the following table for the age of the universe:

Ω_m	Ω_Λ	t_0
0	0	H_0^{-1}
0.1	0	$0.90H_0^{-1}$
0.3	0	$0.81H_0^{-1}$
0.5	0	$0.75H_0^{-1}$
1	0	$(2/3)H_0^{-1}$

There are many ways of estimating the matter density Ω_m of the universe, some of which are discussed in Chapter 6. These estimates give $\Omega_m \sim 0.3$. With $\Omega_m = 0.3$, $\Omega_\Lambda = 0$ (no dark energy), and the HST Key Project value $h = 0.72$, we get $t_0 = 12.2 \times 10^9$ years. This is about the same as the lowest estimates for the ages of the oldest stars. Since it should take hundreds of millions of years for the first stars to form, the open universe (or in general, a no-dark-energy universe, $\Omega_\Lambda = 0$) seems also to have an age problem.

The cases ($\Omega_m > 1$, $\Omega_\Lambda = 0$) and ($\Omega_0 = \Omega_m + \Omega_\Lambda = 1$, $\Omega_\Lambda > 0$) are left as exercises. The more general case ($\Omega_0 \neq 1$, $\Omega_\Lambda \neq 0$) leads to elliptic functions.

3.2.4 Distance-redshift relation

From Eq. (32), the comoving distance to redshift z is

$$d^c(z) = \int_{t_1}^{t_0}\frac{dt}{a(t)} = \int\frac{da}{a}\frac{1}{da/dt} = \int_0^z \frac{dz'}{H(z')}. \tag{115}$$

We have da/dt from Eq. (110), giving

$$\begin{aligned} d^c(z) &= H_0^{-1}\int_{\frac{1}{1+z}}^1 \frac{da}{\sqrt{\Omega_\Lambda a^4 + \Omega_k a^2 + \Omega_m a + \Omega_r}} \\ &= H_0^{-1}\int_0^z \frac{dz'}{\sqrt{\Omega_r(1+z)^4 + \Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}}. \end{aligned} \tag{116}$$

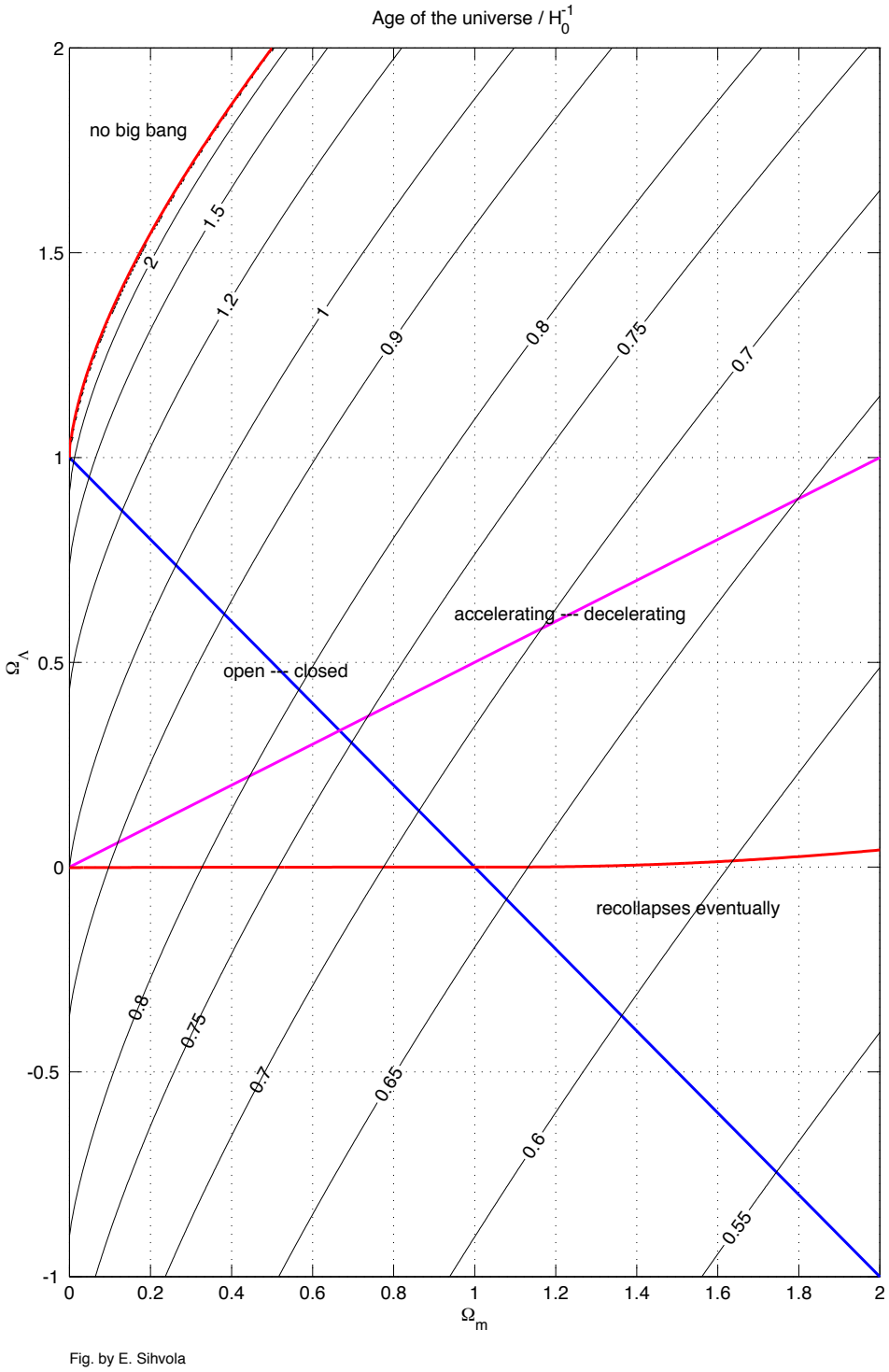


Figure 6: The age of the universe as a function of Ω_m and Ω_Λ.

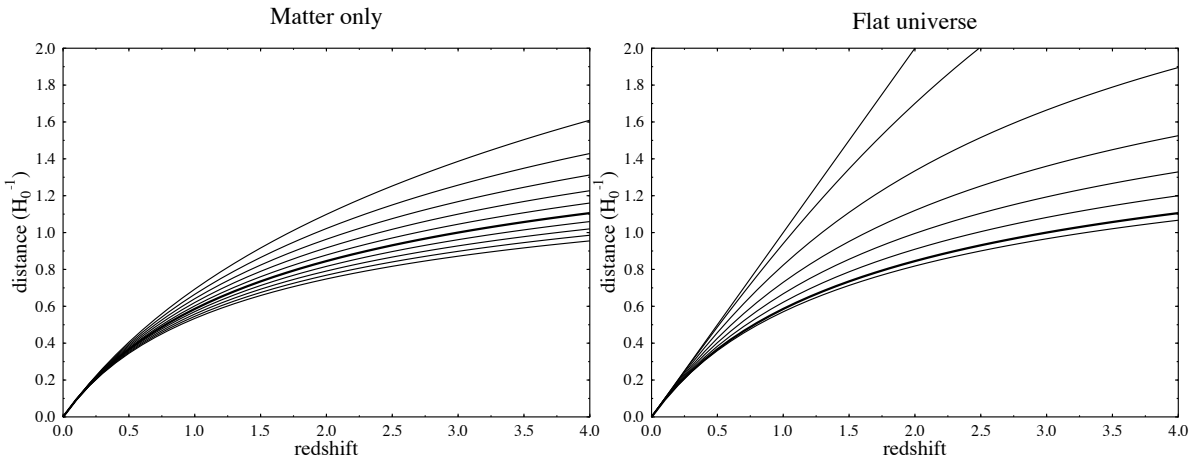


Figure 7: The distance-redshift relation, Eq. (116), for a) the matter-only universe $\Omega_\Lambda = 0$, $\Omega_m = 0$, 0.2,...,1.8 (from top to bottom) b) the flat universe $\Omega_0 = 1$ ($\Omega_\Lambda = 1 - \Omega_m$), $\Omega_m = 0$, 0.05, 0.2, 0.4, 0.6, 0.8, 1.0, 1.05 (from top to bottom). The thick line in both cases is the $\Omega_m = 1$, $\Omega_\Lambda = 0$ model.

This is the *distance-redshift relation.*

Example: How does the comoving distance depend on cosmological parameters. We can ignore Ω_r, since it makes such a small contribution. Noting that $\Omega_k = 1 - \Omega_0$ and $\Omega_m = \Omega_0 - \Omega_\Lambda$ we write (116) as

$$d^c(z) \;=\; H_0^{-1} \int_{\frac{1}{1+z}}^{1} \frac{da}{\sqrt{\Omega_0(a-a^2) - \Omega_\Lambda(a-a^4) + a^2}}\,. \tag{117}$$

We see that it depends on three independent cosmological parameters, for which we have taken H_0, Ω_0, and Ω_Λ. In this parametrization, the distance at a given redshift is proportional to the Hubble distance, H_0^{-1}. If we give the distance in units of H_0^{-1}, then it depends only on the two remaining parameters, Ω_0 and Ω_Λ. If we increase Ω_0 keeping Ω_Λ constant (meaning that we increase Ω_m), the distance corresponding to a given redshift decreases. This is because the universe has expanded faster in the past (see Fig. 5), so that there is less time between a given $a = 1/(1+z)$ and the present. The distance to the galaxy with redshift z is shorter, because photons have had less time to travel. Whereas if we increase Ω_Λ with a fixed Ω_0 (meaning that we decrease Ω_m), we have the opposite situation and the distance increases. Note that $(a - a^2)$ and $(a - a^4)$ are always positive since $0 < a \leq 1$.

If a galaxy (with some redshift z) has stayed at the same coordinate value r, i.e., it has no peculiar velocity, then the comoving distance to it is equal to its present distance. The actual distance to the galaxy at the time t_1 the light left the galaxy is

$$d_1^p(z) = \frac{d^c(z)}{1+z}. \tag{118}$$

We encounter the beginning of time, $t = 0$, at $a = 0$ or $z = \infty$. Thus the comoving distance light has travelled during the entire age of the universe, the horizon distance, is

$$d^c_{\rm hor} = H_0^{-1} \int_0^1 \frac{da}{\sqrt{\Omega_\Lambda a^4 + \Omega_k a^2 + \Omega_m a + \Omega_r}}\,. \tag{119}$$

The simplest cases[15], where only one of the terms under the square root is nonzero, give:

radiation dominated	$(\Omega_r = \Omega_0 = 1)$:	$d^c_{\rm hor} = H_0^{-1} = 2t_0$
matter dominated	$(\Omega_m = \Omega_0 = 1)$:	$d^c_{\rm hor} = 2H_0^{-1} = 3t_0$
curvature dominated	$(\Omega_0 = 0)$:	$d^c_{\rm hor} = \infty$
vacuum dominated	$(\Omega_\Lambda = \Omega_0 = 1)$:	$d^c_{\rm hor} = \infty\,.$

These results can be applied also at other times, e.g., during the radiation-dominated epoch the horizon distance was related to the Hubble parameter and age by $d^p_{\rm hor} = H^{-1} = 2t$ and during the matter-dominated epoch by $d^p_{\rm hor} = 2H^{-1} = 3t$ (assuming that epoch had already lasted long enough so that we can ignore the effect of the earlier epochs on the age and horizon distance). Returning to the present time, we know that Ω_r is so small that ignoring that term causes negligible error.

Example: Distance and redshift in the flat matter-dominated universe. Let us look at the simplest case, $(\Omega_m, \Omega_\Lambda) = (1, 0)$ (with $\Omega_r \approx 0$), in more detail. Now Eq. (116) is just

$$d^c(z) = H_0^{-1} \int_{\frac{1}{1+z}}^{1} \frac{da}{a^{1/2}} = 2H_0^{-1} \left(1 - \frac{1}{\sqrt{1+z}}\right). \tag{120}$$

Expanding $1/\sqrt{1+z} = 1 - \frac{1}{2}z + \frac{3}{8}z^2 - \frac{5}{16}z^3 \cdots$ we get

$$d^c(z) = H_0^{-1}(z - \frac{3}{4}z^2 + \frac{5}{8}z^3 - \cdots) \tag{121}$$

so that for small redshifts, $z \ll 1$ we get the Hubble law, $z = H_0 d_0$. At the time when the light we see left the galaxy, its distance was

$$d^p_1(z) = \frac{1}{1+z} d^c(z) = a(t)r = 2H_0^{-1} \left(\frac{1}{1+z} - \frac{1}{(1+z)^{3/2}}\right) \tag{122}$$

$$= H_0^{-1}(z - \frac{7}{4}z^2 + \frac{19}{8}z^3 - \cdots) \tag{123}$$

so the Hubble law is valid for small z independent of our definition of distance.

The distance $d^p(t) = a(t)r$ to the galaxy grows with the velocity $\dot{d}^p = r\dot{a} = raH$, so that today $\dot{d}^p = rH_0 = d^c H_0 = 2(1 - 1/\sqrt{1+z})$. This equals 1 (the speed of light) at $z = 3$.

We note that $d^p_1(z)$ has a maximum $d^p_1(z) = \frac{8}{27}H_0^{-1}$ at $z = \frac{5}{4}$ $(1 + z = \frac{9}{4})$. This corresponds to the comoving distance $d^c(z) = \frac{2}{3}H_0^{-1}$. See Fig. 9. Galaxies that are further out were thus closer when the light left, since the universe was then so much smaller.

The distance to the horizon in this simplest case is

$$d^c_{\rm hor} \equiv d^c(z = \infty) = 2H_0^{-1} = 3t_0. \tag{124}$$

Example: Effect of radiation. Consider the flat universe ($\Omega_k = 0$). Ignoring radiation, with $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, we get for the age of the universe, time since photon decoupling ($z = 1090$), time

[15] Of these cases, the strict forms of the two last ones, pure curvature ($\Omega_0 = 0$) and pure vacuum ($\Omega_\Lambda = \Omega_0 = 1$) do not actually fit in the FRW framework, where the starting assumption was *spatial* homogeneity that formed the basis of separation between time and space. This separation requires a physical quantity that evolves in time, in practice the energy density $\rho(t)$, so that the $t = const$ slices can be defined as the $\rho = const$ hypersurfaces. Now in these two cases, $\rho = const$ (either 0 or the vacuum value) also in time, and does not provide this separation. These cases are called the *Milne universe* and the *de Sitter space* (or anti-de Sitter space for $\rho_{\rm vac} < 0$) and are discussed in the General Relativity course. For our purposes, we should instead consider these as limiting cases where there is also a density component that is just very small (a nonzero Ω_m or Ω_r that is $\ll 1$). Then this other component necessarily becomes important in the early universe, as $a \to 0$. This means that $d_{\rm hor}$ is not ∞, just very large. The same applies to the "infinite" age of the vacuum-dominated universe.

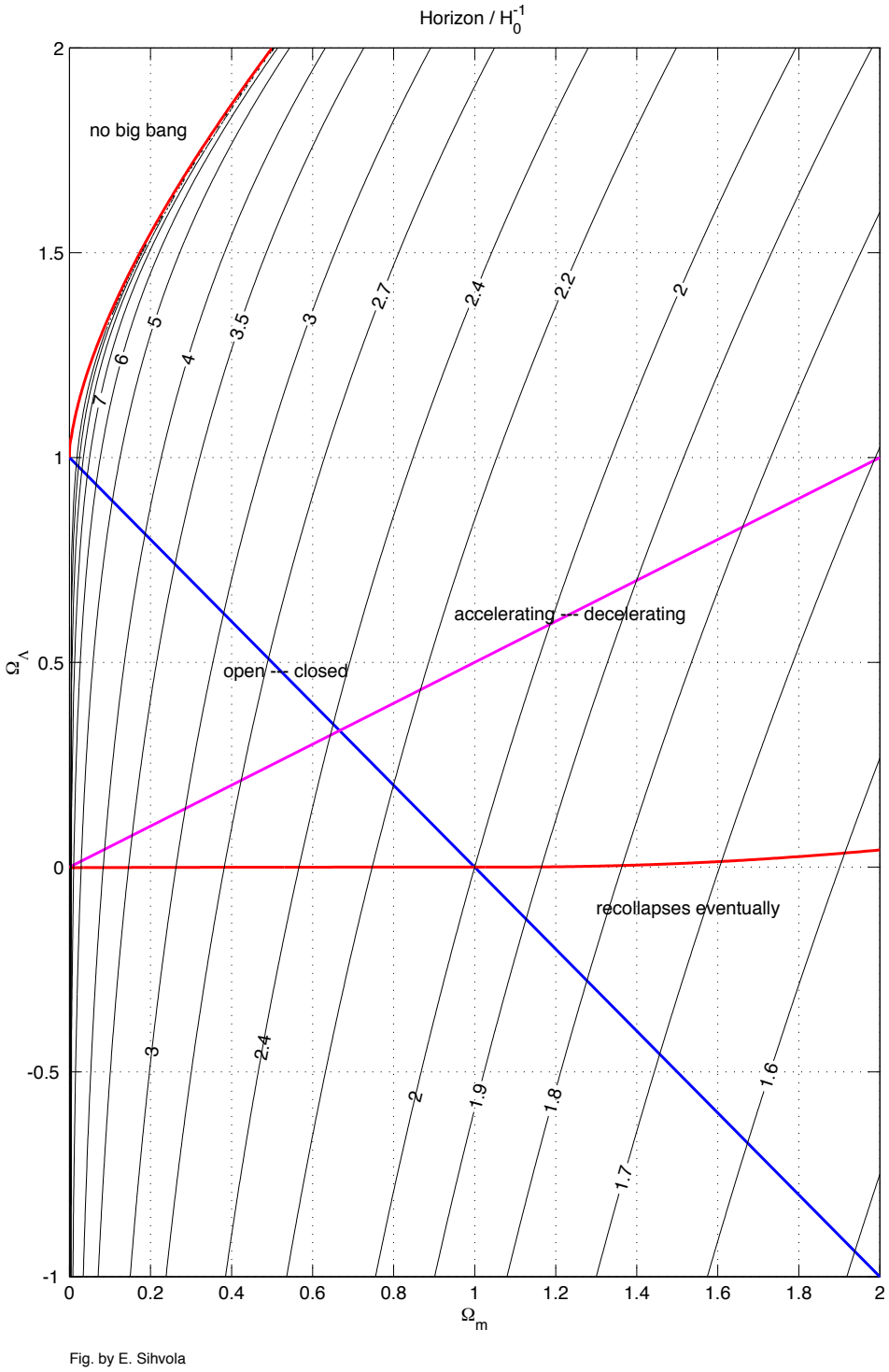


Figure 8: The horizon as a function of Ω_m and Ω_Λ.

since $z = 10$, horizon distance, distance to last scattering sphere, and distance to $z = 10$ (the most distant galaxies observed):

$$\begin{aligned}
t_0 &= H_0^{-1}\int_0^1 \frac{da}{\sqrt{\Omega_m a^{-1}+\Omega_\Lambda a^2}} = 0.964099 H_0^{-1}\\
t_0 - t_{\rm dec} &= H_0^{-1}\int_{1/1091}^1 \frac{da}{\sqrt{\Omega_m a^{-1}+\Omega_\Lambda a^2}} = 0.964066 H_0^{-1}\\
t_0 - t(z=10) &= H_0^{-1}\int_{1/11}^1 \frac{da}{\sqrt{\Omega_m a^{-1}+\Omega_\Lambda a^2}} = 0.930747 H_0^{-1}\\
d^c_{\rm hor} &= H_0^{-1}\int_0^1 \frac{da}{\sqrt{\Omega_m a+\Omega_\Lambda a^4}} = 3.30508 H_0^{-1}\\
d^c(z=1090) &= H_0^{-1}\int_{1/1091}^1 \frac{da}{\sqrt{\Omega_m a+\Omega_\Lambda a^4}} = 3.19453 H_0^{-1}\\
d^c(z=10) &= H_0^{-1}\int_{1/11}^1 \frac{da}{\sqrt{\Omega_m a+\Omega_\Lambda a^4}} = 2.20425 H_0^{-1}\,. \qquad (125)
\end{aligned}$$

Include then radiation with $\Omega_r = 0.000085$ (we learn in Chapter 4 that this value corresponds to $h = 0.7$) and subtract it from matter so that $\Omega_m = 0.299915$, $\Omega_\Lambda = 0.7$. Now we get

$$\begin{aligned}
t_0 &= H_0^{-1}\int_0^1 \frac{da}{\sqrt{\Omega_r a^{-2}+\Omega_m a^{-1}+\Omega_\Lambda a^2}} = 0.963799 H_0^{-1}\\
t_0 - t_{\rm dec} &= H_0^{-1}\int_{1/1091}^1 \frac{da}{\sqrt{\Omega_r a^{-2}+\Omega_m a^{-1}+\Omega_\Lambda a^2}} = 0.963772 H_0^{-1}\\
t_0 - t(z=10) &= H_0^{-1}\int_{1/11}^1 \frac{da}{\sqrt{\Omega_r a^{-2}+\Omega_m a^{-1}+\Omega_\Lambda a^2}} = 0.930586 H_0^{-1}\\
d^c_{\rm hor} &= H_0^{-1}\int_0^1 \frac{da}{\sqrt{\Omega_r+\Omega_m a+\Omega_\Lambda a^4}} = 3.244697 H_0^{-1}\\
d^c(z=1090) &= H_0^{-1}\int_{1/1091}^1 \frac{da}{\sqrt{\Omega_r+\Omega_m a+\Omega_\Lambda a^4}} = 3.17967 H_0^{-1}\\
d^c(z=10) &= H_0^{-1}\int_{1/11}^1 \frac{da}{\sqrt{\Omega_r+\Omega_m a+\Omega_\Lambda a^4}} = 2.20348 H_0^{-1}\,. \qquad (126)
\end{aligned}$$

(All integrals were done numerically with WolframAlpha, although the first three in (125) could have been done analytically.) The effect of radiation on these numbers is thus rather small compared to accuracy of observations in cosmology.

Just like any planar map of the surface of Earth must be distorted, so is it for the curved spacetime. Even in the flat-universe case, the spacetime is curved due to the expansion. Thus any spacetime diagram is a distortion of the true situation. In Figs. 9 and 10 there are three different ways of drawing the same spacetime diagram. In the first one the vertical distance is proportional to the cosmic time t, the horizontal distance to the proper distance at that time, $d^p(t)$. The second one is in the comoving coordinates (t, r), so that the horizontal distance is proportional to the comoving distance d^c (Note that for $\Omega = 1$, i.e., $K = 0$, we have $d^c = r$, see Eq. (29)). The third one is in the conformal coordinates (η, r). This one has the advantage that light cones are always at a 45° angle. This is thus the "Mercator projection"[16] spacetime.

Angular diameter distance: The comoving angular diameter distance is given by the coordinate $r = f_K(d^c) = d^c_A$ so that using the distance-redshift relation, Eq. (116), and dropping

[16]The Mercator projection is a way of drawing a map of the Earth so that the points of compass correspond to the same direction everywhere on the map, e.g., northeast and northwest are always 45° from the north direction.

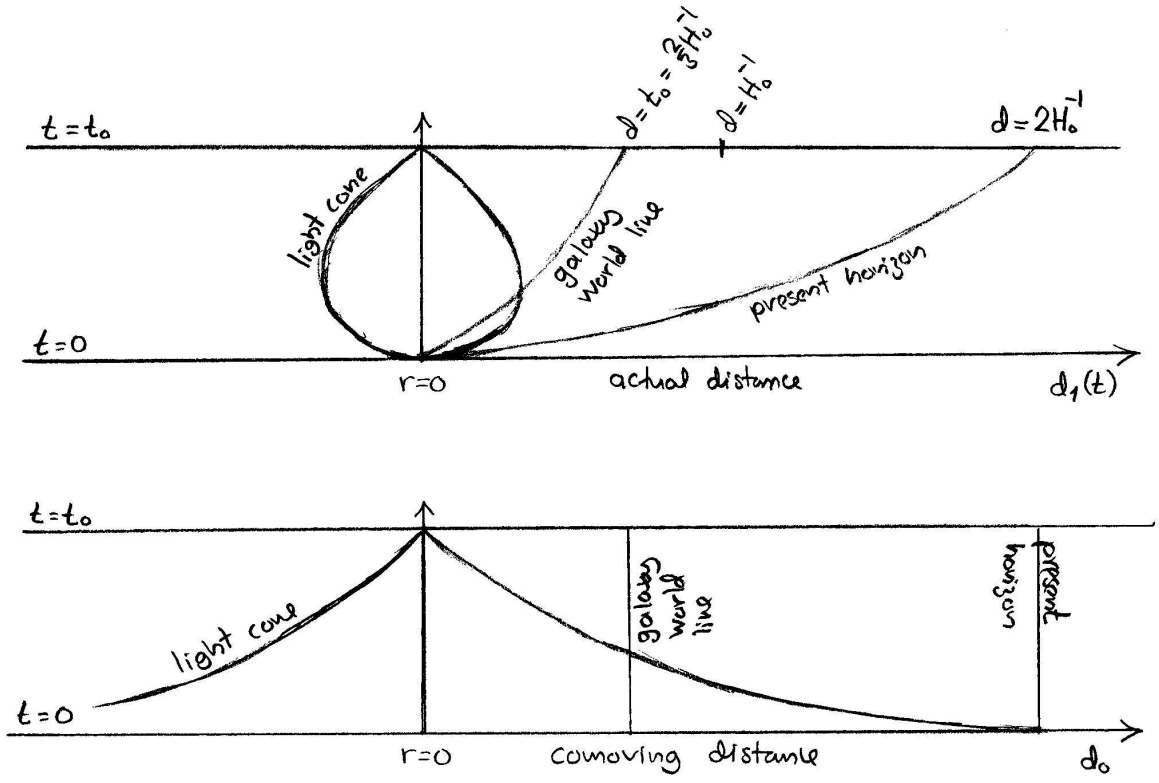


Figure 9: Spacetime diagrams for a flat matter-dominated universe giving a) the proper distance (denoted here as $d_1(t)$ b) the comoving distance (denoted d_0) from origin as a function of cosmic time.

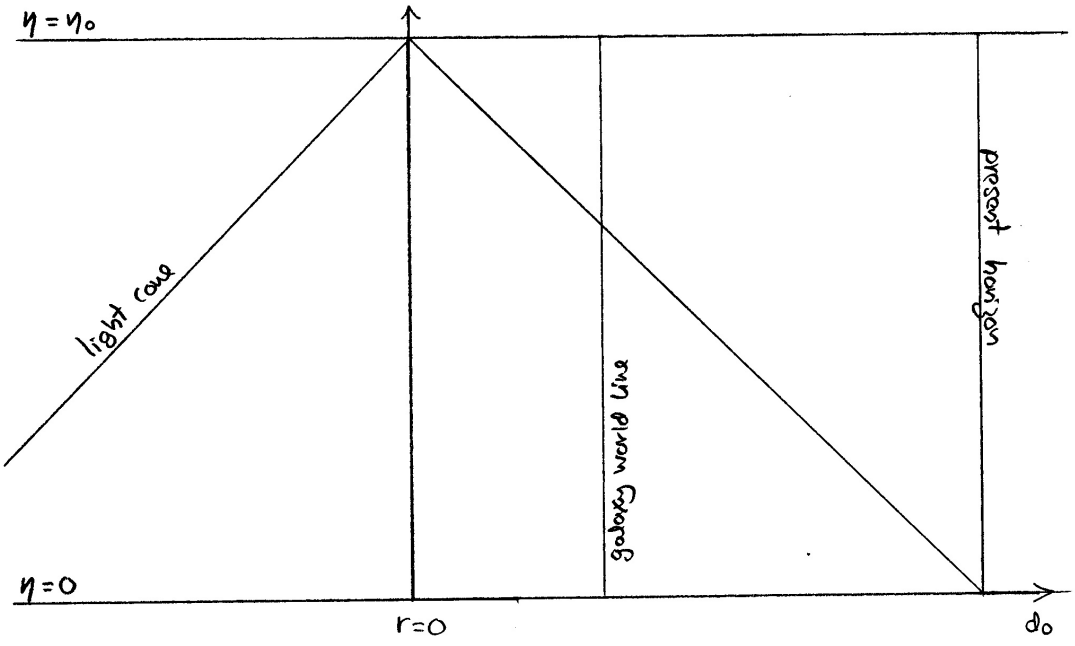


Figure 10: Spacetime diagram for a flat matter-dominated universe in conformal coordinates.

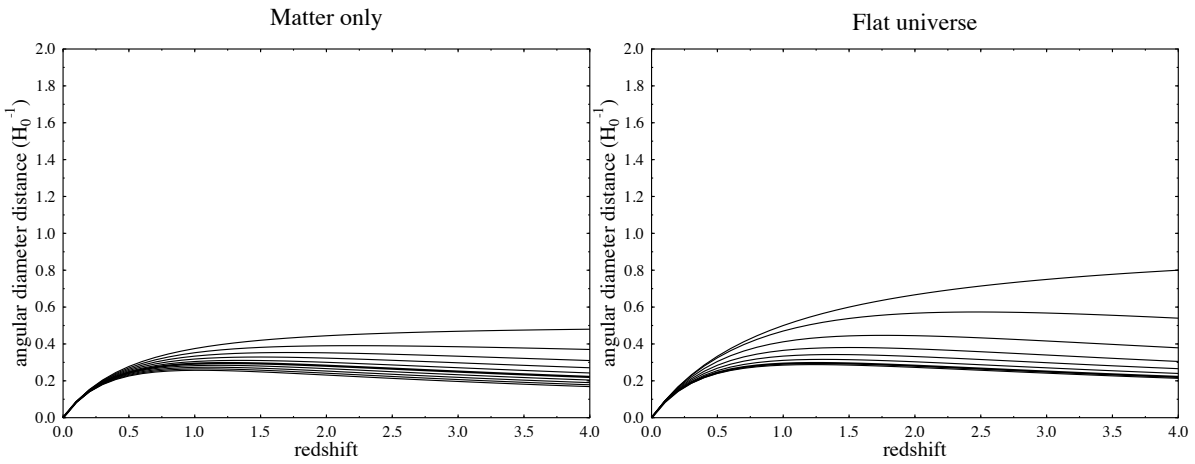


Figure 11: The angular diameter distance -redshift relation, Eq. (129), for a) the matter-only universe $\Omega_\Lambda = 0$, $\Omega_m = 0$, 0.2,...,1.8 (from top to bottom) b) the flat universe $\Omega_0 = 1$ ($\Omega_\Lambda = 1 - \Omega_m$), $\Omega_m = 0$, 0.05, 0.2, 0.4, 0.6, 0.8, 1.0, 1.05 (from top to bottom). The thick line in both cases is the $\Omega_m = 1$, $\Omega_\Lambda = 0$ model. Note how the angular diameter distance decreases for large redshifts, meaning that the object that is farther away may appear larger on the sky. In the flat case, this is an expansion effect, an object with a given size occupies a larger comoving volume in the earlier, smaller universe. In the matter-only case, the effect is enhanced by space curvature effects for the closed ($\Omega_m > 1$) models.

Ω_r, we have

$$\begin{aligned} d_A^c(z) &= f_K\left[\frac{1}{H_0}\int_{\frac{1}{1+z}}^{1}\frac{da}{\sqrt{\Omega_0(a-a^2)-\Omega_\Lambda(a-a^4)+a^2}}\right] \\ &= H_0^{-1}\frac{1}{\sqrt{\Omega_k}}f_k\left(\sqrt{\Omega_k}H_0\int_0^z\frac{dz'}{H(z')}\right), \end{aligned} \tag{127}$$

where we defined (note k instead of K)

$$f_k(x) \equiv \begin{cases} \sin(x), & (K>0) \\ x, & (K=0) \\ \sinh(x), & (K<0) \end{cases} \tag{128}$$

for the comoving angular diameter distance and

$$d_A(z) = d_A^c(z)/(1+z) = \frac{1}{1+z}f_K\left[\frac{1}{H_0}\int_{\frac{1}{1+z}}^{1}\frac{da}{\sqrt{\Omega_0(a-a^2)-\Omega_\Lambda(a-a^4)+a^2}}\right], \tag{129}$$

for the angular diameter distance.

For a flat universe the comoving angular diameter distance is equal to the comoving distance,

$$d_A^c(z) = d^c(z) = H_0^{-1}\int_{\frac{1}{1+z}}^{1}\frac{da}{\sqrt{\Omega_0(a-a^2)-\Omega_\Lambda(a-a^4)+a^2}}. \tag{130}$$

We shall later (in Cosmology II) use the angular diameter distance to relate the observed anisotropies of the cosmic microwave background to the physical length scale of the density fluctuations they represent. Since this length scale can be calculated from theory, their observed angular size gives us information of the cosmological parameters.

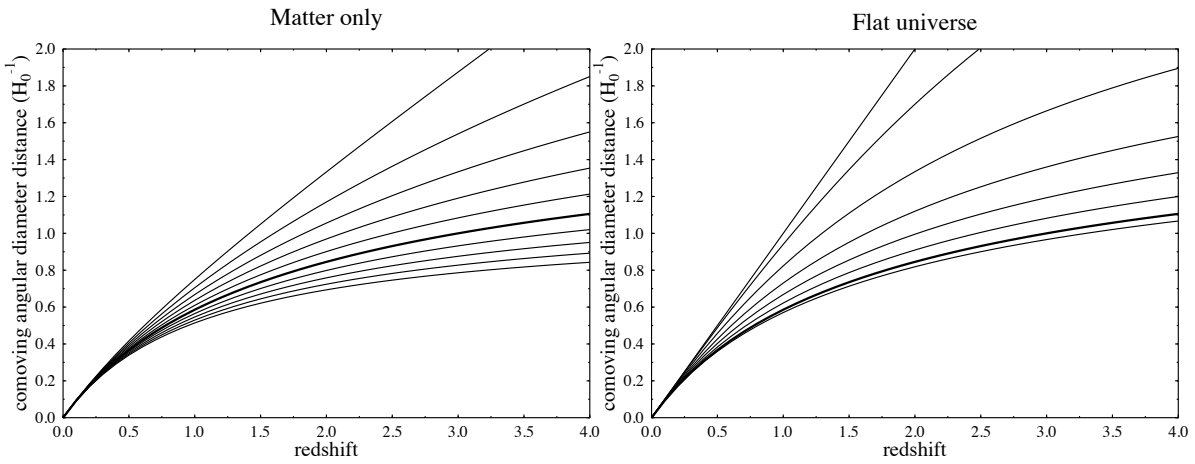


Figure 12: Same as Fig. 11, bur for the *comoving* angular diameter distance. Now the expansion effect is eliminated. For the closed models (for $\Omega_m > 1$ in the case of $\Omega_\Lambda = 0$) even the comoving angular diameter distance may begin to decrease at large enough redshifts. This happens when we are looking beyond $\sqrt{K}\chi = \pi/2$, where the universe "begins to close up". The figure does not go to high enough z to show this for the parameters used. Note how for the flat universe the comoving angular diameter distance is equal to the comoving distance (see Fig. 7).

Luminosity distance: From Eq. (46),

$$d_L \equiv \sqrt{\frac{L}{4\pi l}} = (1+z)r = (1+z)d_A^c(z) = (1+z)^2 d_A(z)$$

As we discussed in Chapter 1, astronomers have the habit of giving luminosities as magnitudes. From the definitions of the absolute and apparent magnitude,

$$M \equiv -2.5 \lg \frac{L}{L_0}\,, \qquad m \equiv -2.5 \lg \frac{l}{l_0}\,, \tag{131}$$

and Eq. (41), we have that the distance modulus $m - M$ is given by the luminosity distance as

$$m - M = -2.5 \lg \frac{l}{L}\frac{L_0}{l_0} = 5 \lg d_L + 2.5 \lg 4\pi \frac{l_0}{L_0} = -5 + 5 \lg d_L(\mathrm{pc})\,. \tag{132}$$

(As explained in Chapter 1, the constants L_0 and l_0 are chosen so as to give the value -5 for the constant term, when d_L is given in parsecs.) For a set of standard candles, all having the same absolute magnitude M, we find that their apparent magnitudes m should be related to their redshift z as

$$\begin{aligned}
m(z) &= M - 5 + 5 \lg d_L(\mathrm{pc}) \\
&= M - 5 - 5 \lg H_0 + 5 \lg \left\{ (1+z) H_0 f_K \left(H_0^{-1} \int_{\frac{1}{1+z}}^{1} \frac{dx}{\sqrt{\Omega_0(x - x^2) - \Omega_\Lambda(x - x^4) + x^2}} \right) \right\} \\
&= M - 5 - 5 \lg H_0 + 5 \lg \left\{ (1+z) \sqrt{\frac{-K}{\Omega_k}} \times \right. \\
&\quad \left. \times f_K \left[\sqrt{\frac{\Omega_k}{-K}} \int_{\frac{1}{1+z}}^{1} \frac{dx}{\sqrt{\Omega_0(x - x^2) - \Omega_\Lambda(x - x^4) + x^2}} \right] \right\} \\
&= M - 5 - 5 \lg H_0 + 5 \lg \left\{ \frac{1+z}{\sqrt{|\Omega_k|}} f_k \left(\sqrt{|\Omega_k|} \int_{\frac{1}{1+z}}^{1} \frac{dx}{\sqrt{\Omega_0(x - x^2) - \Omega_\Lambda(x - x^4) + x^2}} \right) \right\}.
\end{aligned} \tag{133}$$

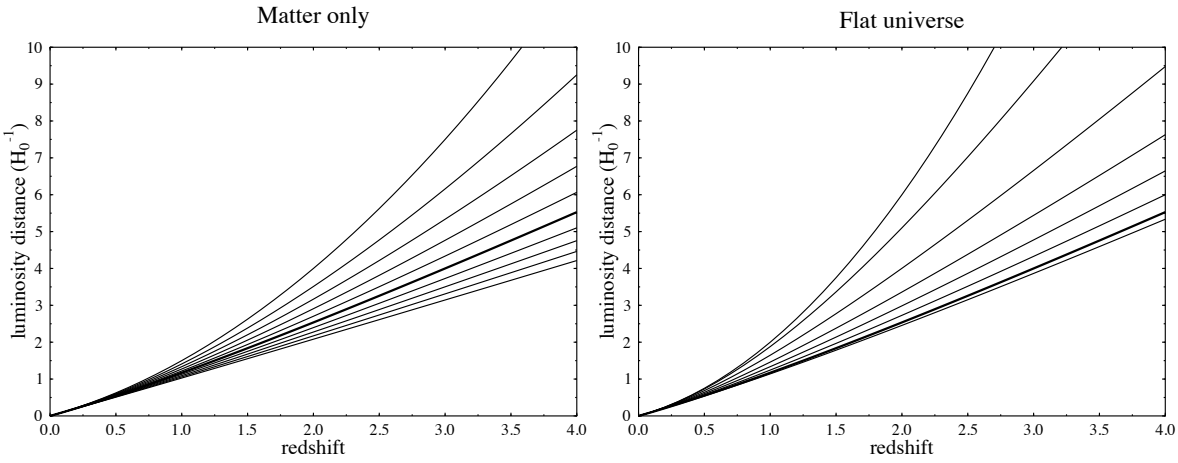


Figure 13: Same as Fig. 11, bur for the *luminosity* distance. Note how the vertical scale now extends to 10 Hubble distances instead of just 2, to have room for the much more rapidly increasing luminosity distance.

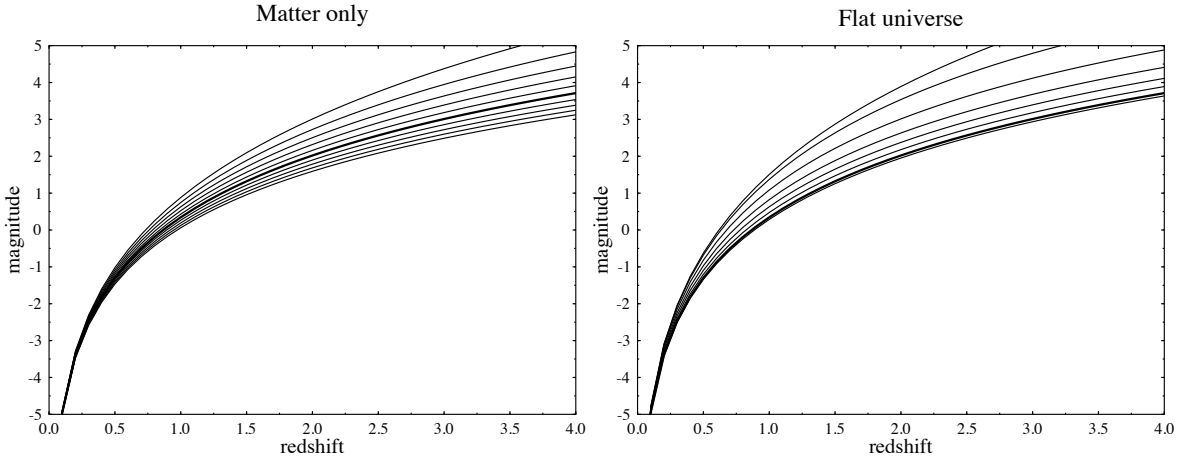


Figure 14: Same as Fig. 11, bur for the *magnitude-redshift* relation. The constant $M - 5 - 5 \lg H_0$ in Eq. (134), which is different for different classes of standard candles, has been arbitrarily set to 0.

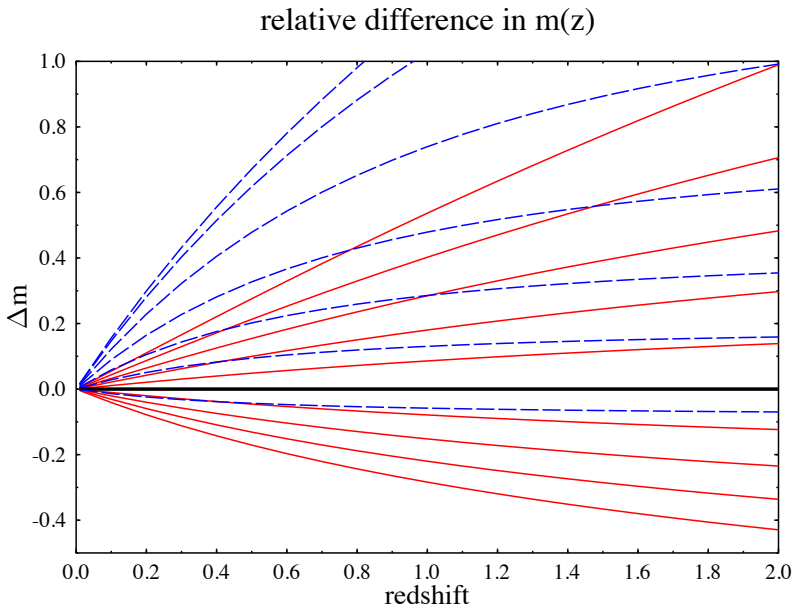


Figure 15: The difference between the magnitude-redshift relation of the different models in Fig. 14 from the reference model $\Omega_m = 1$, $\Omega_\Lambda = 0$ (which appears as the horizontal thick line). The red (solid) lines are for the matter-only ($\Omega_\Lambda = 0$) models and the blue (dashed) lines are for the flat ($\Omega_0 = 1$) models.

We find that the Hubble constant H_0 contributes only to a constant term in this *magnitude-redshift relation*. If we just know that all the objects have the same M, but do not know the value of M, we cannot use the observed $m(z)$ to determine H_0, since both M and H_0 contribute to this constant term. On the other hand, the *shape* of the $m(z)$ curve depends only on the two parameters Ω_0 and Ω_Λ (see Fig. 15).

Type Ia supernovae (SNIa) are fairly good standard candles.[17] Two groups, the Supernova Cosmology Project[18] and the High-Z Supernova Search Team[19] have been using observations of such supernovae up to redshifts $z \sim 2$ to try to determine the values of the cosmological parameters Ω_0 and Ω_Λ.

In 1998 they announced [3, 4] that their observations are inconsistent with a matter-dominated universe, i.e., with $\Omega_\Lambda = 0$. In fact their observations required that the expansion of the universe is accelerating. This result was named the "Breakthrough of the Year" by the Science magazine [5]. Later more accurate observations by these and other groups have confirmed this result. This SNIa data is one of the main arguments for the existence of dark energy in the universe.[20] See Fig. 16 for SNIa data from 2004, and Fig. 17 for a determination of Ω_m and Ω_Λ from this data. As you can see, the data is not good enough for a simultaneous accurate determination of both Ω_m and Ω_Λ. But by assuming a flat universe, $\Omega_0 = 1$, Riess et al. [6] found $\Omega_\Lambda = 0.71^{+0.03}_{-0.05}$ ($\Rightarrow \Omega_m = 0.29^{+0.05}_{-0.03}$). (The main evidence for a flat universe, $\Omega_0 \approx 1$ comes from the CMB anisotropy, which we shall discuss later, in Cosmology II)

[17] To be more precise, they are "standardizable candles", i.e., their peak absolute magnitudes M vary, but these are related to their observable properties in a way that can be determined.

[18] http://supernova.lbl.gov/

[19] http://cfa-www.harvard.edu/cfa/oir/Research/supernova/HighZ.html, http://www.nu.to.infn.it/exp/all/hzsnst/

[20] The other main argument comes from combining CMB anisotropy and large-scale-structure data, and will be discussed in Cosmology II.

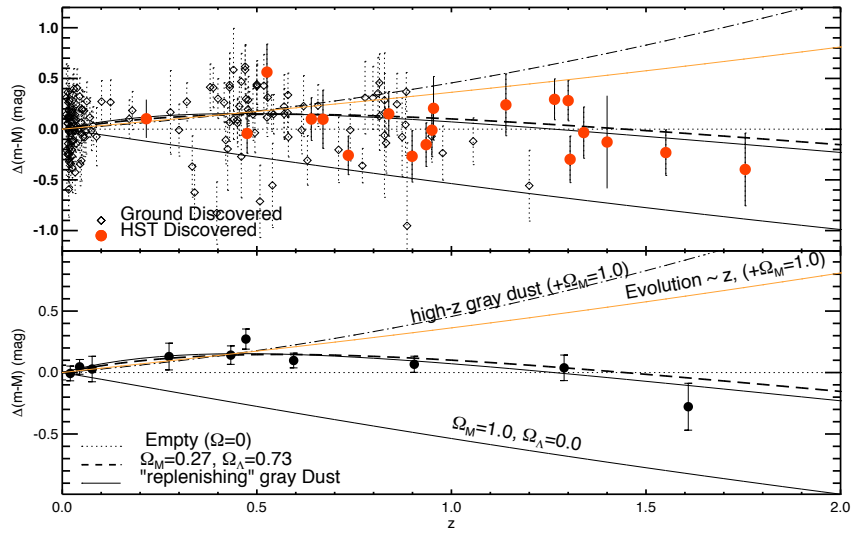


Figure 16: Supernova Ia luminosity-redshift data. The top panel shows all supernova of the data set. The bottom panel show the averages from different redshift bins. The curves corresponds to three different FRW cosmologies, and some alternative explanations: "dust" refers to the possibility that the universe is not transparent, but some photons get absorbed on the way; "evolution" to the possibility that the SNIa are not standard candles, but were different in the younger universe, so that $M = M(z)$. From Riess et al., astro-ph/0402512 [6].

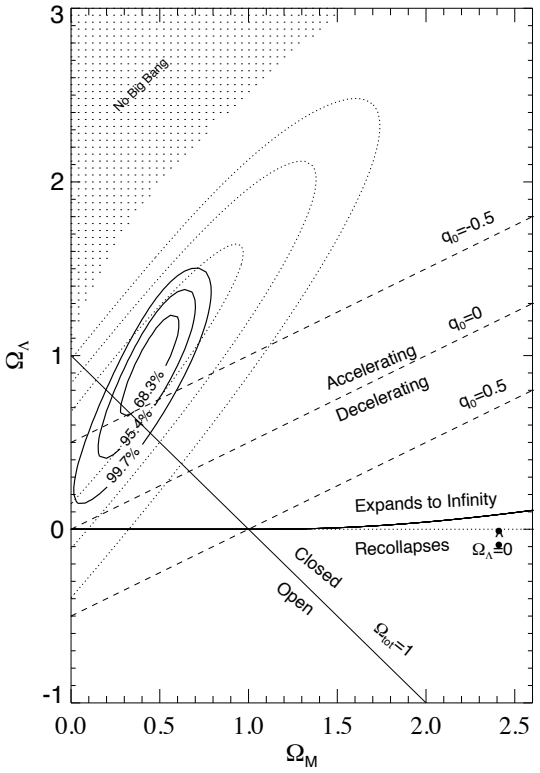


Figure 17: Ω_m and Ω_Λ determined from the Supernova Ia data. The dotted contours are the old 1998 results[3]. From Riess et al., astro-ph/0402512 [6].

We have in the preceding assumed that the mysterious dark energy component of the universe is vacuum energy, or indistinguishable from a cosmological constant, so that $p_{\rm de} = -\rho_{\rm de}$. Making the assumption[21] that the equation-of-state parameter $w_{\rm de} \equiv p_{\rm de}/\rho_{\rm de}$ for dark energy is a constant, but not necessarily equal to -1, Riess et al. [6] found the limits $-1.48 < w_{\rm de} < -0.72$, when they assumed a flat universe, and used an independent limit on Ω_m from other cosmological observations.

[21] There is no theoretical justification for this assumption. It is just done for simplicity since the present data is not good enough for determining a larger number of free parameters in the dark-energy equation of state to a meaningful accuracy.

3.3 Concordance Model

Currently the simplest cosmological model that fits all cosmological observations reasonably well is the ΛCDM model, also called the *Concordance Model* (since it fits different kinds of observations) or the *standard model of cosmology* (since it is often now assumed in studies that relate to cosmology but focus on other questions than the cosmological model). In the name, Λ stands for the cosmological constant, i.e., dark energy is assumed to be vacuum energy and to dominate the energy density of the universe today, and CDM for cold dark matter, which is assumed to make up most of the matter in the universe.

The ΛCDM model includes a number of assumptions related to *primordial perturbations*, i.e., deviations from the homogeneous FRW model, that we will discuss in Cosmology II, but for the present discussion the relevant part of the ΛCDM model is that the "unperturbed" homogeneous "background" model, a good approximation for large distance scales, is the flat FRW universe with $\Omega_0 = 1 \approx \Omega_\Lambda + \Omega_m$. Often the term "Concordance Model" is used for this FRW model, and the term ΛCDM model is used when also the other assumptions are included. We adopt this usage.

The expansion law $a(t)$ of the Concordance Model is solved from

$$\frac{da}{dt} \;=\; H_0\sqrt{\Omega_m a^{-1} + \Omega_\Lambda a^2}\,, \tag{134}$$

which is easier to integrate from

$$\begin{aligned} t(a) \;&=\; H_0^{-1}\int_0^a \frac{da}{\sqrt{\Omega_m a^{-1} + \Omega_\Lambda a^2}} \;=\; H_0^{-1}\int_0^a \frac{a^{1/2}da}{\sqrt{\Omega_m + \Omega_\Lambda a^3}} \\ &=\; \frac{2}{3}H_0^{-1}\frac{1}{\sqrt{\Omega_\Lambda}}\int_0^y \frac{dy}{\sqrt{1+y^2}} \;=\; \frac{2}{3}H_0^{-1}\frac{1}{\sqrt{\Omega_\Lambda}}\,\mathrm{arsinh}\left[\sqrt{\frac{\Omega_\Lambda}{\Omega_m}}a^{3/2}\right], \end{aligned} \tag{135}$$

where we used the substitution $y = \sqrt{\Omega_\Lambda/\Omega_m}\,a^{3/2}$. Inverting this, we have the expansion law

$$\boxed{a(t) \;=\; \left(\frac{\Omega_m}{\Omega_\Lambda}\right)^{1/3}\sinh^{2/3}\left(\tfrac{3}{2}\sqrt{\Omega_\Lambda}H_0 t\right).} \tag{136}$$

At early times, $t \ll (2/3\sqrt{\Omega_\Lambda})H_0^{-1}$, the expansion is decelerating and $a \propto t^{2/3}$ (the matter-dominated era):

$$a(t) \;\approx\; \left(\frac{9\Omega_m}{4}\right)^{1/3} H_0^{2/3}t^{2/3}\,. \tag{137}$$

At late times, $t \gg (2/3\sqrt{\Omega_\Lambda})H_0^{-1}$, the expansion is exponential and accelerating (the vacuum-dominated era):

$$a(t) \;\approx\; \left(\frac{\Omega_m}{4\Omega_\Lambda}\right)^{1/3} e^{\sqrt{\Omega_\Lambda}H_0 t}\,. \tag{138}$$

From above, the age-redshift relation is

$$\boxed{t(z) \;=\; \frac{2}{3}H_0^{-1}\frac{1}{\sqrt{\Omega_\Lambda}}\,\mathrm{arsinh}\left[\sqrt{\frac{\Omega_\Lambda}{\Omega_m}}(1+z)^{-3/2}\right],} \tag{139}$$

and the present age of the universe is

$$t_0 \;=\; \frac{2}{3}H_0^{-1}\frac{1}{\sqrt{\Omega_\Lambda}}\,\mathrm{arsinh}\sqrt{\frac{\Omega_\Lambda}{\Omega_m}} \tag{140}$$

In the concordance model, there are two energy-density components, $\rho_{\rm vac} = const$ and $\rho_m \propto a^{-3}$, so that $\rho = \rho_{\rm vac} + \rho_m$ and $p = -\rho_{\rm vac}$. From the second Friedmann equation

$$\frac{\ddot{a}}{a} \;=\; -\frac{4\pi G}{3}(\rho + 3p) \;=\; -\frac{4\pi G}{3}(\rho_m - 2\rho_{\rm vac}) \tag{141}$$

we see that the deceleration turns into accelaration when $\rho_{\rm vac} = \frac{1}{2}\rho_m$. Since

$$\frac{\rho_{\rm vac}}{\rho_m} \;=\; \frac{\Omega_\Lambda}{\Omega_m}\left(\frac{a}{a_0}\right)^3 \;=\; \sinh^2\left(\tfrac{3}{2}\sqrt{\Omega_\Lambda}H_0 t\right)\,, \tag{142}$$

we get that this happens when

$$a \;=\; \frac{1}{1+z} \;=\; \left(\frac{\Omega_m}{2\Omega_\Lambda}\right)^{1/3} \quad \text{and} \quad t_{\rm acc} \;=\; \frac{2}{3}H_0^{-1}\frac{1}{\sqrt{\Omega_\Lambda}}\,{\rm arsinh}\frac{1}{\sqrt{2}}\,. \tag{143}$$

The vacuum and matter energy densities become equal later, when

$$a \;=\; \frac{1}{1+z} \;=\; \left(\frac{\Omega_m}{\Omega_\Lambda}\right)^{1/3} \quad \text{and} \quad t_{\rm eq} \;=\; \frac{2}{3}H_0^{-1}\frac{1}{\sqrt{\Omega_\Lambda}}\,{\rm arsinh}(1)\,. \tag{144}$$

Here

$${\rm arsinh}\frac{1}{\sqrt{2}} = 0.65848 \quad \text{and} \quad {\rm arsinh}(1) = 0.88137\,. \tag{145}$$

For $\Omega_\Lambda = 0.7$ and $\Omega_m = 0.3$,

$${\rm arsinh}\sqrt{\frac{\Omega_\Lambda}{\Omega_m}} \;=\; {\rm arsinh}(1.5275) \;=\; 1.2099 \tag{146}$$

and

$$t_{\rm acc} \;=\; 0.5247H_0^{-1} \qquad t_{\rm eq} \;=\; 0.7023H_0^{-1} \qquad t_0 \;=\; 0.9641H_0^{-1}\,. \tag{147}$$

In the concordance model the distance-redshift relation appears not to have a closed form in terms of elementary functions, so we need to integrate it numerically. Then we can also include the effect of the small Ω_r, which is important only at early times (the $z = 1000$ and $z = 1089$ lines in Table 1). The best-fit ΛCDM model to the first release of Planck satellite data [7] (we discuss this in Cosmology II) has

$$\Omega_\Lambda = 0.6830 \qquad \Omega_m = 0.3169 \qquad \Omega_r = 9.323\times10^{-5} \qquad H_0 = 67.15\,{\rm km/s/Mpc} \tag{148}$$

The Hubble distance and time are then

$$H_0^{-1} = 4464.5\,{\rm Mpc} = 14.571\times10^9\,{\rm yr}\,, \tag{149}$$

the age of the universe is $t_0 = 0.949H_0^{-1} = 13.83\,{\rm Gyr}$, $t_{\rm acc} = 0.5311H_0^{-1} = 7.74\,{\rm Gyr}$, $t_{\rm eq} = 0.7110H_0^{-1} = 10.36\,{\rm Gyr}$, i.e., the expansion started accelerating 6.1 billion years ago, and vacuum energy density exceeded the matter energy density 3.5 billion years ago. The horizon distance is 46 billion light years. We provide a table of different quantities (age and comoving distance) as a function of redshift for this model in Table 1. These numbers are for this particular model. They can be taken to represent the real universe with a few % accuracy. The reason for giving them with so many digits in the table is that one can then take differences between the values for different redshifts. Table 2 is for our reference model $h = 0.7$, $\Omega_\Lambda = 0.7$ and Table 3 for the Planck 2018 best-fit model [9].

The distant future in the concordance model has interesting properties because of the accelerating expansion:

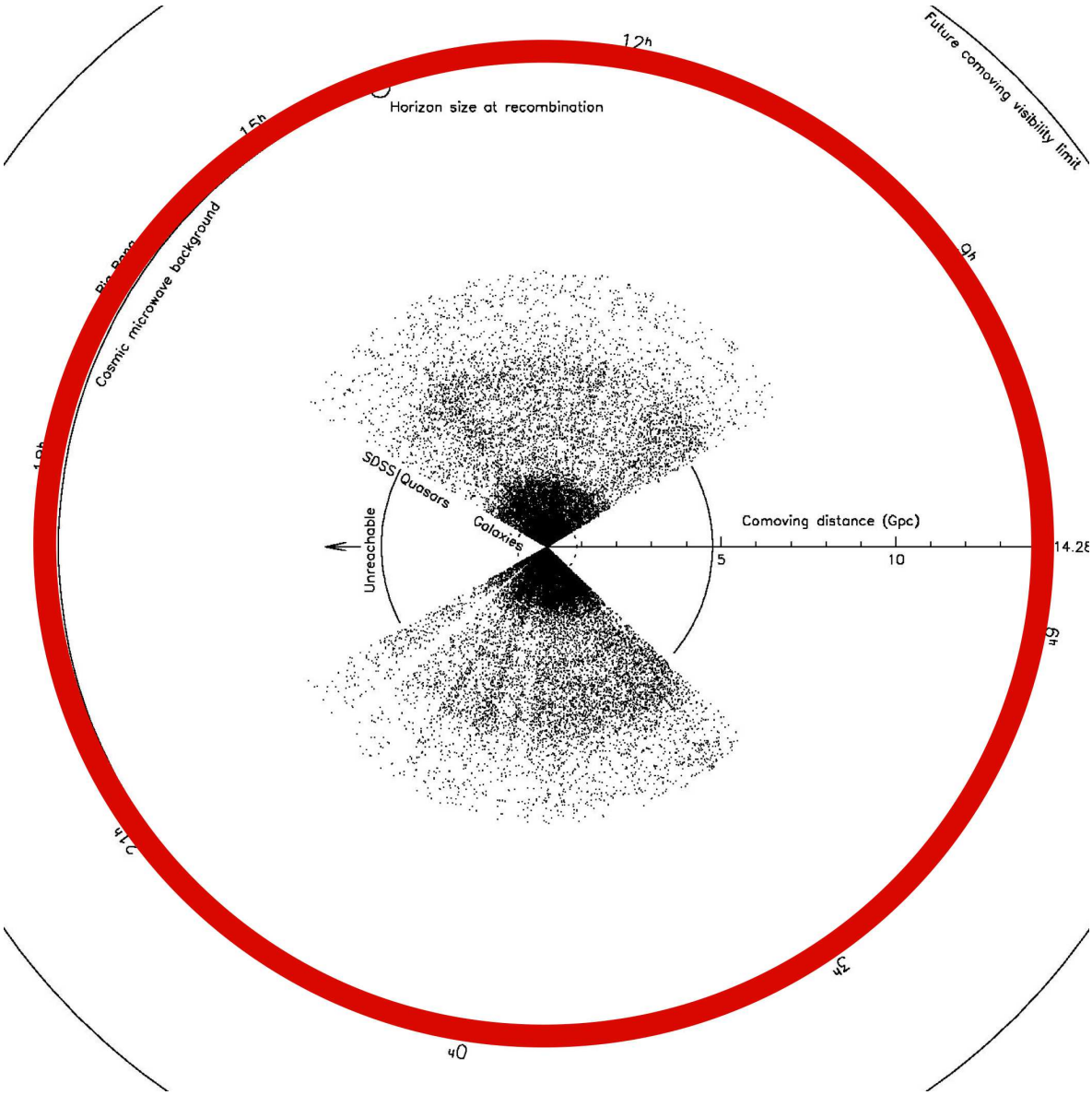


Figure 18: We had this figure already in Chapter 1, but let us look at it again. The figure is in comoving coordinates, so the galaxies do not move in time, except for their peculiar velocity. As time goes on the horizon recedes and we can see further out. The "Future comoving visibility limit" is how far one can eventually see in the very distant future, assuming the "Concordance Model" for the universe (Sec. 3.3). Because of the accelerated expansion of the universe it is not possible to reach the most distant galaxies we see (beyond the circle marked "Unreachable"), even if traveling at (arbitrarily close to) the speed of light. Figure from Gott et al: "Map of the Universe" (2005) [8].

1. Distant galaxies will recede from us faster and faster, with the result that it is not possible to travel from here to the most distant galaxies ("Unreachable" in Fig. 18) we can observe now, even if it there were means to travel with speeds arbitrarily close to the speed of light. Also light rays from here will never reach those galaxies, and similarly, light rays sent from those galaxies today will never reach us.

2. The (comoving) horizon distance $d^c_{\rm hor}$ will approach asymptotically a maximum value (the "Future comoving visibility limit" in Fig. 18). Galaxies beyond that will never become observable from here. We already see a sizable fraction of that part of the universe that will ever become observable from here. Instead, because the redshifts of distant galaxies will keep increasing with time, eventually they will disappear from sight because they will become so faint (they will still stay within the horizon, since their d^c stays constant, and $d^c_{\rm hor}$ does not decrease with time). The relevant time scale here is of course cosmological; we are referring to a future tens of billions of years from today.

Example: Future comoving visibility limit. The comoving distance traveled by a light ray since $t = 0$ until $t = \infty$ is

$$\begin{aligned} d^c &= \int d\chi = \int_0^\infty \frac{dt}{a(t)} = \left(\frac{\Omega_\Lambda}{\Omega_m}\right)^{1/3} \frac{2}{3\sqrt{\Omega_\Lambda} H_0} \int_0^\infty \sinh^{-2/3} x\, dx \\ &= \Omega_\Lambda^{-1/6} \Omega_m^{-1/3} \frac{\Gamma(\frac{1}{6})\Gamma(\frac{1}{3})}{3\sqrt{\pi}} H_0^{-1} = 2.8044\, \Omega_\Lambda^{-1/6} \Omega_m^{-1/3} H_0^{-1}\,. \end{aligned} \tag{150}$$

With $\Omega_m = 0.3$ this gives $d^c = 4.4457 H_0^{-1}$. The integral was done by converting it to the Euler B function

$$B(p,q) \equiv \int_0^1 t^{p-1}(1-t)^{q-1}\, dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \tag{151}$$

by the substitution $t = \tanh^2 x$, which gives

$$\int_0^\infty \sinh^\mu x\, dx = \tfrac{1}{2} \int_0^1 t^{\mu/2-1/2}(1-t)^{-\mu/2-1} = \tfrac{1}{2} B\left(\tfrac{\mu}{2}+\tfrac{1}{2}, -\tfrac{\mu}{2}\right) = \frac{\Gamma(\frac{\mu}{2}+\frac{1}{2})\Gamma(-\frac{\mu}{2})}{2\Gamma(\frac{1}{2})} \tag{152}$$

where $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

References

[1] A. Friedmann, *Über die Krümmung des Raumes*, Zeitschrift für Physik, **10**, 377 (1924).

[2] A. Friedmann, *Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes*, Zeitschrift für Physik, **21**, 326 (1924).

[3] A.G. Riess et al., Astron. J. **116**, 1009 (1998)

[4] S. Perlmutter et al., Astrophys. J. **517**, 565 (1999)

[5] J. Glantz, Science **282**, 2156 (18 Dec 1998)

[6] A.G. Riess et al., Astrophys. J. **607**, 665 (2004), astro-ph/0402512

[7] Planck Collaboration, A&A **571**, A16 (2014), arXiv:1303.5076v2

[8] J. Richard Gott III et al., *A Map of the Universe*, Astrophys. J. **624**, 463 (2005), astro-ph/0310571.

[9] Planck Collaboration, arXiv:1807.06209v1

Age and distance in the Planck 2013 best-fit concordance model

z	t (Gyr)	$t_0 - t$ (H_0^{-1})	d^c (H_0^{-1})	d^c (Mpc)	d^c (Gyr)
0	13.83	0	0	0	0
0.1	12.47	0.0930	0.0976	436	1.421
0.2	11.30	0.1737	0.1902	849	2.770
0.3	10.27	0.2439	0.2779	1241	4.046
0.4	9.38	0.3051	0.3605	1610	5.250
0.5	8.60	0.3589	0.4384	1957	6.384
0.6	7.91	0.4062	0.5117	2284	7.450
0.7	7.30	0.4480	0.5806	2592	8.454
0.8	6.76	0.4850	0.6454	2881	9.397
0.9	6.28	0.5180	0.7064	3154	10.285
1	5.85	0.5475	0.7638	3409	11.121
1.5	4.27	0.6562	1.0061	4492	14.650
2	3.27	0.7243	1.1922	5322	17.359
2.5	2.61	0.7699	1.3397	5981	19.507
3	2.14	0.8020	1.4597	6517	21.255
4	1.54	0.8436	1.6449	7343	23.951
5	1.17	0.8687	1.7823	7957	25.952
6	0.928	0.8853	1.8894	8435	27.511
7	0.760	0.8969	1.9758	8821	28.769
8	0.636	0.9053	2.0474	9141	29.811
9	0.543	0.9117	2.1080	9411	30.694
10	0.471	0.9167	2.1602	9644	31.454
20	0.178	0.9368	2.4553	10962	35.752
100	0.0164	0.9479	2.8741	12831	41.850
1000	0.000430	0.948955	3.1051	13863	45.213
1089	0.000375	0.948959	3.1090	13880	45.271

Table 1: The cosmological parameters assumed here are $\Omega_\Lambda = 0.6830$, $\Omega_m = 0.3169$, $\Omega_r = 9.323 \times 10^{-5}$, $H_0 = 67.15\,\mathrm{km/s/Mpc}$. Credit: Elina Palmgren.

Age and distance in our reference model

z	$t\,(\mathrm{Gyr})$	$t_0 - t\ (H_0^{-1})$	$d^c\,(H_0^{-1})$	$d^c\,(\mathrm{Mpc})$	$d^c\,(\mathrm{Gyr})$	$d^L\,(H_0^{-1})$
0.0	13.463	0.0	0.0	0.0	0.0	0.0
0.1	12.162	0.0932	0.0977	418.5	1.365	0.1075
0.2	11.031	0.1741	0.1907	816.7	2.664	0.2288
0.3	10.045	0.2447	0.2789	1194.4	3.896	0.3625
0.4	9.181	0.3065	0.3623	1551.5	5.061	0.5072
0.5	8.422	0.3609	0.441	1888.6	6.16	0.6615
0.6	7.753	0.4087	0.5152	2206.3	7.196	0.8243
0.7	7.161	0.4511	0.5851	2505.6	8.172	0.9946
0.8	6.636	0.4888	0.6509	2787.6	9.092	1.1716
0.9	6.167	0.5223	0.7129	3053.2	9.959	1.3545
1.0	5.748	0.5523	0.7714	3303.7	10.775	1.5428
1.1	5.372	0.5792	0.8266	3540.0	11.546	1.7358
1.2	5.033	0.6035	0.8787	3763.2	12.274	1.9331
1.3	4.727	0.6254	0.928	3974.3	12.963	2.1344
1.4	4.449	0.6453	0.9746	4174.1	13.615	2.3391
1.5	4.197	0.6633	1.0189	4363.6	14.232	2.5472
1.6	3.967	0.6798	1.0608	4543.3	14.819	2.7582
1.7	3.757	0.6949	1.1007	4714.2	15.376	2.972
1.8	3.564	0.7087	1.1387	4876.8	15.906	3.1884
1.9	3.387	0.7214	1.1749	5031.7	16.411	3.4071
2.0	3.223	0.7331	1.2094	5179.4	16.893	3.6281
2.5	2.57	0.7798	1.3605	5826.9	19.005	4.7619
3.0	2.109	0.8128	1.4838	6354.8	20.727	5.9353
3.5	1.771	0.837	1.5866	6795.1	22.163	7.1398
4.0	1.513	0.8555	1.674	7169.1	23.383	8.3698
4.5	1.312	0.8698	1.7493	7491.8	24.435	9.6211
5.0	1.152	0.8813	1.8151	7773.7	25.355	10.8908
6.0	0.915	0.8983	1.9251	8244.9	26.892	13.476
7.0	0.749	0.9102	2.014	8625.3	28.132	16.1116
8.0	0.627	0.9189	2.0876	8940.5	29.161	18.7881
9.0	0.535	0.9255	2.1499	9207.3	30.031	21.4987
10.0	0.464	0.9306	2.2035	9436.9	30.78	24.2383
20.0	0.175	0.9513	2.5069	10736.4	35.018	52.6446
100.0	0.0162	0.9626	2.9376	12581.1	41.035	296.6988
1000.0	0.000425	0.963767	3.1753	13599.2	44.356	3178.5241
1090.0	0.000368	0.963771	3.1796	13617.5	44.415	3468.9642

Table 2: The cosmological parameters assumed here are $H_0 = 70.0\,\mathrm{km/s/Mpc}$, $\Omega_\Lambda = 0.7$, $\Omega_r = 4.18 \times 10^{-5} h^{-2} = 8.531 \times 10^{-5}$, $\Omega_m = 1 - \Omega_\Lambda - \Omega_r$.

Age and distance in in the Planck 2018 best-fit concordance model

z	$t\,(\mathrm{Gyr})$	$t_0 - t\ (H_0^{-1})$	$d^c\,(H_0^{-1})$	$d^c\,(\mathrm{Mpc})$	$d^c\,(\mathrm{Gyr})$	$d^L\,(H_0^{-1})$
0.0	13.797	0.0	0.0	0.0	0.0	0.0
0.1	12.446	0.0931	0.0976	434.6	1.417	0.1074
0.2	11.274	0.1737	0.1903	847.2	2.763	0.2283
0.3	10.255	0.2439	0.2779	1237.7	4.037	0.3613
0.4	9.364	0.3052	0.3607	1606.1	5.238	0.5049
0.5	8.583	0.359	0.4386	1953.1	6.37	0.6579
0.6	7.895	0.4064	0.5119	2279.6	7.435	0.819
0.7	7.288	0.4482	0.5809	2586.7	8.437	0.9875
0.8	6.749	0.4853	0.6457	2875.6	9.379	1.1623
0.9	6.27	0.5183	0.7068	3147.4	10.266	1.3428
1.0	5.841	0.5478	0.7642	3403.4	11.101	1.5285
1.1	5.457	0.5742	0.8184	3644.7	11.888	1.7187
1.2	5.111	0.598	0.8696	3872.5	12.631	1.9131
1.3	4.799	0.6195	0.9179	4087.7	13.332	2.1112
1.4	4.517	0.639	0.9636	4291.3	13.997	2.3127
1.5	4.26	0.6566	1.0069	4484.2	14.626	2.5174
1.6	4.026	0.6728	1.048	4667.2	15.223	2.7249
1.7	3.812	0.6875	1.0871	4841.0	15.79	2.9351
1.8	3.615	0.701	1.1242	5006.3	16.329	3.1477
1.9	3.435	0.7134	1.1596	5163.8	16.842	3.3627
2.0	3.269	0.7248	1.1933	5313.9	17.332	3.5798
2.5	2.606	0.7705	1.3409	5971.6	19.477	4.6933
3.0	2.138	0.8027	1.4613	6507.3	21.225	5.845
3.5	1.795	0.8264	1.5615	6953.9	22.681	7.027
4.0	1.534	0.8443	1.6467	7333.2	23.918	8.2336
4.5	1.33	0.8584	1.7202	7660.4	24.985	9.461
5.0	1.168	0.8695	1.7844	7946.2	25.918	10.7062
6.0	0.927	0.8861	1.8916	8423.8	27.475	13.2413
7.0	0.758	0.8977	1.9782	8809.3	28.733	15.8253
8.0	0.636	0.9062	2.0499	9128.8	29.775	18.4492
9.0	0.542	0.9126	2.1106	9399.1	30.657	21.1063
10.0	0.47	0.9176	2.1629	9631.8	31.415	23.7916
20.0	0.178	0.9377	2.4585	10948.5	35.71	51.6294
100.0	0.0164	0.9488	2.8782	12817.4	41.806	290.6998
1000.0	0.000429	0.949888	3.1097	13848.0	45.167	3112.7597
1090.0	0.000371	0.949892	3.1138	13866.5	45.227	3397.1543

Table 3: The cosmological parameters assumed here are $H_0 = 67.32\,\mathrm{km/s/Mpc}$, $\Omega_\Lambda = 0.6841$, $\Omega_r = 4.18 \times 10^{-5} h^{-2} = 9.223 \times 10^{-5}$, $\Omega_m = 1 - \Omega_\Lambda - \Omega_r$.

4 Thermal history of the Early Universe

4.1 Relativistic thermodynamics

As we look out in space we can see the history of the universe unfolding in front of our telescopes. However, at redshift $z = 1090$ our line of sight hits the *last scattering surface*, from which the cosmic microwave background (CMB) radiation originates. This corresponds to $t = 370\,000$ years. Before that the universe was not transparent, so we cannot see further back in time, into the *early universe*. As explained in Sec. 3, we can ignore curvature and vacuum/dark energy in the early universe and concern ourselves only with radiation and matter. The isotropy of the CMB shows that matter was distributed homogeneously in the early universe, and the spectrum of the CMB shows that this matter, the "primordial soup" of particles, was in thermodynamic equilibrium. Therefore we can use thermodynamics to calculate the history of the early universe. As we shall see, this calculation leads to predictions (especially the BBN, big bang nucleosynthesis) testable by observation. We shall now discuss the thermodynamics of the primordial soup.

From elementary quantum mechanics we are familiar with the "particle in a box". Let us consider a cubic box, whose edge is L (volume $V = L^3$), with periodic boundary conditions. Solving the Schrödinger equation gives us the energy and momentum eigenstates, where the possible momentum values are

$$\vec{p} = \frac{h}{L}(n_1\hat{x} + n_2\hat{y} + n_3\hat{z}) \qquad (n_i = 0, \pm 1, \pm 2, \ldots)\,, \tag{1}$$

where h is the Planck constant. (The wave function will have an integer number of wavelengths in each of the three directions.) The state density in momentum space (number of states / $\Delta p_x \Delta p_y \Delta p_z$) is thus

$$\frac{L^3}{h^3} = \frac{V}{h^3}, \tag{2}$$

and the state density in the 6-dimensional phase space $\{(\vec{x}, \vec{p})\}$ is $1/h^3$. If the particle has g internal degrees of freedom (e.g., spin),

$$\text{density of states} = \frac{g}{h^3} = \frac{g}{(2\pi)^3} \qquad \left(\hbar \equiv \frac{h}{2\pi} \equiv 1\right). \tag{3}$$

This result is true even for relativistic momenta. The state density in phase space is independent of the volume V, so we can apply it for arbitrarily large systems (e.g., the universe).

For much of the early universe, we can ignore the interaction energies between the particles. Then the particle energy is (according to special relativity)

$$E(\vec{p}) = \sqrt{p^2 + m^2}\,, \tag{4}$$

where $p \equiv |\vec{p}|$ (not pressure!), and the states available for the particles are the free particle states discussed above.

Particles fall into two classes, *fermions* and *bosons*. Fermions obey the Pauli exclusion principle: no two fermions can be in the same state.

In thermodynamic equilibrium the *distribution function*, or the expectation value f of the occupation number of a state, depends only on the energy of the state. According to statistical physics, it is

$$f(\vec{p}) = \frac{1}{e^{(E-\mu)/T} \pm 1} \tag{5}$$

where $+$ is for fermions and $-$ is for bosons. (For fermions, where $f \leq 1$, f gives the probability that a state is occupied.) This equilibrium distribution has two parameters, the *temperature* T,

and the *chemical potential* μ. The temperature is related to the energy density in the system and the chemical potential is related to the number density n of particles in the system. Note that, since we are using the relativistic formula for the particle energy E, which includes the mass m, it is also "included" in the chemical potential μ. Thus in the nonrelativistic limit, both E and μ differ from the corresponding quantities of nonrelativistic statistical physics by m, so that $E - \mu$ and the distribution functions remain the same.

If there is no conserved particle number in the system (e.g., a photon gas), then $\mu = 0$ in equilibrium.

The particle density in phase space is the density of states times their occupation number,

$$\frac{g}{(2\pi)^3} f(\vec{p}). \tag{6}$$

We get the particle density in (ordinary 3D) space by integrating over the momentum space. Thus we find the following quantities:

$$\text{number density} \qquad n = \frac{g}{(2\pi)^3} \int f(\vec{p}) d^3p \tag{7}$$

$$\text{energy density} \qquad \rho = \frac{g}{(2\pi)^3} \int E(\vec{p}) f(\vec{p}) d^3p \tag{8}$$

$$\text{pressure} \qquad p = \frac{g}{(2\pi)^3} \int \frac{|\vec{p}|^2}{3E} f(\vec{p}) d^3p \,. \tag{9}$$

Different particle species i have different masses m_i; so the preceding is applied separately to each particle species. If particle species i has the above distribution for some μ_i and T_i, we say the species is in *kinetic equilibrium*. If the system is in *thermal equilibrium*, all species have the same temperature, $T_i = T$. If the system is in *chemical equilibrium* ("chemistry" here refers to reactions where particles change into other species), the chemical potentials of different particle species are related according to the reaction formulas. For example, if we have a reaction

$$i + j \quad \leftrightarrow \quad k + l \,, \tag{10}$$

then

$$\mu_i + \mu_j = \mu_k + \mu_l \,. \tag{11}$$

Thus all chemical potentials can be expressed in terms of the chemical potentials of conserved quantities, e.g., the baryon number chemical potential, μ_B. There are thus as many independent chemical potentials, as there are independent conserved particle numbers. For example, if the chemical potential of particle species i is μ_i, then the chemical potential of the corresponding antiparticle is $-\mu_i$. We can also have a situation that some reactions are in chemical equilibrium but others are not.

Thermodynamic equilibrium refers to having all these equilibria, but I will also use the term more loosely to refer to some subset of them.

As the universe expands, T and μ change, so that energy continuity and particle number conservation are satisfied. In principle, an expanding universe is not in equilibrium. The expansion is however so slow, that the particle soup usually has time to settle close to local equilibrium. (And since the universe is homogeneous, the local values of thermodynamic quantities are also global values).

From the remaining numbers of fermions (electrons and nucleons) in the present universe, we can conclude that in the early universe we had $|\mu| \ll T$ when $T \gg m$. (We don't know the chemical potential of neutrinos, but it is usually assumed to be small too). If the temperature is much greater than the mass of a particle, $T \gg m$, the *ultrarelativistic limit*, we can approximate $E = \sqrt{p^2 + m^2} \approx p$.

For $|\mu| \ll T$ and $m \ll T$, we approximate $\mu = 0$ and $m = 0$ to get the following formulae

$$n = \frac{g}{(2\pi)^3}\int_0^\infty \frac{4\pi p^2 dp}{e^{p/T} \pm 1} = \begin{cases} \frac{3}{4\pi^2}\zeta(3) g T^3 & \text{fermions} \\ \frac{1}{\pi^2}\zeta(3) g T^3 & \text{bosons} \end{cases} \tag{12}$$

$$\rho = \frac{g}{(2\pi)^3}\int_0^\infty \frac{4\pi p^3 dp}{e^{p/T} \pm 1} = \begin{cases} \frac{7}{8}\frac{\pi^2}{30} g T^4 & \text{fermions} \\ \frac{\pi^2}{30} g T^4 & \text{bosons} \end{cases} \tag{13}$$

$$p = \frac{g}{(2\pi)^3}\int_0^\infty \frac{\frac{4}{3}\pi p^3 dp}{e^{p/T} \pm 1} = \frac{1}{3}\rho \approx \begin{cases} 1.0505 nT & \text{fermions} \\ 0.9004 nT & \text{bosons}\,. \end{cases} \tag{14}$$

For the average particle energy we get

$$\langle E \rangle = \frac{\rho}{n} = \begin{cases} \frac{7\pi^4}{180\zeta(3)} T \approx 3.151 T & \text{fermions} \\ \frac{\pi^4}{30\zeta(3)} T \approx 2.701 T & \text{bosons}\,. \end{cases} \tag{15}$$

In the above, ζ is the Riemann zeta function, and $\zeta(3) \equiv \sum_{n=1}^\infty (1/n^3) = 1.20206$.

If the chemical potential $\mu = 0$, there are equal numbers of particles and antiparticles. If $\mu \neq 0$, we find for fermions in the ultrarelativistic limit $T \gg m$ (i.e., for $m = 0$, but $\mu \neq 0$) the "net particle number"

$$\begin{aligned} n - \bar{n} &= \frac{g}{(2\pi)^3}\int_0^\infty dp\, 4\pi p^2 \left(\frac{1}{e^{(p-\mu)/T} + 1} - \frac{1}{e^{(p+\mu)/T} + 1} \right) \\ &= \frac{g T^3}{6\pi^2}\left(\pi^2 \left(\frac{\mu}{T}\right) + \left(\frac{\mu}{T}\right)^3 \right) \end{aligned} \tag{16}$$

and the total energy density

$$\begin{aligned} \rho + \bar{\rho} &= \frac{g}{(2\pi)^3}\int_0^\infty dp\, 4\pi p^3 \left(\frac{1}{e^{(p-\mu)/T} + 1} + \frac{1}{e^{(p+\mu)/T} + 1} \right) \\ &= \frac{7}{8} g \frac{\pi^2}{15} T^4 \left(1 + \frac{30}{7\pi^2}\left(\frac{\mu}{T}\right)^2 + \frac{15}{7\pi^4}\left(\frac{\mu}{T}\right)^4 \right). \end{aligned} \tag{17}$$

Note that the last forms in Eqs. (16) and (17) are exact, not just truncated series. (The difference $n - \bar{n}$ and the sum $\rho + \bar{\rho}$ lead to a nice cancellation between the two integrals. We don't get such an elementary form for the individual n, $\bar{n}$, ρ, $\bar{\rho}$, or the sum $n + \bar{n}$ and the difference $\rho - \bar{\rho}$ when $\mu \neq 0$.)

In the nonrelativistic limit, $T \ll m$ and $T \ll m - \mu$, the typical kinetic energies are much below the mass m, so that we can approximate $E = m + p^2/2m$. The second condition, $T \ll m - \mu$, leads to occupation numbers $\ll 1$, a *dilute* system. This second condition is usually satisfied in cosmology when the first one is. (It is violated in systems of high density, like white dwarf stars and neutrons stars.) We can then approximate

$$e^{(E-\mu)/T} \pm 1 \approx e^{(E-\mu)/T}, \tag{18}$$

so that the boson and fermion expressions become equal,[1] and we get (exercise)

$$n = g\left(\frac{mT}{2\pi}\right)^{3/2} e^{-\frac{m-\mu}{T}} \tag{19}$$

$$\rho = n\left(m + \frac{3T}{2}\right) \tag{20}$$

$$p = nT \ll \rho \tag{21}$$

$$\langle E \rangle = m + \frac{3T}{2} \tag{22}$$

$$n - \bar{n} = 2g\left(\frac{mT}{2\pi}\right)^{\frac{3}{2}} e^{-\frac{m}{T}} \sinh\frac{\mu}{T}. \tag{23}$$

In the general case, where neither $T \ll m$, nor $T \gg m$, the integrals don't give elementary functions, but $n(T)$, $\rho(T)$, etc. need to be calculated numerically for the region $T \sim m$.[2]

By comparing the ultrarelativistic ($T \gg m$) and nonrelativistic ($T \ll m$) limits we see that the number density, energy density, and pressure of a particle species falls exponentially as the temperature falls below the mass of the particle. What happens is that the particles and antiparticles annihilate each other. (Other reactions may also be involved, and if these particles are unstable, also their decay contributes to their disappearance.) At higher temperatures these annihilation reactions are also constantly taking place, but they are balanced by particle-antiparticle pair production. At lower temperatures the thermal particle energies are no more sufficient for pair production. This *particle-antiparticle annihilation* takes place mainly (about 80%) during the temperature interval $T = m \to \frac{1}{6}m$. See Fig. 1. It is thus not an instantaneous event, but takes several Hubble times.

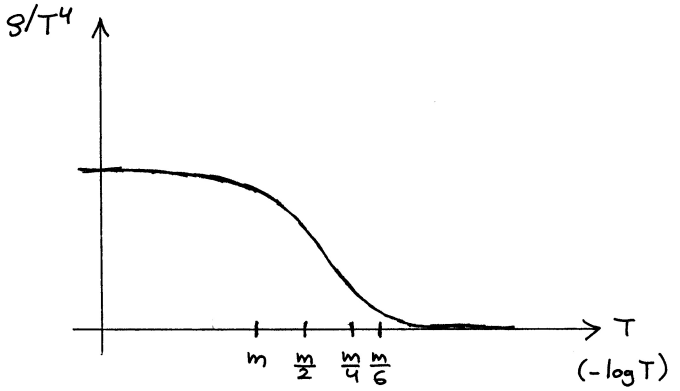


Figure 1: The fall of energy density of a particle species, with mass m, as a function of temperature (decreasing to the right).

4.2 Primordial soup

We shall now apply the thermodynamics discussed in the previous section to the evolution of the early universe.

The primordial soup initially consists of all the different species of elementary particles. Their masses range from the heaviest known elementary particle, the top quark ($m = 173$ GeV) down to the lightest particles, the electron ($m = 511$ keV), the neutrinos ($m = ?$) and the

[1] This approximation leads to what is called Maxwell–Boltzmann statistics; whereas the previous exact formulas give Fermi–Dirac (for fermions) and Bose–Einstein (for bosons) statistics.

[2] If we use Maxwell–Boltzmann statistics, i.e., we drop the term ± 1, the integrals give modified Bessel functions, e.g., $K_2(m/T)$, and the error is often less than 10%.

photon ($m = 0$). In addition to the particles of the standard model of particle physics (given in Table 1), there are other, so far undiscovered, species of particles, at least those that make up the CDM. As the temperature falls, the various particle species *become nonrelativistic and annihilate* at different times.

Another central theme is *decoupling*: as the number densities and particle energies fall with the expansion, some reaction rates become too low to keep up with the changing equilibrium and therefore some quantities are "frozen" at their pre-decoupling values. We will encounter neutrino and photon decoupling later in this chapter; decoupling is also important in BBN (Chapter 5) and for dark matter (Chapter 6).

Table 1: The particles in the standard model of particle physics
Particle Data Group, 2018

Quarks	t	$173.0 \pm 0.4\,\mathrm{GeV}$	$\bar{t}$	spin=$\frac{1}{2}$	$g = 2 \cdot 3 = 6$	
	b	4.15–4.22 GeV	$\bar{b}$	3 colors		
	c	$1.27 \pm 0.03\,\mathrm{GeV}$	$\bar{c}$			
	s	92–104 MeV	$\bar{s}$			
	d	4.4–5.2 MeV	$\bar{d}$			
	u	1.8–2.7 MeV	$\bar{u}$			
						72
Gluons	8 massless bosons			spin=1	$g = 2$	16
Leptons	τ^-	$1776.86 \pm 0.12\,\mathrm{MeV}$	τ^+	spin=$\frac{1}{2}$	$g = 2$	
	μ^-	105.658 MeV	μ^+			
	e^-	510.999 keV	e^+			
						12
	ν_τ	$< 2\,\mathrm{eV}$	$\bar{\nu}_\tau$	spin=$\frac{1}{2}$	$g = 1$	
	ν_μ	$< 2\,\mathrm{eV}$	$\bar{\nu}_\mu$			
	ν_e	$< 2\,\mathrm{eV}$	$\bar{\nu}_e$			
						6
Electroweak	W^+	$80.379 \pm 0.012\,\mathrm{GeV}$	W^-	spin=1	$g = 3$	
gauge bosons	Z^0	$91.1876 \pm 0.0021\,\mathrm{GeV}$				
	γ	0 $(< 1 \times 10^{-18}\,\mathrm{eV})$			$g = 2$	
						11
Higgs boson (SM)	H^0	$125.18 \pm 0.16\,\mathrm{GeV}$		spin=0	$g = 1$	1
					$g_f = 72 + 12 + 6 = 90$	
					$g_b = 16 + 11 + 1 = 28$	

The mass limits for neutrinos come from a direct laboratory upper limit for ν_e and evidence from neutrino oscillations that the differences in neutrino masses are much smaller. We can use cosmology to put tighter limits to neutrino masses. Neutrinos are special in that the antineutrino is just the other spin state of the neutrino. Therefore we put $g = 1$ for their internal degrees of freedom when we count antineutrinos separately.

According to the Friedmann equation the expansion of the universe is governed by the total energy density

$$\rho(T) = \sum \rho_i(T)\,,$$

where i runs over the different particle species. Since the energy density of relativistic species is much greater than that of nonrelativistic species, it suffices to include the relativistic species only. (This is true in the early universe, during the radiation-dominated era, but not at later times. Eventually the rest masses of the particles left over from annihilation begin to dominate and we enter the matter-dominated era.) Thus we have

$$\rho(T) = \frac{\pi^2}{30} g_*(T) T^4, \tag{24}$$

where

$$g_*(T) = g_b(T) + \frac{7}{8} g_f(T),$$

and $g_b = \sum_i g_i$ over relativistic bosons and $g_f = \sum_i g_i$ over relativistic fermions. These results assume thermal equilibrium. For pressure we have $p(T) \approx \frac{1}{3}\rho(T)$.

The above is a simplification of the true situation: Since the annihilation takes a long time, often the annihilation of some particle species is going on, and the contribution of this species disappears gradually. Using the exact formula for ρ we define *the effective number of degrees of freedom* $g_*(T)$ by

$$g_*(T) \equiv \frac{30}{\pi^2} \frac{\rho}{T^4}\,. \tag{25}$$

We can also define

$$g_{*p}(T) \equiv \frac{90}{\pi^2} \frac{p}{T^4} \approx g_*(T)\,. \tag{26}$$

These can then be calculated numerically (see Figure 1).

We see that when there are no annihilations taking place, $g_{*p} = g_* = \text{ const} \Rightarrow p = \frac{1}{3}\rho \Rightarrow \rho \propto a^{-4}$ and $\rho \propto T^4$, so that $T \propto a^{-1}$. Later in this chapter we shall calculate the $T(a)$ relation more exactly (including the effects of annihilations).

For $T > m_t = 173$ GeV, all known particles are relativistic. Adding up their internal degrees of freedom we get

g_b	=	28	gluons 8×2, photons 2, $W^\pm$ and Z^0 3×3, and Higgs 1
g_f	=	90	quarks 12×6, charged leptons 6×2, neutrinos 3×2
g_*	=	106.75.	

The electroweak (EW) transition[3] took place close to this time ($T_c \sim 100$ GeV). It appears that g_* was the same before and after this transition. Going to earlier times and higher temperatures, we expect g_* to get larger than 106.75 as new physics (new unknown particle species) comes to play.[4]

[3] This is usually called the electroweak *phase transition*, but the exact nature of the transition is not known. Technically it may be a cross-over rather than a phase transition, meaning that it occurs over a temperature range rather than at a certain critical temperature T_c.

[4] A popular form of such new physics is *supersymmetry*, which provides supersymmetric partners, whose spin differ by $\frac{1}{2}$, for the known particle species, so that fermions have supersymmetric boson partners and bosons have supersymmetric fermion partners. Since these partners have not been so far observed, supersymmetry must be *broken*, allowing these partners to have much higher masses. In the minimal supersymmetric standard model (MSSM) the new internal degrees of freedom are as follows: Spin-0 bosons (scalars): sleptons $9 \cdot 2 = 18$, squarks $6 \cdot 2 \cdot 2 \cdot 3 = 72$ (although there is only one spin degree instead of 2, there is another degree of freedom, so that we get the same 18+72 as for leptons and quarks), and a new complex Higgs doublet $2 \cdot 2 = 4$. Spin-$\frac{1}{2}$ fermions: neutralinos $4 \cdot 2 = 8$, charginos $2 \cdot 2 \cdot 2$ (two charge degrees and two spin degrees), and gluinos $8 \cdot 2 = 16$. This gives $g_* = 106.75 + 94 + \frac{7}{8} \cdot 32 = 228.75$. Other supersymmetric models have somewhat more degrees of freedom but some of the new degrees of freedom may be very heavy ($10^{10} \ldots 10^{16}$ GeV).[1]

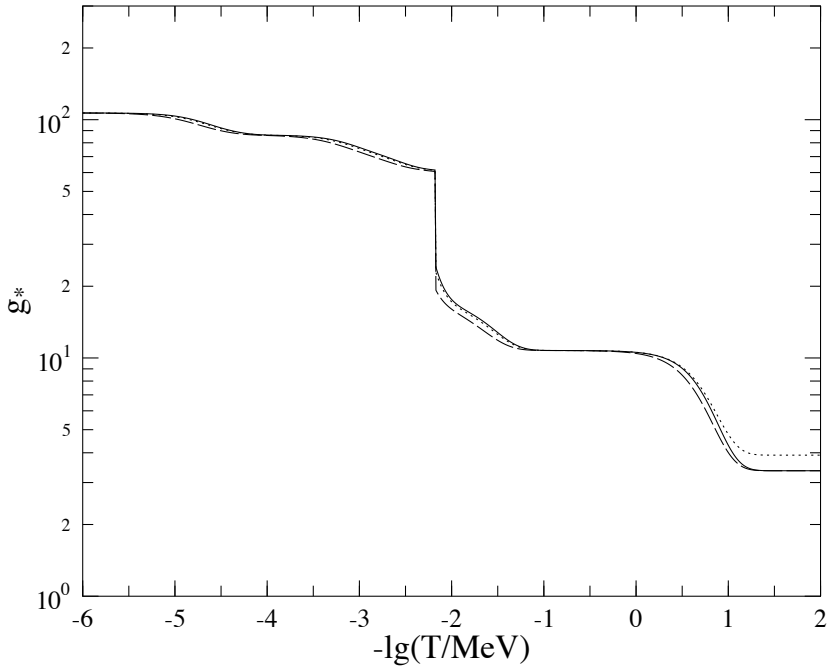


Figure 2: The functions $g_*(T)$ (solid), $g_{*p}(T)$ (dashed), and $g_{*s}(T)$ (dotted) calculated for the standard model particle content.

Let us now follow the history of the universe starting at the time when the EW transition has already happened. We have $T \sim 100$ GeV, $t \sim 20$ ps, and the t quark annihilation is on the way. The Higgs boson and the gauge bosons $W^{\pm}$, Z^0 annihilate next. At $T \sim 10$ GeV, we have $g_* = 86.25$. Next the b and c quarks annihilate and then the τ meson, so that $g_* = 61.75$.

4.3 QCD transition

Before s quark annihilation would take place, something else happens: the *QCD transition* (also called the quark–hadron transition). This takes place at $T \sim 150$ MeV, $t \sim 20$ μs. The temperature and thus the quark energies have fallen so that the quarks lose their so called *asymptotic freedom*, which they have at high energies. The interactions between quarks and gluons (the strong nuclear force, or the color force) become important (so that the formulae for the energy density in Sec. 4.1 no longer apply) and soon a phase transition takes place. There are no more free quarks and gluons; the *quark-gluon plasma* has become a *hadron gas*. The quarks and gluons have formed bound three-quark systems, called *baryons*, and quark-antiquark pairs, called *mesons*. The lightest baryons are the nucleons: the proton and the neutron. The lightest mesons are the pions: $\pi^{\pm}$, π^0. Baryons are fermions, mesons are bosons.

There are very many different species of baryons and mesons, but all except pions are non-relativistic below the QCD transition temperature. Thus the only particle species left in large numbers are the pions, muons, electrons, neutrinos, and the photons. For pions, $g = 3$, so now $g_* = 17.25$.

Table 2: History of $g_*(T)$

$T \sim 200$ GeV	all present	106.75	
$T \sim 100$ GeV	EW transition	(no effect)	
$T < 170$ GeV	top annihilation	96.25	
$T < 80$ GeV	$W^\pm$, Z^0, H^0	86.25	
$T < 4$ GeV	bottom	75.75	
$T < 1$ GeV	charm, τ^-	61.75	
$T \sim 150$ MeV	QCD transition	17.25	(u,d,s,g$\to \pi^{\pm,0}$, $47.5 \to 3$)
$T < 100$ MeV	$\pi^\pm$, π^0, μ^-	10.75	$e^\pm$, ν, $\bar{\nu}$, γ left
$T < 500$ keV	e^- annihilation	(7.25)	$2 + 5.25(4/11)^{4/3} = 3.36$

This table gives what value $g_*(T)$ would have after the annihilation of a particle species is over assuming the annihilation of the next species had not begun yet. In reality they overlap in many cases. The temperature value at the left is the approximate mass of the particle in question and indicates roughly when annihilation begins. The temperature is much smaller when the annihilation is over. Therefore top annihilation is placed after the EW transition. The top quark receives its mass in the EW transition, so annihilation only begins after the transition.

4.4 Neutrino decoupling and electron-positron annihilation

Soon after the QCD phase transition the pions and muons annihilate and for $T = 20$ MeV $\to 1$ MeV, $g_* = 10.75$. Next the electrons annihilate, but to discuss the e^+e^--annihilation we need more physics.

So far we have assumed that all particle species have the same temperature, i.e., the interactions among the particles are able to keep them in thermal equilibrium. Neutrinos, however, feel the weak interaction only. The weak interaction is actually not so weak when particle energies are close to the masses of the $W^\pm$ and Z^0 bosons, which mediate the weak interaction. But as the temperature falls, the weak interaction becomes rapidly weaker and weaker. Finally, close to $T \sim 1$ MeV, the neutrinos *decouple*, after which they move practically freely without interactions.

The momentum of a freely moving neutrino redshifts as the universe expands,

$$p(t_2) = (a_1/a_2)p(t_1)\,. \tag{27}$$

From this follows that neutrinos stay in kinetic equilibrium. This is true in general for ultrarelativistic ($m \ll T \Rightarrow p = E$) noninteracting particles. Let us show this:

At time t_1 a phase space element $d^3p_1dV_1$ contains

$$dN = \frac{g}{(2\pi)^3} f(\vec{p}_1) d^3p_1 dV_1 \tag{28}$$

particles, where

$$f(\vec{p}_1) = \frac{1}{e^{(p_1-\mu_1)/T_1} \pm 1}$$

is the distribution function at time t_1. At time t_2 these same dN particles are in a phase space element $d^3p_2dV_2$. Now how is the distribution function at t_2, given by

$$\frac{g}{(2\pi)^3} f(\vec{p}_2) = \frac{dN}{d^3p_2dV_2}\,,$$

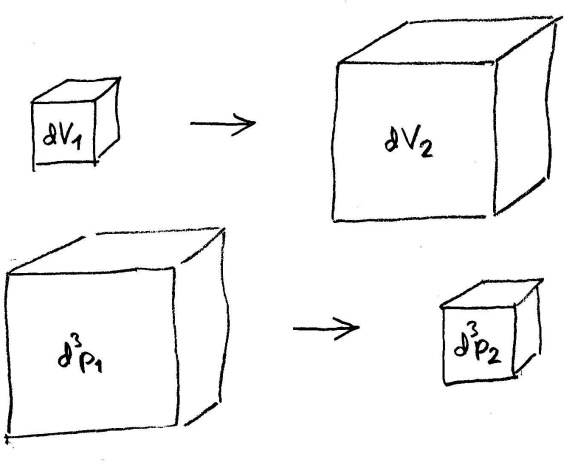


Figure 3: The expansion of the universe increases the volume element dV and decreases the momentum space element d^3p so that the phase space element $d^3p dV$ stays constant.

related to $f(\vec{p}_1)$? Since $d^3p_2 = (a_1/a_2)^3 d^3p_1$ and $dV_2 = (a_2/a_1)^3 dV_1$, we have

$$\begin{aligned} dN &= \frac{g}{(2\pi)^3}\frac{d^3p_1\, dV_1}{e^{(p_1-\mu_1)/T_1}\pm 1} && (dN \text{ evaluated at } t_1) \\ &= \frac{g}{(2\pi)^3}\frac{(\frac{a_2}{a_1})^3 d^3p_2\,(\frac{a_1}{a_2})^3 dV_2}{e^{(\frac{a_2}{a_1}p_2-\mu_1)/T_1}\pm 1} && \text{(rewritten in terms of } p_2,\, dp_2, \text{ and } dV_2) \\ &= \frac{g}{(2\pi)^3}\frac{d^3p_2\, dV_2}{e^{(p_2-\frac{a_1}{a_2}\mu_1)/\frac{a_1}{a_2}T_1}\pm 1} \\ &= \frac{g}{(2\pi)^3}\frac{d^3p_2\, dV_2}{e^{(p_2-\mu_2)/T_2}\pm 1} && (\text{defining } \mu_2 \text{ and } T_2)\,, \end{aligned} \tag{29}$$

where $\mu_2 \equiv (a_1/a_2)\mu_1$ and $T_2 \equiv (a_1/a_2)T_1$. Thus the particles keep the shape of a thermal distribution; the temperature and the chemical potential just redshift $\propto a^{-1}$. (**Exercise:** For nonrelativistic particles, $m \gg T \Rightarrow E = m + p^2/2m$, there is a corresponding, but different result. Derive this.)

Thus for as long as $T \propto a^{-1}$ for the particle soup, the neutrino distribution evolves exactly as if it were in thermal equilibrium with the soup, i.e., $T_\nu = T$. However, annihilations will cause a deviation from $T \propto a^{-1}$. The next annihilation event is the electron-positron annihilation.

The easiest way to obtain the relation between the temperature T and the scale factor a is to use *entropy conservation*.

From the fundamental equation of thermodynamics,

$$E = TS - pV + \sum \mu_i N_i$$

we have

$$s = \frac{\rho + p - \sum \mu_i n_i}{T}, \tag{30}$$

for the entropy density $s \equiv S/V$. Since $|\mu_i| \ll T$, and the relativistic species dominate, we approximate

$$s = \frac{\rho + p}{T} = \begin{cases} \frac{7\pi^2}{180} g T^3 & \text{fermions} \\ \frac{2\pi^2}{45} g T^3 & \text{bosons}\,. \end{cases} \tag{31}$$

Adding up all the relativistic species and allowing now for the possibility that some species may have a kinetic temperature T_i, which differs from the temperature T of those species which remain in thermal equilibrium, we get

$$\begin{aligned} \rho(T) &= \frac{\pi^2}{30} g_*(T) T^4 \\ s(T) &= \frac{2\pi^2}{45} g_{*s}(T) T^3 , \end{aligned} \tag{32}$$

where now

$$\begin{aligned} g_*(T) &= \sum_{\text{bos}} g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8} \sum_{\text{fer}} g_i \left(\frac{T_i}{T}\right)^4 \\ g_{*s}(T) &= \sum_{\text{bos}} g_i \left(\frac{T_i}{T}\right)^3 + \frac{7}{8} \sum_{\text{fer}} g_i \left(\frac{T_i}{T}\right)^3 , \end{aligned} \tag{33}$$

and the sums are over all relativistic species of bosons and fermions.

If some species are "semirelativistic", i.e., $m = \mathcal{O}(T)$, $\rho(T)$ and $s(T)$ are to be calculated from the integral formulae in Sec. 4.1, and Eq. (32) *defines* $g_*(T)$ and $g_{*s}(T)$.

For as long as all species have the same temperature and $p \approx \frac{1}{3}\rho$, we have

$$g_{*s}(T) \approx g_*(T). \tag{34}$$

The electron annihilation, however, forces us to make a distinction between $g_*(T)$ and $g_{*s}(T)$.

According to the second law of thermodynamics the total entropy of the universe never decreases; it either stays constant or increases. An increase in entropy is always related to a deviation from thermodynamic equilibrium. It turns out that any entropy production in the various known processes in the universe is totally insignificant compared to the total entropy of the universe[5], which is huge, and dominated by the relativistic species. Thus it is an excellent approximation to treat the expansion of the universe as *adiabatic*, so that the total entropy stays constant, i.e.,

$$d(sa^3) = 0. \tag{35}$$

This now gives us the relation between a and T,

$$\boxed{g_{*s}(T) T^3 a^3 = \text{const.}} \tag{36}$$

We shall have much use for this formula.

In the electron annihilation g_{*s} changes from

$$\begin{aligned} g_{*s} \;=\; g_* \;&=\; 2 + 3.5 + 5.25 \;=\; 10.75 \\ &\quad\; \gamma \quad e^{\pm} \qquad \nu \end{aligned} \tag{37}$$

to

$$g_{*s} = 2 + 5.25 \left(\frac{T_\nu}{T}\right)^3 , \tag{38}$$

where

$$T_\nu^3 a^3 = \text{const} = T^3 a^3 (\text{before annihilation}). \tag{39}$$

[5]There may be exceptions to this in the very early universe, most notably *inflation*, where essentially all the entropy of the universe supposedly was produced.

(since the neutrinos have decoupled, T_ν redshifts $T_\nu \propto a^{-1}$). As the number of relativistic degrees of freedom is reduced, energy density and entropy are transferred from electrons and positrons to photons, but not to neutrinos, in the annihilation reactions

$$e^+ + e^- \quad \to \quad \gamma + \gamma.$$

The photons are thus heated (the photon temperature does not fall as much) relative to neutrinos.

Dividing Eq. (36) with Eq. (39) we get that

$$g_{*s}(T) \left(\frac{T}{T_\nu}\right)^3 = \text{const}$$

or (Eqs. (37) and (38))

$$10.75 = 2\left(\frac{T}{T_\nu}\right)^3 + 5.25 \qquad \text{(before = after)}$$

from which we solve the neutrino temperature after e^+e^--annihilation,[6]

$$\begin{aligned} T_\nu &= \left(\frac{4}{11}\right)^{\frac{1}{3}} T \;=\; 0.714\,T \\ g_{*s}(T) &= 2 + 5.25 \cdot \frac{4}{11} \;=\; 3.909 \\ g_*(T) &= 2 + 5.25\left(\frac{4}{11}\right)^{\frac{4}{3}} \;=\; 3.363. \end{aligned} \tag{40}$$

These relations remain true for the photon+neutrino background as long as the neutrinos stay ultrarelativistic ($m_\nu \ll T$). It used to be the standard assumption that neutrinos are massless or that their masses are so small that they can be ignored, in which case the above relation would apply even today, when the photon (the CMB) temperature is $T = T_0 = 2.725$ K $= 0.2348$ meV, giving the neutrino background the temperature $T_{\nu 0} = 0.714 \cdot 2.725$ K $= 1.945$ K $= 0.1676$ meV today. However, *neutrino oscillation* experiments suggest a neutrino mass in the meV range, so that the neutrino background could be nonrelativistic today. In any case, the CMB (photon) temperature keeps redshifting as $T \propto a^{-1}$, so we can use Eq. (36) to relate the scale factor a and the CMB temperature T, keeping $g_{*s}(T) = 3.909$ all the way to the present time (and into the future).

Regardless of the question of neutrino masses, these relativistic backgrounds do not dominate the energy density of the universe any more today (photons + neutrinos still dominate the entropy density), as we shall discuss in Sec. 4.6.

4.5 Time scale of the early universe

The curvature term K/a^2 and dark energy can be ignored in the early universe, so the metric is

$$ds^2 = -dt^2 + a^2(t)\left[dr^2 + r^2 d\vartheta^2 + r^2 \sin^2\vartheta\, d\varphi^2\right]. \tag{41}$$

[6]To be more precise, neutrino decoupling was not complete when e^+e^--annihilation began; so that some of the energy and entropy leaked to the neutrinos. Therefore the neutrino energy density after e^+e^--annihilation is about 1.3% higher (at a given T) than the above calculation gives. The neutrino distribution also deviates slightly from kinetic equilibrium.

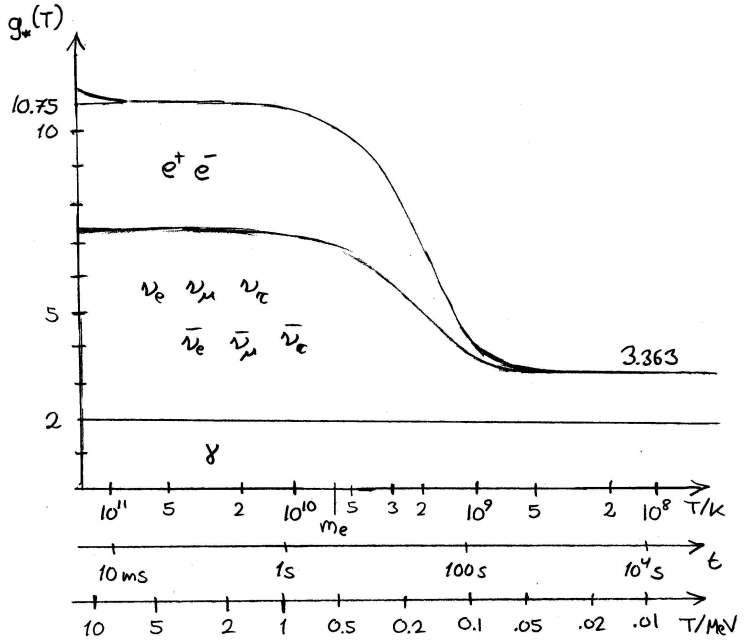


Figure 4: The evolution of the energy density, or rather, $g_*(T)$, and its different components through electron-positron annihilation. Since $g_*(T)$ is defined as $\rho/(\pi^2 T^4/30)$, where T is the photon temperature, the photon contribution appears constant. If we had plotted $\rho/(\pi^2 T_\nu^4/30) \propto \rho a^4$ instead, the neutrino contribution would appear constant, and the photon contribution would increase at the cost of the electron-positron contribution, which would better reflect what is going on.

and the Friedmann equation is

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho(T) = \frac{8\pi G}{3}\frac{\pi^2}{30}g_*(T)T^4. \tag{42}$$

To integrate this equation exactly we would need to calculate numerically the function $g_*(T)$ with all the annihilations[7]. For most of the time, however, $g_*(T)$ is changing slowly, so we can approximate $g_*(T) = \text{const}$. Then $T \propto a^{-1}$ and $H(t) = H(t_1)(a_1/a)^2$. Thus

$$dt = \frac{da}{a}H^{-1}(t_1)\left(\frac{a}{a_1}\right)^2 \qquad \Rightarrow \qquad t_1 = H^{-1}(t_1)\int_0^{a_1}\left(\frac{a}{a_1}\right)\frac{da}{a_1} = \frac{1}{2}H^{-1}(t_1)$$

and we get the relation

$$\boxed{t \;=\; \frac{1}{2}H^{-1} \;=\; \sqrt{\frac{45}{16\pi^3 G}}\frac{T^{-2}}{\sqrt{g_*}} \;=\; 0.301 g_*^{-1/2}\frac{m_{\rm Pl}}{T^2} \;=\; \frac{2.4}{\sqrt{g_*}}\left(\frac{\rm T}{\rm MeV}\right)^{-2}\rm s} \tag{43}$$

between the age of the universe t and the Hubble parameter H. Here

$$m_{\rm Pl} = \frac{1}{\sqrt{G}} = 1.22 \times 10^{19}\ \text{GeV}$$

is the Planck mass. Thus

$$a \propto T^{-1} \propto t^{1/2}.$$

[7]During electron annihilation one needs to calculate $g_{*s}(T)$ also, to get $T_\nu(T)$, needed for $g_*(T)$.

Except for a few special stages (like the QCD transition) the error from ignoring the time-dependence of $g_*(T)$ is small, since the time scales of earlier events are so much shorter, so the approximate result, Eq. (43), will be sufficient for us, as far as the time scale is concerned, when we use for each time t the value of g_* at that time. But for the relation between a and T, we need to use the more exact result, Eq. (36). Table 4 gives the times of the different events in the early universe.

Let us calculate (was already done in Chapter 3) the distance to the horizon $d_{\rm hor}(t_1) = a_1 r_{\rm hor}(t_1)$ at a given time t_1. For a radial light ray $dt = a(t)dr$ and from above $a(t) = (t/t_1)^{1/2} a_1$. Thus

$$t_1^{1/2} \int_0^{t_1} \frac{dt}{t^{1/2}} = a_1 \int_0^{r_{\rm hor}} dr \qquad \Rightarrow \qquad 2t_1 = a_1 r_{\rm hor} = d_{\rm hor}(t_1)$$

and we find for the horizon

$$d_{\rm hor} = 2t = H^{-1}. \tag{44}$$

Thus in the radiation-dominated early universe the distance to the horizon is equal to the Hubble length.

4.6 Matter

We noted that the early universe is dominated by the relativistic particles, and we can forget the nonrelativistic particles when we are considering the dynamics of the universe. We followed one species after another becoming nonrelativistic and disappearing from the picture, until only photons (the cosmic background radiation) and neutrinos were left, and even the latter of these had stopped interacting.

We must now return to the question what happened to the nucleons and the electrons. We found that they annihilated with their antiparticles when the temperature fell below their respective rest masses. For nucleons, the annihilation began immediately after they were formed in the QCD phase transition. There were however slightly more particles than antiparticles, and this small excess of particles was left over. (This must be so since we observe electrons and nucleons today). This means that the chemical potential $\mu_{\rm B}$ associated with baryon number differs from zero (is positive). Baryon number is a conserved quantity. Since nucleons are the lightest baryons, the baryon number resides today in nucleons (protons and neutrons; since the proton is lighter than the neutron, free neutrons have decayed into protons, but there are neutrons in atomic nuclei, whose mass/baryon is even smaller). The universe is electrically neutral, and the negative charge lies in the electrons, the lightest particles with negative charge. Therefore the number of electrons must equal the number of protons.

The number densities etc. of the electrons and the nucleons we get from the equations of Sec. 4.1. But what is the chemical potential μ in them? For each species, we get $\mu(T)$ from the conserved quantities.[8] The baryon number resides in the nucleons,

$$n_{\rm B} \,=\, n_{\rm N} - n_{\bar{\rm N}} \,=\, n_{\rm p} + n_{\rm n} - n_{\bar{\rm p}} - n_{\bar{\rm n}}\,. \tag{45}$$

Let us define the parameter η, the baryon-photon ratio today,

$$\eta \equiv \frac{n_{\rm B}(t_0)}{n_\gamma(t_0)}\,. \tag{46}$$

[8] In general, the recipe to find how the thermodynamical parameters, temperature and the chemical potentials, evolve in the expanding FRW universe, is to use the conservation laws of the conserved numbers, entropy conservation, and energy continuity, to find how the number densities and energy densities must evolve. The thermodynamical parameters will then evolve to satisfy these requirements.

From observations we know that $\eta = 10^{-10}$–10^{-9}. Since baryon number is conserved, $n_{\rm B}V \propto n_{\rm B}a^3$ stays constant, so

$$n_{\rm B} \propto a^{-3}\,. \tag{47}$$

After electron annihilation $n_\gamma \propto a^{-3}$, so we get

$$n_{\rm B}(T) = \eta n_\gamma = \eta\frac{2\zeta(3)}{\pi^2}T^3 \qquad \text{for } T \ll m_{\rm e}\,, \tag{48}$$

and for all times (as long as the universe expands adiabatically and the baryon number is conserved), using Eqs. (36), (47), and (48),

$$n_{\rm B}(T) = \eta\frac{2\zeta(3)}{\pi^2}\frac{g_{*s}(T)}{g_{*s}(T_0)}T^3\,. \tag{49}$$

For $T < 10$ MeV we have in practice

$$n_{\bar{\rm N}} \ll n_{\rm N} \qquad \text{and} \qquad n_{\rm N} \equiv n_{\rm n} + n_{\rm p} = n_{\rm B}\,.$$

We shall later (Chapter 5) discuss big bang nucleosynthesis—how the protons and neutrons formed atomic nuclei. Approximately one quarter of all nucleons (all neutrons and roughly the same number of protons) form nuclei ($A > 1$) and three quarters remain as free protons. Let us denote by $n_{\rm p}^*$ and $n_{\rm n}^*$ the number densities of protons and neutrons including those in nuclei (and also those in atoms), whereas we shall use $n_{\rm p}$ and $n_{\rm n}$ for the number densities of *free* protons and neutrons. Thus we write instead

$$n_{\rm N}^* \equiv n_{\rm n}^* + n_{\rm p}^* = n_{\rm B}\,.$$

In the same manner, for $T < 10$ keV we have

$$n_{\rm e^+} \ll n_{\rm e^-} \qquad \text{and} \qquad n_{\rm e^-} = n_{\rm p}^*\,.$$

At this time ($T \sim 10$ keV $\to 1$ eV) the universe contains a relativistic photon and neutrino background ("radiation") and nonrelativistic free electrons, protons, and nuclei ("matter"). Since $\rho \propto a^{-4}$ for radiation, but $\rho \propto a^{-3}$ for matter, the energy density in radiation falls eventually below the energy density in matter—the universe becomes *matter-dominated*.

The above discussion is in terms of the known particle species. Today there is much indirect observational evidence for the existence of what is called *cold dark matter* (CDM), which is supposedly made out of some yet undiscovered species of particles (this is discussed in Chapter 6). The CDM particles should be very weakly interacting (they decouple early), and their energy density contribution should be small when we are well in the radiation-dominated era, so they do not affect the above discussion much. They become nonrelativistic early and they are supposed to dominate the matter density of the universe (there appears to be about five times as much mass in CDM as in baryons). Thus the CDM causes the universe to become matter-dominated earlier than if the matter consisted of nucleons and electrons only. The CDM will be important later when we discuss (in Cosmology II) the formation of structure in the universe. The time of matter-radiation equality $t_{\rm eq}$ is calculated in an exercise at the end of this chapter.

4.7 Neutrino masses

The observed phenomenon of *neutrino oscillations*, where neutrinos change their flavor (i.e., whether they are ν_e, ν_μ, or ν_τ) periodically, is an indication of differences in the neutrino masses and therefore the neutrinos cannot all be massless. The oscillation phenomenon is a quantum mechanical effect, and is due to the mass eigenstates of neutrinos (a quantum state with definite

Normal				Inverted			
m_1	m_2	m_3	$\sum m_i$	m_3	m_1	m_2	$\sum m_i$
0	8.7 meV	50 meV	59 meV	0	50 meV	50.7 meV	101 meV
100 meV	100.4 meV	112 meV	312 meV	100 meV	111.8 meV	112.1 meV	324 meV
2 eV	2 eV	2 eV	6 eV	2 eV	2 eV	2 eV	6 eV

Table 3: Possibilities for neutrino masses.

mass) not being the same as the flavor eigenstates (a quantum state with definite flavor). The key point is that how the period of oscillation depends on the neutrino energy is related to a difference in mass squared, Δm^2, between these mass eigenstates. There are two different observed oscillation phenomena, solar neutrino oscillations (neutrinos coming from the Sun, produced as ν_e) and atmospheric neutrino oscillations (neutrinos produced as ν_μ and ν_e in the atmosphere by cosmic rays), and they provide a measurement of two differences:

$$\begin{aligned} \Delta m^2_{21} &\equiv m_2^2 - m_1^2 \approx 7.5\times10^{-5}\,\mathrm{eV}^2 \qquad \text{(solar)} \\ |\Delta m^2_{31}| &\equiv |m_3^2 - m_1^2| \approx |\Delta m^2_{32}| \approx 2.5\times10^{-3}\,\mathrm{eV}^2 \qquad \text{(atmospheric)}\,. \end{aligned} \tag{50}$$

Two of the mass eigenstates, labeled m_1 and m_2, are thus close to each other and $m_1 < m_2$; but we do not know whether the third mass eigenstate has a larger or smaller mass. These two possibilities are called the *normal* ($m_1 < m_2 < m_3$) and *inverted* ($m_3 < m_1 < m_2$) hierarchies. The neutrino mixing matrix, which relates the mass and flavor eigenstates, is not known well, but it appears that m_2 is a roughly equal mixture of all three flavors, and if we have the normal hierarchy, m_3 is mostly ν_μ and ν_τ.[2]

Since we have a laboratory upper limit $m < 2\,\mathrm{eV}$ for ν_e, the smallest of these mass eigenstates must be $< 2\,\mathrm{eV}$. (Measurement of the mass of a neutrino flavor projects the flavor state into a mass state, giving m_1, m_2, or m_3 with different probabilities; the upper limit presumably refers to the mass expectation value of the flavor state.) To have an idea what these Δm^2 mean for neutrino masses, consider three possibilities for the lowest mass eigenstate: $m = 0$, $m = 100\,\mathrm{meV}$, and $m = 2\,\mathrm{eV}$. This gives Table 3 and we conclude that the sum of the three neutrino masses must lie between $\sim 0.06\,\mathrm{eV}$ and $\sim 6\,\mathrm{eV}$, and that if we have the inverted hierarchy, it should be at least $0.1\,\mathrm{eV}$. The smallest possibility, where $m_1 \ll m_2$ and $\sum m_i = 0.06\,\mathrm{eV}$, is perhaps the most natural one and is considered as part of the standard model of cosmology (and the other possibilities are "extensions" of this standard model).

4.8 Recombination

Radiation (photons) and matter (electrons, protons, and nuclei) remained in thermal equilibrium for as long as there were lots of free electrons. When the temperature became low enough the electrons and nuclei combined to form neutral atoms (*recombination*), and the density of free electrons fell sharply. The *photon mean free path* grew rapidly and became longer than the horizon distance. Thus the universe became *transparent*. Photons and matter *decoupled*, i.e., their interaction was no more able to maintain them in thermal equilibrium with each other. After this, by T we refer to the photon temperature. Today, these photons are the CMB, and $T = T_0 = 2.725$ K. (After photon decoupling, the matter temperature fell at first faster than the photon temperature, but structure formation then heated up the matter to different temperatures at different places.)

To simplify the discussion of recombination, let us forget other nuclei than protons (in reality over 90% (by number) of the nuclei are protons, and almost all the rest are ^{4}He nuclei). Let us denote the number density of *free* protons by n_p, free electrons by n_e, and hydrogen atoms

by $n_{\rm H}$. Since the universe is electrically neutral, $n_{\rm p} = n_{\rm e}$. The conservation of baryon number gives $n_{\rm B} = n_{\rm p} + n_{\rm H}$. From Sec. 4.1 we have

$$n_i = g_i \left(\frac{m_i T}{2\pi}\right)^{3/2} e^{\frac{\mu_i - m_i}{T}}. \tag{51}$$

For as long as the reaction

$$p + e^- \;\leftrightarrow\; {\rm H} + \gamma \tag{52}$$

is in chemical equilibrium the chemical potentials are related by $\mu_{\rm p} + \mu_{\rm e} = \mu_{\rm H}$ (since $\mu_\gamma = 0$). Using this we get the relation

$$n_{\rm H} = \frac{g_{\rm H}}{g_{\rm p} g_{\rm e}} n_{\rm p} n_{\rm e} \left(\frac{m_{\rm e} T}{2\pi}\right)^{-3/2} e^{B/T}, \tag{53}$$

between the number densities. Here $B = m_{\rm p} + m_{\rm e} - m_{\rm H} = 13.6$ eV is the *binding energy* of hydrogen. The numbers of internal degrees of freedom are $g_{\rm p} = g_{\rm e} = 2$, $g_{\rm H} = 4$. Outside the exponent we approximated $m_{\rm H} \approx m_{\rm p}$. Defining the *fractional ionization*

$$x \equiv \frac{n_{\rm p}}{n_{\rm B}} \quad \Rightarrow \quad \frac{n_{\rm H}}{n_{\rm p} n_{\rm e}} = \frac{(1-x)}{x^2 n_{\rm B}}. \tag{54}$$

Using (48), Eq. (53) becomes

$$\frac{1-x}{x^2} = \frac{4\sqrt{2}\,\zeta(3)}{\sqrt{\pi}} \eta \left(\frac{T}{m_{\rm e}}\right)^{3/2} e^{B/T}, \tag{55}$$

the *Saha equation* for ionization in thermal equilibrium. When $B \ll T \ll m_{\rm e}$, the RHS $\ll 1$ so that $x \sim 1$, and almost all protons and electrons are free. As temperature falls, $e^{B/T}$ grows, but since both η and $(T/m_{\rm e})^{3/2}$ are $\ll 1$, the temperature needs to fall to $T \ll B$, before the whole expression becomes large (~ 1 or $\gg 1$).

The ionization fraction at first follows the equilibrium result of Eq. (55) closely, but as this equilibrium fraction begins to fall rapidly, the true ionization fraction begins to lag behind. As the number densities of free electrons and protons fall, it becomes more difficult for them to find each other to "recombine", and they are no longer able to maintain chemical equilibrium for the reaction (52). To find the correct ionization evolution, $x(t)$, requires then a more complicated calculation involving the reaction cross section of this reaction. See Figs. 5 and 6.

Although the equilibrium formula is thus not enough to give us the true ionization evolution, its benefit is twofold:

1. It tells us when recombination begins. While the equilibrium ionization changes slowly, it is easy to stay in equilibrium. Thus things won't start to happen until the equilibrium fraction begins to change a lot.

2. It gives the initial conditions for the more complicated calculation that will give the true evolution.

A similar situation holds for many other events in the early universe, e.g., big bang nucleosynthesis.

The recombination is not instantaneous. Let us define the recombination temperature $T_{\rm rec}$ as the temperature where $x = 0.5$. Now $T_{\rm rec} = T_0(1 + z_{\rm rec})$ since $1 + z = a^{-1}$ and the photon temperature falls as $T \propto a^{-1}$. (Since $\eta \ll 1$, the energy release in recombination is negligible compared to ρ_γ; and after photon decoupling photons travel freely maintaining kinetic equilibrium with $T \propto a^{-1}$.)

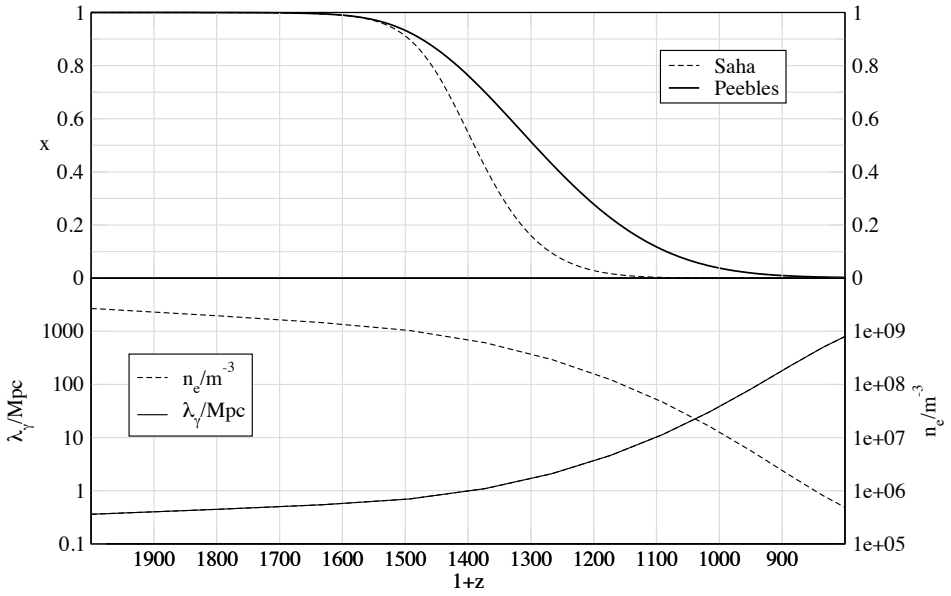


Figure 5: Recombination. In the top panel the dashed curve gives the equilibrium ionization fraction as given by the Saha equation. The solid curve is the true ionization fraction, calculated using the actual reaction rates (original calculation by Peebles). You can see that the equilibrium fraction is followed at first, but then the true fraction lags behind. The bottom panel shows the free electron number density $n_{\rm e}$ and the photon mean free path λ_γ. The latter is given in comoving units, i.e., the distance is scaled to the corresponding present distance. This figure is for $\eta = 8.22 \times 10^{-10}$. (Figure by R. Keskitalo.)

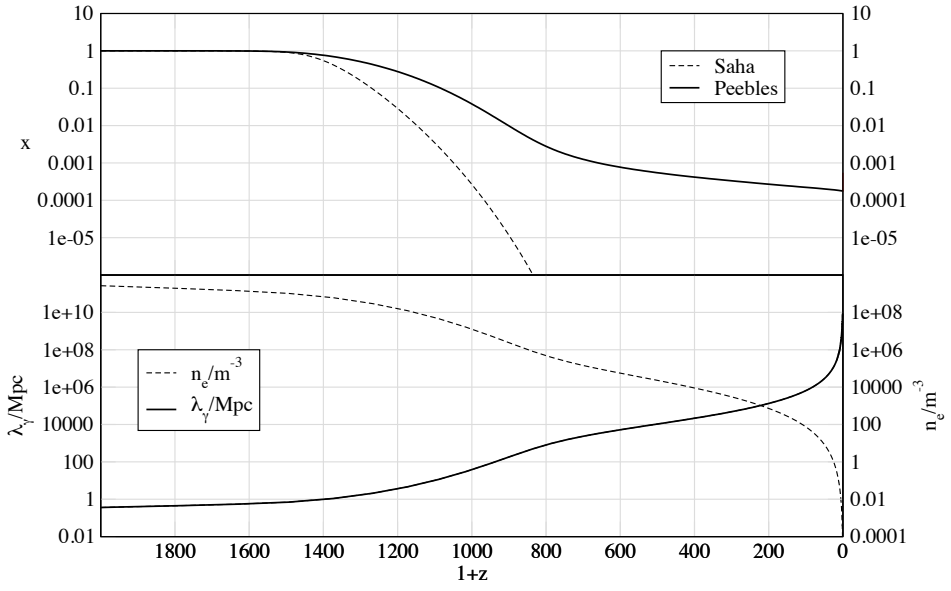


Figure 6: Same as Fig. 5, but with a logarithmic scale for the ionization fraction, and the time (actually redshift) scale extended to present time ($z = 0$ or $1 + z = 1$). You can see how a residual ionization $x \sim 10^{-4}$ remains. This figure does not include the reionization which happened at around $z \sim 10$. (Figure by R. Keskitalo.)

We get (for $\eta \sim 10^{-9}$)

$$\begin{aligned} T_{\rm rec} &\sim 0.3 \text{ eV} \\ z_{\rm rec} &\sim 1300. \end{aligned}$$

You might have expected that $T_{\rm rec} \sim B$. Instead we found $T_{\rm rec} \ll B$. The main reason for this is that $\eta \ll 1$. This means that there are very many photons for each hydrogen atom. Even when $T \ll B$, the high-energy tail of the photon distribution contains photons with energy $E > B$ so that they can ionize a hydrogen atom.

The photon decoupling takes place somewhat later, at $T_{\rm dec} \equiv (1+z_{\rm dec})T_0$, when the ionization fraction has fallen enough. We define the photon decoupling time to be the time when the photon mean free path exceeds the Hubble distance. The numbers are roughly

$$\begin{aligned} T_{\rm dec} &\sim 3000 \text{ K} \sim 0.26 \text{ eV} \\ z_{\rm dec} &\sim 1090. \end{aligned}$$

The decoupling means that the recombination reaction can not keep the ionization fraction on the equilibrium track, but instead we are left with a residual ionization of $x \sim 10^{-4}$.

A long time later ($z \sim 10$) the first stars form, and their radiation *reionizes* the gas that is left in interstellar space. The gas has now such a low density however, that the universe remains transparent.

Exercise: Transparency of the universe. We say the universe is transparent when the photon mean free path λ_γ is larger than the Hubble length $l_H = H^{-1}$, and opaque when $\lambda_\gamma < l_H$. The photon mean free path is determined mainly by the scattering of photons by free electrons, so that $\lambda_\gamma = 1/(\sigma_T n_e)$, where $n_e = x n_e^*$ is the number density of free electrons, n_e^* is the total number density of electrons, and x is the ionization fraction. The cross section for photon-electron scattering is independent of energy for $E_\gamma \ll m_e$ and is then called the Thomson cross section, $\sigma_T = \frac{8\pi}{3}(\alpha/m_e)^2$, where α is the fine-structure constant. In recombination x falls from 1 to 10^{-4}. Show that the universe is opaque before recombination and transparent after recombination. (Assume the recombination takes place between $z = 1300$ and $z = 1000$. You can assume a matter-dominated universe—see below for parameter values.) The interstellar matter gets later reionized (to $x \sim 1$) by the light from the first stars. What is the earliest redshift when this can happen without making the universe opaque again? (You can assume that most ($\sim$ all) matter has remained interstellar). Calculate for $\Omega_m = 1.0$ and $\Omega_m = 0.3$ (note that Ω_m includes nonbaryonic matter). Use $\Omega_\Lambda = 0$, $h = 0.7$ and $\eta = 6 \times 10^{-10}$.

The photons in the cosmic background radiation have thus traveled without scattering through space all the way since we had $T = T_{\rm dec} = 1091\, T_0$. When we look at this cosmic background radiation we thus see the universe (its faraway parts near our horizon) as it was at that early time. Because of the redshift, these photons which were then largely in the visible part of the spectrum, have now become microwave photons, so this radiation is now called the *cosmic microwave background* (CMB). It still maintains the kinetic equilibrium distribution. This was confirmed to high accuracy by the FIRAS (Far InfraRed Absolute Spectrophotometer) instrument on the COBE (Cosmic Background Explorer) satellite in 1989. John Mather received the 2006 Physics Nobel Prize for this measurement of the CMB frequency (photon energy) spectrum (see Fig. 7).[9]

We shall now, for a while, stop the detailed discussion of the history of the universe at these events, recombination and photon decoupling. The universe is about 400 000 years old now. What will happen next, is that the structure of the universe (galaxies, stars) begins to form, as gravity begins to draw matter into overdense regions. Before photon decoupling the radiation pressure from photons prevented this. But before going to the physics of *structure formation* (discussed in Cosmology II) we shall discuss some earlier events (big bang nucleosynthesis, ...) in more detail.

[9] He shared the Nobel Prize with George Smoot, who got it for the discovery of the CMB anisotropy with the DMR instrument on the same satellite. The CMB anisotropy will be discussed in Cosmology II.

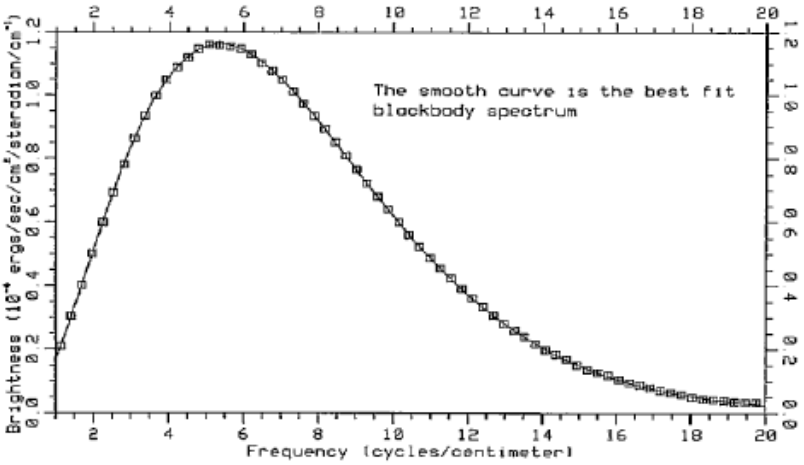


Figure 7: The CMB frequency spectrum as measured by the FIRAS instrument on the COBE satellite[3]. This first spectrum from FIRAS is based on just 9 minutes of measurements. The CMB temperature estimated from it was $T = 2.735 \pm 0.060\,\mathrm{K}$. The final result[4] from FIRAS is $T = 2.725 \pm 0.001\,\mathrm{K}$ (68% confidence; [4] actually gives this as $T = 2.725 \pm 0.002\,\mathrm{K}$ at 95% confidence).

Electroweak Transition	$T \sim 100$ GeV	$t \sim 20$ ps
QCD Transition	$T \sim 150$ MeV	$t \sim 20\mu$s
Neutrino Decoupling	$T \sim 1$ MeV	$t \sim 1$ s
Electron-Positron Annihilation	$T < m_e \sim 0.5$ MeV	$t \sim 10$ s
Big Bang Nucleosynthesis	$T \sim 50$–100 keV	$t \sim 10$ min
Matter-Radiation Equality	$T \sim 0.8$ eV ~ 9000 K	$t \sim 60000$ yr
Recombination + Photon Decoupling	$T \sim 0.3$ eV ~ 3000 K	$t \sim 380000$ yr

Table 4: Early universe events.

4.9 The Dark Age

How would the universe after recombination appear to an observer with human eyes? At first one would see a uniform red glow everywhere, since the wavelengths of the CMB photons are in the visible range. (It would also feel rather hot, 3000 K). As time goes on this glow gets dimmer and dimmer as the photons redshift towards the infrared, and after a few million years it gets completely dark, as the photons become invisible infrared (heat) radiation. There are no stars yet. This is often called the *dark age* of the universe. It lasts several hundred million years. While it lasts, it gradually gets cold. In the dark, however, masses are gathering together. And then, one by one, the first stars light up.

The decoupling of photon from baryonic matter (electrons, protons, nuclei, ions, atoms) is actually very asymmetric, since there is over 10^9 photons for each nucleus. The photon decoupling redshift $z = 1090$ is when photons decouple from baryons. After that, most photons will never scatter. However, some do, and these are enough to keep the temperature of the baryonic matter the same as photon temperature down to $z \sim 200$. After that, the decoupling is complete also from the baryonic point of view.[10] The baryonic matter (mainly hydrogen and helium gas) remains in internal kinetic equilibrium, but its temperature T_b falls now as a^{-2} (momentum redshifts as a^{-1}, and for nonrelativistic particles kinetic energy is $p^2/2m$ and mean kinetic energy is related to temperature by $\langle E_k \rangle = 3T/2$). So at $z \sim 20$, the baryon temperature is only a few K, about 1/10 of the photon temperature then. This is their coldest moment, since sometime after $z \sim 20$ the first stars form and begin to heat up the interstellar gas.

It seems that the star-formation rate peaked between redshifts $z = 1$ and $z = 2$. Thus the universe at a few billion years was brighter than it is today, since the brightest stars are short-lived, and the galaxies were closer to each other then.[11]

4.10 The radiation and neutrino backgrounds

While the starlight is more visible to us than the cosmic microwave background, its average energy density and photon number density in the universe is much less. Thus the photon density is essentially given by the CMB. The number density of CMB photons today ($T_0 = 2.725$ K) is

$$n_{\gamma 0} = \frac{2\zeta(3)}{\pi^2} T_0^3 = 410.5 \text{ photons/cm}^3 \tag{56}$$

and the energy density is

$$\rho_{\gamma 0} = 2\frac{\pi^2}{30} T_0^4 = 2.701 T_0 n_{\gamma 0} = 4.641 \times 10^{-31} \text{kg/m}^3 \,. \tag{57}$$

Since the critical density is

$$\rho_{\text{cr}0} = \frac{3H_0^2}{8\pi G} = h^2 \cdot 1.8788 \times 10^{-26} \text{kg/m}^3 \tag{58}$$

we get for the photon density parameter

$$\Omega_\gamma \equiv \frac{\rho_{\gamma 0}}{\rho_{\text{cr}0}} = 2.47 \times 10^{-5} h^{-2} \,. \tag{59}$$

[10] There is a similar asymmetry in neutrino decoupling. From the neutrino point of view, the decoupling temperature is $T \sim 3\,\text{MeV}$, from the baryonic point of view $T \sim 0.8\,\text{MeV}$.

[11] To be fair, galaxies seen from far away are rather faint objects, difficult to see with the unaided eye. In fact, if you were suddenly transported to a random location in the present universe, you might not be able to see anything. Thus, to enjoy the spectacle, our hypothetical observer should be located within a forming galaxy, or equipped with a good telescope.

While relativistic, neutrinos contribute another radiation component

$$\rho_\nu = \frac{7N_\nu}{4}\frac{\pi^2}{30}T_\nu^4\,. \tag{60}$$

After e^+e^--annihilation this gives

$$\rho_\nu = \frac{7N_\nu}{8}\left(\frac{4}{11}\right)^{\frac{4}{3}}\rho_\gamma\,, \tag{61}$$

where $N_\nu = 3$ is the number of neutrino species.

When the number of neutrino species was not yet known, cosmology (BBN) was used to constrain it. Big bang nucleosynthesis is sensitive to the expansion rate in the early universe, and that depends on the energy density. Observations of abundances of light element isotopes combined with BBN calculations require $N_\nu = 2$–4. Actually any new light particle species that would be relativistic at nucleosynthesis time ($T \sim 50$ keV – 1 MeV) and would thus contribute to the expansion rate through its energy density, but which would not interact directly with nuclei and electrons, would have the same effect. Thus such hypothetical unknown particles (called *dark radiation*) may not contribute to the energy density of the universe at that time more than one neutrino species does.

If we take Eq. (61) to define N_ν, but then take into account the extra contribution to ρ_ν from energy leakage during e^+e^--annihilation (and some other small effects), we get (as a result of years of hard work by many theorists)

$$N_\nu = 3.04\,. \tag{62}$$

(So this does not mean that there are 3.04 neutrino species. It means that the total energy density in neutrinos is 3.04 times as much as the energy density one neutrino species would contribute had it decoupled completely before e^+e^--annihilation.)

If neutrinos are still relativistic today, the neutrino density parameter is

$$\Omega_\nu = \frac{7N_\nu}{22}\left(\frac{4}{11}\right)^{\frac{1}{3}}\Omega_\gamma = 1.71\times10^{-5}h^{-2}\,, \tag{63}$$

so that the total radiation density parameter is

$$\Omega_r = \Omega_\gamma + \Omega_\nu = 4.18\times10^{-5}h^{-2} \sim 10^{-4}\,. \tag{64}$$

We thus confirm the claim in Chapter 3, that the radiation component can be ignored in the Friedmann equation, except in the early universe. The combination $\Omega_i h^2$ is often denoted by ω_i, so we have

$$\omega_\gamma = 2.47\times10^{-5} \tag{65}$$

$$\omega_\nu = 1.71\times10^{-5} \tag{66}$$

$$\omega_r = \omega_\gamma + \omega_\nu = 4.18\times10^{-5}\,. \tag{67}$$

Neutrino oscillation experiments indicate a neutrino mass in the meV–eV range. This means that neutrinos are nonrelativistic today and count as matter, not radiation, except possibly the lightest neutrino species. Then the above result for the neutrino energy density of the present universe does not apply. However, unless the neutrino masses are above 0.2 eV, they would still have been relativistic, and counted as radiation, at the time of recombination and matter-radiation equality. While the neutrinos are relativistic, one still gets the neutrino energy density as

$$\rho_\nu = \Omega_\nu\rho_{\rm cr0}a^{-4} \tag{68}$$

using the Ω_ν of Eq. (63), even though this relation does not hold when the neutrinos become nonrelativistic and thus this Ω_ν is not the density parameter to give the present density of neutrinos (we shall discuss that in Chapter 6).

Exercise: Matter–radiation equality. The present density of matter is $\rho_{m0} = \Omega_m \rho_{\rm cr0}$ and the present density of radiation is $\rho_{r0} = \rho_{\gamma 0} + \rho_{\nu 0}$ (we assume neutrinos are massless). What was the age of the universe t_{eq} when $\rho_m = \rho_r$? (Note that in these early times—but not today—you can ignore the curvature and vacuum terms in the Friedmann equation.) Give numerical value (in years) for the cases $\Omega_m = 0.1$, 0.3, and 1.0, and $H_0 = 70$km/s/Mpc. What was the temperature ($T_{\rm eq}$) then?

References

[1] H. Waltari, private communication (2016).

[2] K. Huitu, private communication (2017).

[3] J.C. Mather et al., Astrophys. J. Lett. **354**, 37 (1990).

[4] J.C. Mather et al., Astrophys. J. **512**, 511 (1999).

5 Big Bang Nucleosynthesis

One quarter (by mass) of the baryonic matter in the universe is helium. Heavier elements make up a few per cent. The rest, i.e., the major part, is hydrogen.

The building blocks of atomic nuclei, the nucleons, or protons and neutrons, formed in the QCD transition at $T \sim 150\,\mathrm{MeV}$ and $t \sim 20\,\mu\mathrm{s}$. The protons are hydrogen (^{1}H) nuclei.

Elements (their nuclei) heavier than helium, and also some of the helium, have mostly been produced by stars in different processes (see Fig. 1). However, the amount of helium and the presence of significant amounts of the heavier hydrogen isotope, deuterium (^{2}H), in the universe cannot be understood by these astrophysical mechanisms. It turns out that ^{2}H, ^{3}He, ^{4}He, and a significant part of ^{7}Li, were mainly produced already in the big bang, in a process we call *Big Bang Nucleosynthesis* (BBN).

The nucleons and antinucleons annihilated each other soon after the QCD transition, and the small excess of nucleons left over from annihilation did not have a significant effect on the expansion and thermodynamics of the universe until much later ($t \sim t_{\mathrm{eq}} = \Omega_m^{-2} h^{-4} 1000\,\mathrm{a}$), when the universe became matter-dominated. The ordinary matter in the present universe comes from this small excess of nucleons. Let us now consider what happened to it in the early universe. We shall focus on the period when the temperature fell from $T \sim 10\,\mathrm{MeV}$ to $T \sim 10\,\mathrm{keV}$ ($t \sim 10\,\mathrm{ms}$ – few h).

5.1 Equilibrium

The total number of nucleons stays constant due to baryon number conservation. This baryon number can be in the form of protons and neutrons or atomic nuclei. Weak nuclear reactions may convert neutrons and protons into each other and strong nuclear reactions may build nuclei from them.

During the period of interest the nucleons and nuclei are nonrelativistic ($T \ll m_p$). Assuming thermal equilibrium we have

$$n_i = g_i \left(\frac{m_i T}{2\pi}\right)^{3/2} e^{\frac{\mu_i - m_i}{T}} \tag{1}$$

for the number density of nucleus type i. If the nuclear reactions needed to build nucleus i (with mass number A and charge Z) from the nucleons,

$$(A-Z)\mathrm{n} + Z\mathrm{p} \quad \leftrightarrow \quad i\,,$$

occur at sufficiently high rate to maintain chemical equilibrium, we have

$$\mu_i = (A-Z)\mu_{\mathrm{n}} + Z\mu_{\mathrm{p}} \tag{2}$$

for the chemical potentials. Since for free nucleons

$$\begin{aligned} n_{\mathrm{p}} &= 2\left(\frac{m_{\mathrm{p}} T}{2\pi}\right)^{3/2} e^{\frac{\mu_{\mathrm{p}} - m_{\mathrm{p}}}{T}} \\ n_{\mathrm{n}} &= 2\left(\frac{m_{\mathrm{n}} T}{2\pi}\right)^{3/2} e^{\frac{\mu_{\mathrm{n}} - m_{\mathrm{n}}}{T}}, \end{aligned} \tag{3}$$

we can express n_i in terms of the neutron and proton densities,

$$n_i = g_i A^{\frac{3}{2}} 2^{-A} \left(\frac{2\pi}{m_{\mathrm{N}} T}\right)^{\frac{3}{2}(A-1)} n_{\mathrm{p}}^Z n_{\mathrm{n}}^{A-Z} e^{B_i/T}, \tag{4}$$

where

$$B_i \equiv Z m_{\mathrm{p}} + (A-Z) m_{\mathrm{n}} - m_i \tag{5}$$

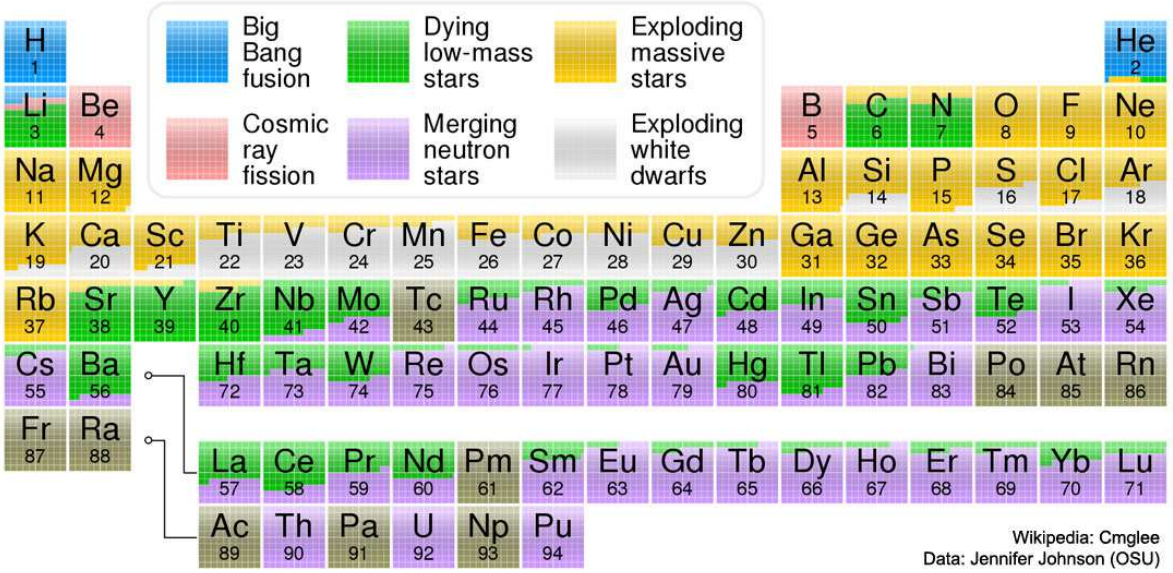


Figure 1: Astronomy Picture of the Day, 2017 October 24 (`https://apod.nasa.gov/apod/ap171024.html`: "Explanation: The hydrogen in your body, present in every molecule of water, came from the Big Bang. There are no other appreciable sources of hydrogen in the universe. The carbon in your body was made by nuclear fusion in the interior of stars, as was the oxygen. Much of the iron in your body was made during supernovas of stars that occurred long ago and far away. The gold in your jewelry was likely made from neutron stars during collisions that may have been visible as short-duration gamma-ray bursts or gravitational wave events. Elements like phosphorus and copper are present in our bodies in only small amounts but are essential to the functioning of all known life. The featured periodic table is color coded to indicate humanity's best guess as to the nuclear origin of all known elements. The sites of nuclear creation of some elements, such as copper, are not really well known and are continuing topics of observational and computational research.") In more scientific terms: During most of their lifetime (the main sequence phase), stars are powered by nuclear fusion of hydrogen into helium in their cores. When hydrogen is exhausted in the core they begin to fuse helium into heavier elements (the giant phase). How far this proceeds depends on the mass of the star. In the heaviest stars fusion proceeds all the way to iron (^{56}Fe). Beyond iron, fusion will no longer produce energy, since ^{56}Fe maximizes binding energy per nucleon. Heavier elements are thus produced in processes which need an energy source to power them. Some of the energy released by nuclear fusion in the stellar cores goes into production of these heavier elements in the giant phase. When the fusion energy is exhausted the star "dies": lighter stars collapse into white dwarfs, heavier stars explode—this explosion is called a supernova. A supernova begins with a collapse as the pressure produced by the fusion longer supports the outer parts. This brings in and raises the temperature of unburnt nuclear fuel from the outer parts. The fusion of this material and the gravitational energy from the collapse release a lot of energy in a short time causing the explosion, which is one source of heavier elements. Also white dwarfs may become supernovae later, if they accrete more mass from companion stars. In all these dying/exploding cases, the outer parts of the stars are ejected and mix into the interstellar material. In a supernova explosion of a massive star the inner part collapses into a neutron star. Collisions of these neutron stars are another source of heavy elements. The main type of nuclear reaction responsible for the production of the heavier elements beyond iron is neutron capture. Since neutrons are neutral it is easy for them to penetrate a nucleus and raise the mass number of the nucleus. The resulting new nucleus may be unstable so that it will β decay, i.e, the neutron releases an electron (and an antineutrino) and becomes a proton. In a slow neutron capture process (s-process) this decay happens before another neutron is captured, and in a rapid neutron capture process (r-process) many neutrons are captured before such decay. Heavy elements are produced by the s-process in the giant phase where fusion reactions provide the required energetic neutrons. The r-process requires a high density of neutrons. It is thought that the main site for the r-process is provided by collisions of neutron stars. Beryllium and boron are mainly produced by cosmic rays breaking up heavier nuclei in interstellar space (cosmic ray spallation).

is the binding energy of the nucleus. Here we have approximated $m_{\rm p} \approx m_{\rm n} \approx m_i/A$ outside the exponent, and denoted it by m_N ("nucleon mass").

${}^A Z$	B(MeV)	B/A(MeV)	g
^{2}H	2.2245	1.11	3
^{3}H	8.4820	2.83	2
^{3}He	7.7186	2.57	2
^{4}He	28.2970	7.07	1
^{6}Li	31.9965	5.33	3
^{7}Li	39.2460	5.61	4
^{7}Be	37.6026	5.37	1
^{12}C	92.1631	7.68	1
^{56}Fe	492.2623	8.79	1

Table 1. Some of the lightest nuclei (+ iron) and their binding energies.

The different number densities add up to the total baryon number density

$$\sum A_i n_i = n_B \,. \tag{6}$$

The baryon number density n_B can be expressed in terms of photon density

$$n_\gamma = \frac{2}{\pi^2}\zeta(3)T^3 \tag{7}$$

and the baryon/photon -ratio

$$\frac{n_B}{n_\gamma} = \frac{g_{*s}(T)}{g_{*s}(T_0)}\eta \tag{8}$$

as

$$n_B = \frac{g_{*s}(T)}{g_{*s}(T_0)}\eta\frac{2}{\pi^2}\zeta(3)T^3. \tag{9}$$

After electron-positron annihilation $g_{*s}(T) = g_{*s}(T_0)$ and $n_B = \eta n_\gamma$. Here η is the *present baryon/photon ratio*. It can be estimated from various observations in a number of ways. It's order of magnitude is 10^{-9}.

We define the *mass fraction* of nucleus i as

$$X_i \equiv \frac{A_i n_i}{n_{\rm B}} \qquad \text{so that} \qquad \sum X_i = 1\,. \tag{10}$$

The equilibrium mass fractions are, from Eq. (4),

$$X_i = \frac{1}{2}X_{\rm p}^Z X_{\rm n}^{A-Z} g_i A^{\frac{5}{2}} \epsilon^{A-1} e^{B_i/T} \tag{11}$$

where

$$\epsilon \equiv \frac{1}{2}\left(\frac{2\pi}{m_{\rm N}T}\right)^{3/2} n_{\rm B} \;=\; \frac{1}{\pi^2}\zeta(3)\left(\frac{2\pi T}{m_{\rm N}}\right)^{3/2}\frac{g_{*s}(T)}{g_{*s}(T_0)}\eta \sim \left(\frac{T}{m_{\rm N}}\right)^{3/2}\eta.$$

The factors which change rapidly with T are $\epsilon^{A-1}e^{B_i/T}$. For temperatures $m_{\rm N} \gg T \gtrsim B_i$ we have $e^{B_i/T} \sim 1$ and $\epsilon \ll 1$. Thus $X_i \ll 1$ for others ($A > 1$) than protons and neutrons. As temperature falls, ϵ becomes even smaller and at $T \sim B_i$ we have $X_i \ll 1$ still. The temperature has to fall below B_i by a large factor before the factor $e^{B_i/T}$ wins and the equilibrium abundance becomes large. We calculate below that, e.g., for ^{4}He this happens at $T \sim 0.3\,$MeV, and for ^{2}H at $T \sim 0.07\,$MeV. Thus we have initially only free neutrons and protons in large numbers.

5.2 Neutron-proton ratio

What can we say about $n_{\rm p}$ and $n_{\rm n}$? Protons and neutrons are converted into each other in the weak reactions

$$\begin{aligned} {\rm n} + \nu_{\rm e} &\leftrightarrow {\rm p} + {\rm e}^- \\ {\rm n} + {\rm e}^+ &\leftrightarrow {\rm p} + \bar{\nu}_{\rm e} \\ {\rm n} &\leftrightarrow {\rm p} + {\rm e}^- + \bar{\nu}_{\rm e}. \end{aligned} \tag{12}$$

If these reactions are in equilibrium, $\mu_{\rm n} + \mu_{\nu_{\rm e}} = \mu_{\rm p} + \mu_{\rm e}$, and the neutron/proton ratio is

$$\frac{n_{\rm n}}{n_{\rm p}} = e^{-Q/T + (\mu_{\rm e} - \mu_{\nu_{\rm e}})/T}, \tag{13}$$

where $Q \equiv m_{\rm n} - m_{\rm p} = 1.293\,{\rm MeV}$.

We need now some estimate for the chemical potentials of electrons and electron neutrinos. The universe is electrically neutral[1], so the net number of electrons $({\rm e}^- - n_{\rm e^+})$ equals the number of protons, and $\mu_{\rm e}$ can be calculated exactly in terms of η and T. We leave the exact calculation as an exercise, but give below a rough estimate for the ultrarelativistic limit ($T > m_{\rm e}$):

In the ultrarelativistic limit

$$n_{\rm e^-} - n_{\rm e^+} = \frac{2T^3}{6\pi^2}\left(\pi^2\left(\frac{\mu_{\rm e}}{T}\right) + \left(\frac{\mu_{\rm e}}{T}\right)^3\right) = n_{\rm p}^* \approx n_{\rm B} \approx \eta n_\gamma = \eta\frac{2}{\pi^2}\zeta(3)T^3. \tag{14}$$

Here n_p^* includes the protons inside nuclei. Since η is small, we must have $\mu_{\rm e} \ll T$, and we can drop the $(\mu_{\rm e}/T)^3$ term to get

$$\frac{\mu_{\rm e}}{T} \approx \frac{6}{\pi^2}\zeta(3)\eta. \tag{15}$$

Thus $\mu_{\rm e}/T \sim \eta \sim 10^{-9}$. The nonrelativistic limit can be done in a similar manner (**exercise**). It turns out that $\mu_{\rm e}$ rises as T falls, and somewhere between $T = 30\,{\rm keV}$ and $T = 10\,{\rm keV}$ $\mu_{\rm e}$ becomes larger than T, and, in fact, comparable to $m_{\rm e}$.

For $T \gtrsim 30\,{\rm keV}$, $\mu_{\rm e} \ll T$, and we can drop the $\mu_{\rm e}$ in Eq. (13).

Since we cannot detect the cosmic neutrino background, we don't know the neutrino chemical potentials. Usually it is *assumed* that also all three $\mu_{\nu_i} \ll T$, so that the difference in the number of neutrinos and antineutrinos is small. Thus we ignore both $\mu_{\rm e}$ and $\mu_{\nu_{\rm e}}$, so that $\mu_{\rm p} = \mu_{\rm n}$ and the equilibrium neutron/proton ratio is

$$\frac{n_{\rm n}}{n_{\rm p}} = e^{-Q/T}. \tag{16}$$

(This is not valid for $T \lesssim 30\,{\rm keV}$, since $\mu_{\rm e}$ is no longer small, but we shall use this formula only at higher temperatures as will be seen below.)

For $T > 0.3\,{\rm MeV}$, we still have $X_{\rm n} + X_{\rm p} \approx 1$, so the equilibrium abundances are

$$X_{\rm n} = \frac{e^{-Q/T}}{1 + e^{-Q/T}} \qquad \text{and} \qquad X_{\rm p} = \frac{1}{1 + e^{-Q/T}}. \tag{17}$$

Nucleons follow these equilibrium abundances until neutrinos decouple at $T \sim 1\,{\rm MeV}$, shutting off the weak ${\rm n} \leftrightarrow {\rm p}$ reactions. After this the neutrons decay into protons, so that

$$X_{\rm n}(t) = X_{\rm n}(t_d)e^{-(t-t_1)/\tau_{\rm n}}, \tag{18}$$

where $\tau_{\rm n} = 880.2 \pm 1.0\,{\rm s}$ is the mean lifetime of a free neutron[1].[2] (The half-life is $\tau_{1/2} = (\ln 2)\tau_{\rm n}$.)

[1] Electromagnetism is stronger than gravity by a factor of about 10^{38} so that the possible excess in positive or negative charge should be much less than one per this number or otherwise it would have been noticed.

[2] This value given for $\tau_{\rm n}$ by the Particle Data Group has quite recently changed by much more than the claimed accuracy. From 2006 to 2010 the given value was $885.7 \pm 0.8\,{\rm s}$.

5.3 Bottlenecks

Using (4), (16), and (6)[3], we get all equilibrium abundances as a function of T (they also depend on the value of the parameter η). There are two items to note, however:

1. The normalization condition, Eq. (6), includes all nuclei up to uranium etc. Thus we would get a huge polynomial equation from which to solve $X_{\rm p}$. (After one has $X_{\rm p}$, one gets the rest easily from (4) and (16).)

2. In practice we don't have to care about the first item, since as the temperature falls the nuclei no longer follow their equilibrium abundances. The reactions are in equilibrium only at high temperatures, when the other equilibrium abundances except $X_{\rm p}$ and $X_{\rm n}$ are small, and we can use the approximation $X_{\rm n} + X_{\rm p} = 1$.

In the early universe the baryon density is too low and the time available is too short for reactions involving three or more incoming nuclei to occur at any appreciable rate. The heavier nuclei have to be built sequentially from lighter nuclei in two-particle reactions, so that deuterium is formed first in the reaction

$$\mathrm{n} + \mathrm{p} \quad \rightarrow \quad \mathrm{d} + \gamma.$$

Only when deuterons are available can helium nuclei be formed, and so on. This process has "bottlenecks": the lack of sufficient densities of lighter nuclei hinders the production of heavier nuclei, and prevents them from following their equilibrium abundances.

As the temperature falls, the equilibrium abundances rise fast. They become large later for nuclei with small binding energies. Since deuterium is formed directly from neutrons and protons it can follow its equilibrium abundance as long as there are large numbers of free neutrons available. Since the deuterium binding energy is rather small, the deuterium abundance becomes large rather late (at $T < 100\,\mathrm{keV}$). Therefore heavier nuclei with larger binding energies, whose equilibrium abundances would become large earlier, cannot be formed. This is the *deuterium bottleneck*. Only when there is lots of deuterium ($X_{\rm d} \sim 10^{-3}$), can helium be produced in large numbers.

The nuclei are positively charged and there is thus an electromagnetic repulsion between them. The nuclei need thus large kinetic energies to overcome this *Coulomb barrier* and get within the range of the strong interaction. Thus the cross sections for these fusion reactions fall rapidly with energy and the nuclear reactions are "shut off" when the temperature falls below $T \sim 30\,\mathrm{keV}$. Thus there is less than one hour available for nucleosynthesis. Because of additional bottlenecks (e.g., there are no stable nuclei with $A = 8$) and the short time available, only very small amounts of elements heavier than helium are formed.

5.4 Calculation of the helium abundance

Let us now calculate the numbers. We saw that because of the deuterium bottleneck, $X_{\rm n}+X_{\rm p} \approx 1$ holds until $T \sim 0.1\,\mathrm{MeV}$. Until then, we get $X_{\rm n}$ and $X_{\rm p}$ at first from (17) and then from (18). In reality, neutrino decoupling and thus the shift from behavior (17) to behavior (18) is not instantaneous, but an approximation where one takes it to be instantaneous at time t_1 when $T = 0.8\,\mathrm{MeV}$, so that $X_{\rm n}(t_1) = 0.1657$, gives a fairly accurate final result.

Deuterium has $B_{\rm d} = 2.22\,\mathrm{MeV}$, and we get $\epsilon e^{B_{\rm d}/T} = 1$ at $T_{\rm d} = 0.06\,\mathrm{MeV}$–$0.07\,\mathrm{MeV}$ (assuming $\eta = 10^{-10} - 10^{-8}$), so the deuterium abundance becomes large near this temperature. Since ^{4}He has a much higher binding energy, $B_4 = 28.3\,\mathrm{MeV}$, the corresponding situation $\epsilon^3 e^{B_4/T} = 1$ occurs at a higher temperature $T_4 \sim 0.3\,\mathrm{MeV}$. But we noted earlier that only deuterium stays

[3]For $n_{\rm p}$ and $n_{\rm n}$ we know just their ratio, since we do not know $\mu_{\rm p}$ and $\mu_{\rm n}$, only that $\mu_{\rm p} = \mu_{\rm n}$. Therefore this extra equation is needed to solve all n_i.

close to its equilibrium abundance once it gets large. Helium begins to form only when there is sufficient deuterium available, in practice slightly above $T_{\rm d}$. Helium forms then rapidly. The available number of neutrons sets an upper limit to ^{4}He production. Since helium has the highest binding energy per nucleon (of all isotopes below A=12), almost all neutrons end up in ^{4}He, and only small amounts of the other light isotopes, ^{2}H, ^{3}H, ^{3}He, ^{7}Li, and ^{7}Be, are produced.

The Coulomb barrier shuts off the nuclear reactions before there is time for heavier nuclei ($A > 8$) to form. One gets a fairly good approximation for the ^{4}He production by assuming instantaneous nucleosynthesis at $T = T_{\rm ns} \sim 1.1 T_{\rm d} \sim 70\,{\rm keV}$, with all neutrons ending up in ^{4}He, so that

$$X_4 \approx 2X_{\rm n}(T_{\rm ns}). \tag{19}$$

After electron annihilation ($T \ll m_{\rm e} = 0.511\,{\rm MeV}$) the time-temperature relation is

$$t = 0.301 g_*^{-1/2} \frac{m_{\rm Pl}}{T^2}\,, \tag{20}$$

where $g_* = 3.363$. Since most of the time in the temperature interval $T = 0.8\,{\rm MeV}$–$0.07\,{\rm MeV}$ is spent at the lower part of this temperature range, this formula gives a good approximation for the time

$$t_{\rm ns} - t_1 = 266.5\ {\rm s} \qquad \text{(in reality 264.3 s)}.$$

Thus we get for the final ^{4}He abundance

$$X_4 = 2X_{\rm n}(t_1) e^{-(t_{\rm ns}-t_1)/\tau_{\rm n}} = 24.5\ \%. \tag{21}$$

Accurate numerical calculations, using the reaction rates of the relevant weak and strong reaction rates give $X_4 = 21$–26 % (for $\eta = 10^{-10} - 10^{-9}$).

As a calculation of the helium abundance X_4 the preceding calculation is of course a cheat, since we have used the results of those accurate numerical calculations to infer that we need to use $T = 0.8\,{\rm MeV}$ as the neutrino decoupling temperature, and $T_{\rm ns} = 1.1 T_{\rm d}$ as the "instantaneous nucleosynthesis" temperature, to best approximate the correct behavior. However, it gives us a quantitative description of what is going on, and an understanding of how the helium yield depends on various things.

Exercise: Using the preceding calculation, find the dependence of X_4 on η, i.e., calculate $dX_4/d\eta$.

5.5 Why so late?

Let us return to the question, why the temperature has to fall so much below the binding energy before the equilibrium abundances become large. From the energetics one might conclude that when typical kinetic energies, $\langle E_k \rangle \approx \frac{3}{2} T$ for nuclei and $\langle E \rangle \approx 2.7\,T$ for photons, are smaller than the binding energy, it would be easy to form nuclei but difficult to break them. Above we saw that the smallness of the factor $\epsilon \sim (T/m_{\rm N})^{3/2}\eta$ is the reason why this is not so. Here $\eta \sim 10^{-9}$ and $(T/m_{\rm N})^{3/2} \sim 10^{-6}$ (for $T \sim 0.1\,{\rm MeV}$). The main culprit is thus the small baryon/photon ratio. Since there are 10^9 photons for each baryon, there is a sufficient amount of photons who can disintegrate a nucleus in the high-energy tail of the photon distribution, even at rather low temperatures. One can also express this result in terms of entropy. A high photon/baryon ratio corresponds to a high entropy per baryon. High entropy favors free nucleons.

5.6 The most important reactions

In reality, neither neutrino decoupling, nor nucleosynthesis, are instantaneous processes. Accurate results require a rather large numerical computation where one uses the cross sections of all the relevant weak and strong interactions. These cross sections are energy-dependent.

Integrating them over the energy and velocity distributions and multiplying with the relevant number densities leads to temperature-dependent reaction rates. The most important reactions are the weak $n \leftrightarrow p$ reactions (12) and the following strong reactions[4](see also Fig. 2):

$$\begin{array}{rcl}
\mathrm{p} + \mathrm{n} & \to & {}^2\mathrm{H} + \gamma \\
{}^2\mathrm{H} + \mathrm{p} & \to & {}^3\mathrm{He} + \gamma \\
{}^2\mathrm{H} + {}^2\mathrm{H} & \to & {}^3\mathrm{H} + \mathrm{p} \\
{}^2\mathrm{H} + {}^2\mathrm{H} & \to & {}^3\mathrm{He} + \mathrm{n} \\
\mathrm{n} + {}^3\mathrm{He} & \to & {}^3\mathrm{H} + \mathrm{p} \\
\mathrm{p} + {}^3\mathrm{H} & \to & {}^4\mathrm{He} + \gamma \\
{}^2\mathrm{H} + {}^3\mathrm{H} & \to & {}^4\mathrm{He} + \mathrm{n} \\
{}^2\mathrm{H} + {}^3\mathrm{He} & \to & {}^4\mathrm{He} + \mathrm{p} \\
{}^4\mathrm{He} + {}^3\mathrm{He} & \to & {}^7\mathrm{Be} + \gamma \\
{}^4\mathrm{He} + {}^3\mathrm{H} & \to & {}^7\mathrm{Li} + \gamma \\
{}^7\mathrm{Be} + \mathrm{n} & \to & {}^7\mathrm{Li} + \mathrm{p} \\
{}^7\mathrm{Li} + \mathrm{p} & \to & {}^4\mathrm{He} + {}^4\mathrm{He}
\end{array}$$

The cross sections of these strong reactions can't be calculated from first principles, i.e., from QCD, since QCD is too difficult. Instead one uses cross sections measured in laboratory. The cross sections of the weak reactions (12) are known theoretically (there is one parameter describing the strength of the weak interaction, which is determined experimentally, in practice by measuring the lifetime τ_n of free neutrons). The relevant reaction rates are now known sufficiently accurately, so that the nuclear abundances produced in BBN (for a given value of η) can be calculated with better accuracy than the present abundances can be measured from astronomical observations.

The reaction chain proceeds along stable and long-lived (compared to the nucleosynthesis timescale—minutes) isotopes towards larger mass numbers. At least one of the two incoming

[4]The reaction chain that produces helium from hydrogen in BBN is not the same that occurs in stars. The conditions is stars are different: there are no free neutrons and the temperatures are lower, but the densities are higher and there is more time available. In addition, second generation stars contain heavier nuclei (C,N,O) which can act as catalysts in helium production. Some of the most important reaction chains in stars are ([2], p. 251):

1. The proton-proton chain

$$\begin{array}{rcl}
\mathrm{p} + \mathrm{p} & \to & {}^2\mathrm{H} + \mathrm{e}^+ + \nu_\mathrm{e} \\
{}^2\mathrm{H} + \mathrm{p} & \to & {}^3\mathrm{He} + \gamma \\
{}^3\mathrm{He} + {}^3\mathrm{He} & \to & {}^4\mathrm{He} + \mathrm{p} + \mathrm{p},
\end{array}$$

2. and the CNO-chain

$$\begin{array}{rcl}
{}^{12}\mathrm{C} + \mathrm{p} & \to & {}^{13}\mathrm{N} + \gamma \\
{}^{13}\mathrm{N} & \to & {}^{13}\mathrm{C} + \mathrm{e}^+ + \nu_\mathrm{e} \\
{}^{13}\mathrm{C} + \mathrm{p} & \to & {}^{14}\mathrm{N} + \gamma \\
{}^{14}\mathrm{N} + \mathrm{p} & \to & {}^{15}\mathrm{O} + \gamma \\
{}^{15}\mathrm{O} & \to & {}^{15}\mathrm{N} + \mathrm{e}^+ + \nu_\mathrm{e} \\
{}^{15}\mathrm{N} + \mathrm{p} & \to & {}^{12}\mathrm{C} + {}^4\mathrm{He}.
\end{array}$$

The cross section of the direct reaction $\mathrm{d}+\mathrm{d} \to {}^4\mathrm{He}+\gamma$ is small (i.e., the ${}^3\mathrm{H}+\mathrm{p}$ and ${}^3\mathrm{He}+\mathrm{n}$ channels dominate $\mathrm{d}+\mathrm{d} \to$), and it is not important in either context.

The triple-α reaction ${}^4\mathrm{He}+{}^4\mathrm{He}+{}^4\mathrm{He} \to {}^{12}\mathrm{C}$, responsible for carbon production in stars, is also not important during big bang, since the density is not sufficiently high for three-particle reactions to occur (the three ${}^4\mathrm{He}$ nuclei would need to come within the range of the strong interaction within the lifetime of the intermediate state, ${}^8\mathrm{Be}$, 2.6×10^{-16} s). (Exercise: calculate the number and mass density of nucleons at $T = 1\,\mathrm{MeV}$.)

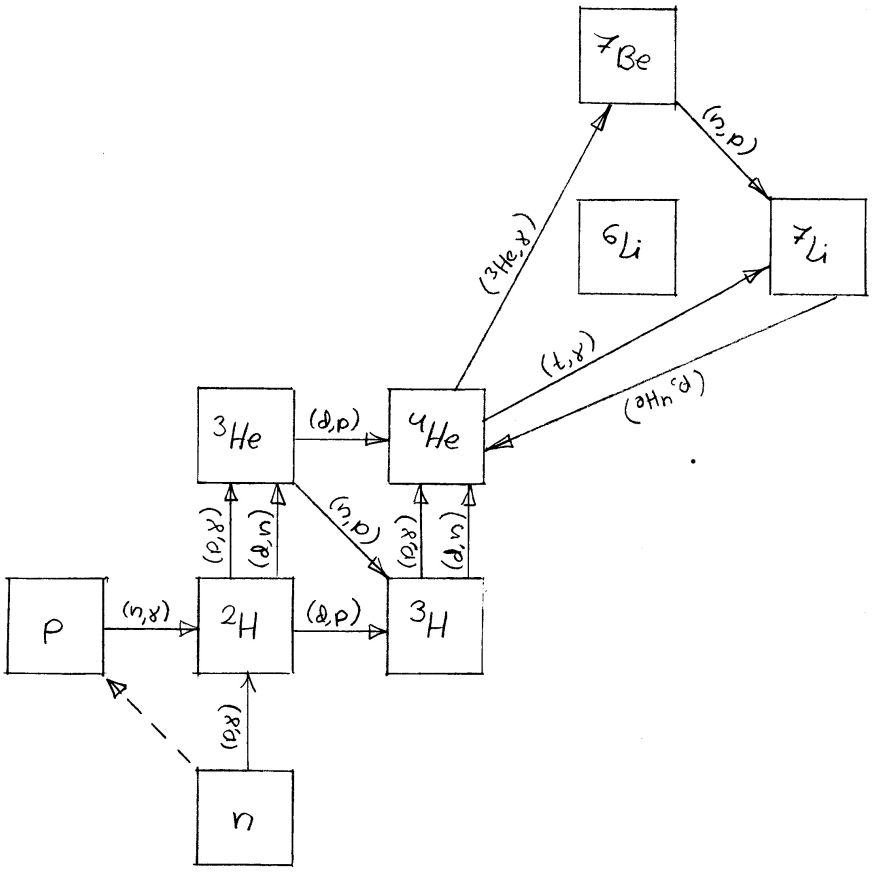


Figure 2: The 12 most important nuclear reactions in big bang nucleosynthesis.

nuclei must be an isotope which is abundant during nucleosynthesis, i.e., n, p, ^{2}H or ^{4}He. The mass numbers $A = 5$ and $A = 8$ form bottlenecks, since they have no stable or long-lived isotopes. These bottlenecks cannot be crossed with n or p. The $A = 5$ bottleneck is crossed with the reactions ^{4}He+^{3}He and ^{4}He+^{3}H, which form a small number of ^{7}Be and ^{7}Li. Their abundances remain so small that we can ignore the reactions (e.g., $^7\mathrm{Be} + {}^4\mathrm{He} \to {}^{11}\mathrm{C} + \gamma$ and $^7\mathrm{Li} + {}^4\mathrm{He} \to {}^{11}\mathrm{B} + \gamma$) which cross the $A = 8$ bottleneck. Numerical calculations also show that the production of the other stable lithium isotope, ^{6}Li is several orders of magnitude smaller than that of ^{7}Li.

Thus BBN produces the isotopes ^{2}H, ^{3}H, ^{3}He, ^{4}He, ^{7}Li and ^{7}Be. Of these, ^{3}H (half life 12.3 a) and ^{7}Be (53 d) are unstable and decay after nucleosynthesis into ^{3}He and ^{7}Li. (^{7}Be actually becomes ^{7}Li through electron capture $^7\mathrm{Be} + \mathrm{e}^- \to {}^7\mathrm{Li} + \nu_\mathrm{e}$.)

In the end BBN has produced cosmologically significant (compared to present abundances) amounts of the four isotopes, ^{2}H, ^{3}He, ^{4}He and ^{7}Li (the fifth isotope $^1\mathrm{H} = \mathrm{p}$ we had already before BBN). Their production in the BBN can be calculated, and there is only one free parameter, the baryon/photon ratio

$$\begin{aligned} \eta &\equiv \frac{n_\mathrm{B}}{n_\gamma} = \frac{\Omega_b \rho_{\mathrm{cr}0}}{m_N n_\gamma} = \frac{\Omega_b}{m_n n_\gamma} \frac{3H_0^2}{8\pi G} \\ &= 274 \times 10^{-10} \omega_b = 1.46 \times 10^{18} \left(\frac{\rho_{b0}}{\mathrm{kgm}^{-3}} \right). \end{aligned} \tag{22}$$

Here ρ_{b0} is the average density of ordinary, or baryonic, matter today, $\Omega_b \equiv \rho_{b0}/\rho_{\mathrm{cr}0}$ is the baryon density parameter, and $\omega_b \equiv \Omega_b h^2$.

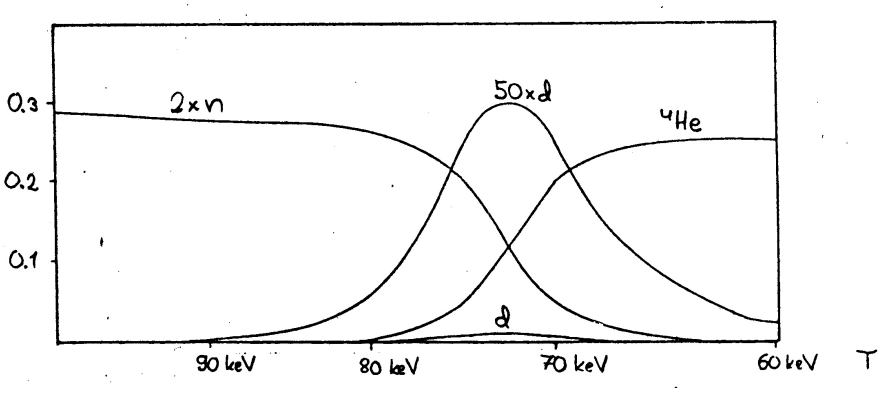


Figure 3: The time evolution of the n, ^{2}H (written as d) and ^{4}He abundances during BBN. Notice how the final ^{4}He abundance is determined by the n abundance before nuclear reactions begin. Only a small part of these neutrons decay or end up in other nuclei. Before becoming ^{4}He, all neutrons pass through ^{2}H. To improve the visibility of the deuterium curve, we have plotted it also as multiplied by a factor of 50. This figure is for $\eta = 6 \times 10^{-10}$. The time at $T = (90, 80, 70, 60)$ keV is $(152, 199, 266, 367)$ s. Thus the action peaks at about $t = 4$ min. The other abundances (except p) remain so low, that to see them the figure must be redrawn in logarithmic scale (see Fig. 4). From [3].

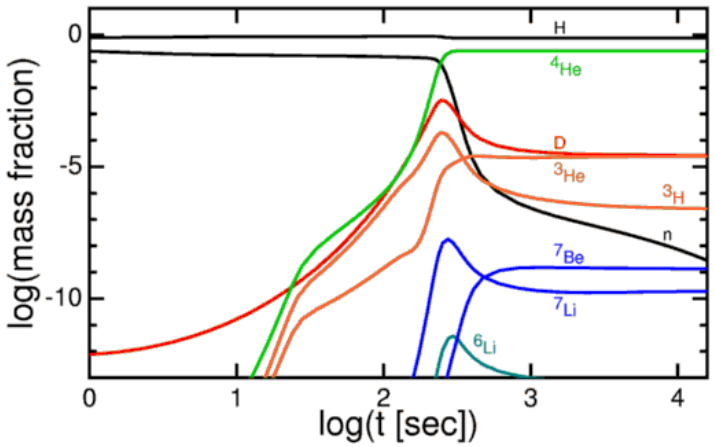


Figure 4: Time evolution of the abundances of the light isotopes during BBN. From `http://www.astro.ucla.edu/~wright/BBNS.html`

5.7 BBN as a function of time

Let us follow nucleosynthesis as a function of time (or decreasing temperature). See Figs. 3 and 4. ^{2}H and ^{3}H are intermediate states, through which the reactions proceed towards ^{4}He. Therefore their abundance first rises, is highest at the time when ^{4}He production is fastest, and then falls as the baryonic matter ends in ^{4}He. ^{3}He is also an intermediate state, but the main channel from ^{3}He to ^{4}He is via ^{3}He+n$\rightarrow^3$H+p , which is extinguished early as the free neutrons are used up. Therefore the abundance of ^{3}He does not fall the same way as ^{2}H and ^{3}H. The abundance of ^{7}Li also rises at first and then falls via ^{7}Li+p$\rightarrow^4$He+^{4}He. Since ^{4}He has a higher binding energy per nucleon, B/A, than ^{7}Li and ^{7}Be have, the nucleons in them also want to return into ^{4}He. This does not happen to ^{7}Be, however, since, just like for ^{3}He, the free neutrons needed for the reaction ^{7}Be+n$\rightarrow^4$He+^{4}He have almost disappeared near the end.

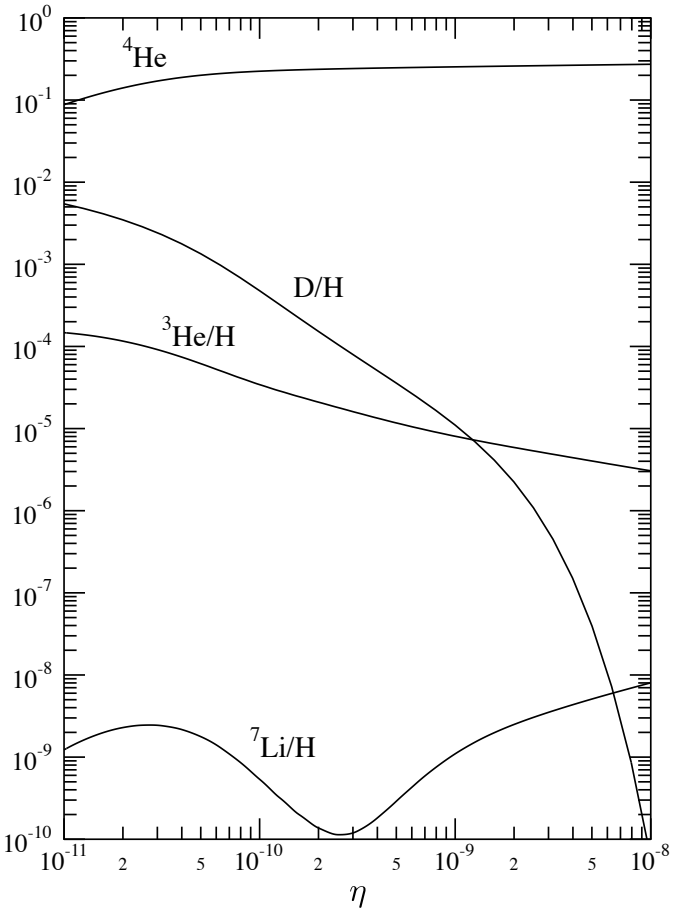


Figure 5: The primordial abundances of the light elements as a function of η. For ^{4}He we give the mass fraction, for D = ^{2}H, ^{3}He, and ^{7}Li the number ratio to H = ^{1}H, i.e., $n_i/n_{\rm H}$.

5.8 Primordial abundances as a function of the baryon-to-photon ratio

Let us then consider BBN as a function of η (see Fig. 5). The larger is η, the higher is the number density of nucleons. The reaction rates are faster and the nucleosynthesis can proceed further. This mean that a smaller fraction of "intermediate nuclei", ^{2}H, ^{3}H, and ^{7}Li are left over—the burning of nuclear matter into ^{4}He is "cleaner". Also the ^{3}He production falls with increasing η. However, ^{7}Be production increases with η. In the figure we have plotted the final BBN yields, so that ^{3}He is the sum of ^{3}He and ^{3}H, and ^{7}Li is the sum of ^{7}Li and ^{7}Be. The complicated shape of the ^{7}Li(η) curve is due to these two contributions: 1) For small η we get lots of "direct" ^{7}Li, whereas 2) for large η there is very little "direct" ^{7}Li left, but a lot of ^{7}Be is produced. In the middle, at $\eta \sim 3 \times 10^{-10}$, there is a minimum of ^{7}Li production where neither way is very effective.

The ^{4}He production increases with η, since with higher density nucleosynthesis begins earlier when there are more neutrons left.

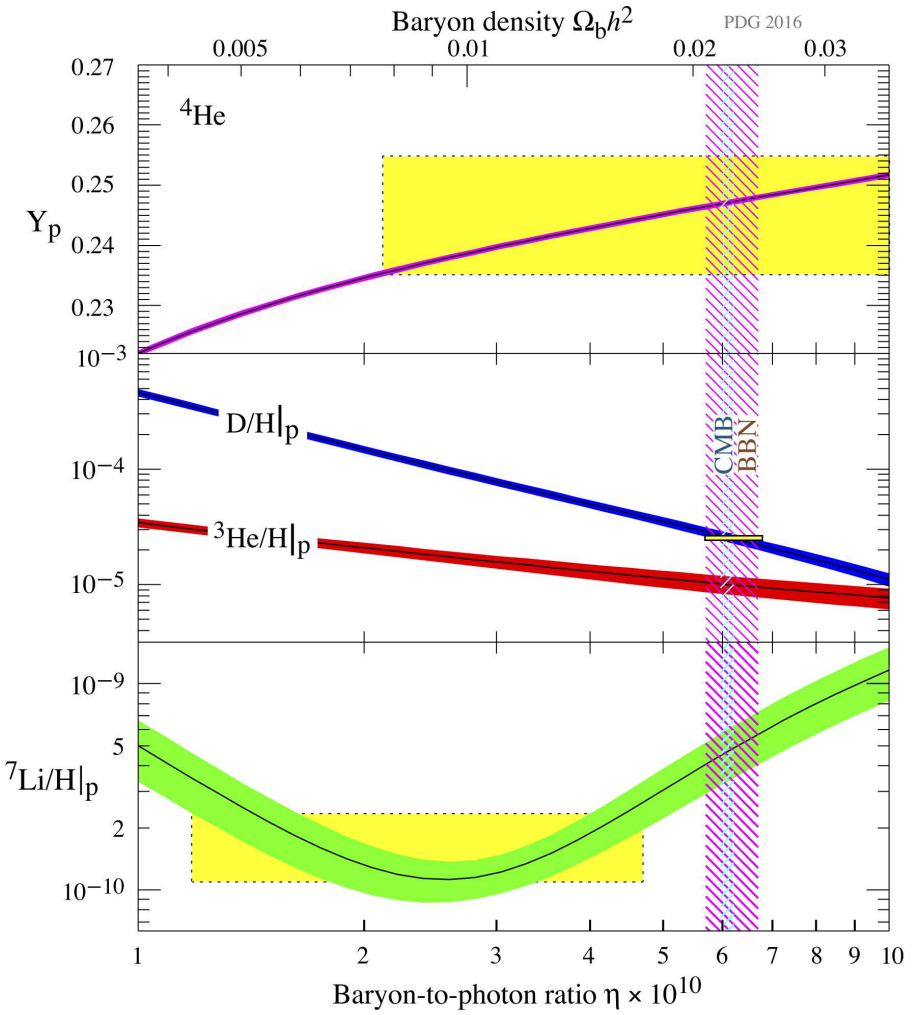


Figure 6: Determining the baryon/photon ration η by comparing BBN predictions to observations. The width of the bands around the the curves represents the uncertainty in BBN prediction due to uncertainty about the reaction rates. The vertical extent of the yellow boxes represent the estimate of the primordial abundance from observations and the horizontal extent the resulting range in η to agree with BBN. Note the small deuterium box. The only observational data on ^{3}He is from our own galaxy, and since ^{3}He is both produced and destroyed in chemical evolution, we cannot infer the primordial abundance from them. From the review by Fields, Molaro, and Sarkar in [1].

5.9 Comparison with observations

The abundances of the various isotopes calculated from BBN can be compared to the observed abundances of these elements. This is one of the most important tests of the big bang theory. A good agreement is obtained for η in the range $\eta = 5.8$–6.6×10^{-10}. This was the best method to estimate the amount of ordinary matter in the universe, until the advent of accurate CMB data, first from the WMAP satellite starting in 2003 and then from the Planck satellite data starting in 2013.[5]

The comparison of calculated abundances with observed abundances is complicated due to *chemical evolution*. The abundances produced in BBN are the *primordial* abundances of these isotopes. The first stars form with this composition. In stars, further fusion reactions take place and the composition of the star changes with time. Towards the end of its lifetime, the star ejects its outer parts into interstellar space, and this processed material mixes with primordial material. From this mixed material later generation stars form, and so on.

The observations of present abundances are based on spectra of interstellar clouds and stellar surfaces. To obtain the primordial abundances from the present abundances the effect of chemical evolution has to be estimated. Since ^{2}H is so fragile (its binding energy is so low), there is hardly any ^{2}H production in stars, rather any pre-existing ^{2}H is destroyed early on in stars. Therefore any interstellar ^{2}H is primordial. The smaller the fraction of processed material in an interstellar cloud, the higher its ^{2}H abundance should be. Thus all observed ^{2}H abundances are lower limits to the primordial ^{2}H abundance.[6] Conversely, stellar production increases the ^{4}He abundance. Thus all ^{4}He observations are upper limits to the primordial ^{4}He. Moreover, stellar processing produces heavier elements, e.g., C, N, O, which are not produced in the BBN. Their abundance varies a lot from place to place, giving a measure of how much chemical evolution has happened in various parts of the universe. Plotting ^{4}He vs. these heavier elements one can extrapolate the ^{4}He abundance to zero chemical evolution to obtain the primordial abundance. Since ^{3}He and ^{7}Li are both produced and destroyed in stellar processing, it is more difficult to make estimates of their primordial abundances based on observed present abundances.

Qualitatively, one can note two clear signatures of big bang in the present universe:

1. All stars and gas clouds observed contain at least 23% ^{4}He. If all ^{4}He had been produced in stars, we would see similar variations in the ^{4}He abundance as we see, e.g., for C, N, and O, with some regions containing just a few % or even less ^{4}He. This universal minimum amount of ^{4}He must signify a primordial abundance produced when matter in the universe was uniform.

2. The existence of significant amounts of ^{2}H in the universe is a sign of BBN, since there are no other known astrophysical sources of large amounts of ^{2}H.

Quantitatively, the observed abundances of all the BBN isotopes, ^{2}H, ^{3}He, ^{4}He and ^{7}Li point towards the range $\eta = 1.5$–7×10^{-10}. See Fig. 6. Since ^{2}H has the steepest dependence on η, it can determine η the most accurately. The best ^{2}H observations for this purpose are from the absorption spectra of distant (high-z) quasars. This absorption is due to gas clouds that lie on the line-of-sight between us and the quasar. Some of these clouds lie also at a high redshift. Thus we observe them as they were when the universe was rather young, and therefore little chemical evolution had yet taken place. These measurements point towards the higher end of

[5] Many cosmological parameters can be estimated from the CMB anisotropy, as will be discussed in Cosmology II. The Planck estimate[5] is $\omega_b = 0.0224 \pm 0.0001$, or $\eta = (6.14 \pm 0.03) \times 10^{-10}$.

[6] This does not apply to sites which have been enriched in ^{2}H due to a separation of ^{2}H from ^{1}H. Deuterium binds into molecules more easily than ordinary hydrogen. Since deuterium is heavier than ordinary hydrogen, deuterium and deuterated molecules have lower thermal velocities and do not escape from gravity as easily. Thus planets tend to have high deuterium-to-hydrogen ratios.

the above range, to $\eta = 5.8$–6.6×10^{-10}. Constraints from ^{3}He and ^{4}He are less accurate but consistent with this range. The estimates based on ^{7}Li abundances in the surfaces of a certain class of old Population II stars, which have been thought to retain the primordial abundance, give lower values $\eta = 1.5$–4.5×10^{-10}. This is known as the "Lithium problem". It is usually assumed that we do not understand well enough of the physics of the stars in question, and the range $\eta = 5.8$–6.6×10^{-10} (at 95% confidence level) is taken as the BBN value for the baryon-to-photon ratio.[1] (This is also consistent with the CMB results.)

The wider range $\eta = 1.5$–7×10^{-10} corresponds to $\omega_b = \Omega_b h^2 = 3.65 \times 10^7 \eta = 0.0055$–$0.026$. With $h = 0.7 \pm 0.07$, this gives $\Omega_b = 0.009$–0.07 for a conservative range of the baryonic density parameter. With $\eta = 5.8$–6.6×10^{-10} and $h = 0.7 \pm 0.07$, the BBN result for the baryonic density parameter is

$$\Omega_b = 0.036\text{–}0.061\,. \tag{23}$$

This is less than cosmological estimates for Ω_m, which are around 0.3. Therefore not all matter can be baryonic. In fact, most of the matter in the universe appears to be nonbaryonic dark matter. This is discussed in Chapter 6.

5.10 BBN as a probe of the early universe

BBN is the earliest event in the history of the universe from which we have quantitative evidence in the form of numbers (primordial abundances of ^{2}H, ^{3}He, ^{4}He, ^{7}Li) that we can calculate from known theory and compare to observations. It can be used to constrain many kinds of speculations about the early universe. For example, suppose there were additional species of particles that were relativistic at BBN time (this is called *dark radiation*). This would increase g_* and speed up the timescale (20), leading to more primordial ^{4}He and a higher primordial abundance of the intermediate isotopes ^{2}H and ^{3}He. Most such modifications of the standard picture will ruin the agreement between theory and observations. Thus we can say that we know well the history of the universe since the beginning of the BBN (from $T \sim 1\,\mathrm{MeV}$ and $t \sim 1\,\mathrm{s}$), but before that there is much more room for speculation.

References

[1] Particle Data Group, Chinese Physics C **40**, 100001 (2016)

[2] H. Karttunen, P. Kröger, H. Oja, M. Poutanen, and K.J. Donner: Fundamental Astronomy (Springer 1987)

[3] H. Kurki-Suonio, *Nukleosynteesi isotrooppisissa ja epäisotrooppisissa kosmologioissa*, Master's thesis, University of Helsinki (1983)

[4] Planck Collaboration, Astronomy & Astrophysics **594**, A13 (2016), arXiv:1502.01589

[5] Planck Collaboration, *Planck 2018 results. VI. Cosmological parameters*, arXiv:1807.06209v1 (2018)

6 Dark Matter

6.1 Observations

The earliest evidence for *dark matter* is due to Zwicky (1933). He observed (from the variation in their redshifts) that the relative velocities of galaxies in galaxy clusters were much larger than the escape velocity due to the mass of the cluster, if that mass was estimated from the amount of light emitted by the galaxies in the cluster. This suggested that there should actually be much more mass in the galaxy clusters than the luminous stars we can see. This was then called the "missing mass" problem. The modern terminology is to talk about dark matter, since it is understood that what is "missing" is not the mass, just the light from that mass.

Similar evidence comes from the *rotation curves* of galaxies. According to Kepler's third law, the velocity of a body orbiting a central mass is related to its distance as

$$v \propto \frac{1}{r^{1/2}} \,. \tag{1}$$

The planets in the Solar System satisfy this relation. For the stars orbiting the center of a galaxy the situation is different, since the mass inside the orbit increases with the distance. Suppose, for example, that the mass density of a galaxy decreases as a power-law

$$\rho \propto r^{-x} \tag{2}$$

with some constant x. Then the mass inside radius r is

$$M(r) \;\propto\; \int r^2 r^{-x} dr \;=\; \frac{r^{3-x}}{3-x} \qquad \text{for} \quad x < 3 \,. \tag{3}$$

Equating the acceleration of circular motion with that caused by Newtonian gravity we have

$$\frac{v^2}{r} \;=\; G\frac{M}{r^2} \;\propto\; r^{1-x} \,. \tag{4}$$

Thus we find that the rotation velocity in our model galaxy should vary with distance from the center as

$$v(r) \;\propto\; r^{1-x/2} \,. \tag{5}$$

The function $v(r)$ is called the rotation curve of a galaxy.

Observed rotation curves increase with r for small r, i.e., near the center of the galaxy, but then typically flatten out, becoming $v(r) \approx const$ up to as large distances as there is anything to observe in the galaxy. From Eq. (5), this would indicate a density profile

$$\rho \propto r^{-2} \,. \tag{6}$$

However, the density of stars appears to fall more rapidly towards the edges of the galaxy. Also, the total mass from stars and other visible objects, like gas and dust clouds, appears to be too small to account for the rotation velocity at large distances. This discrepancy between visible matter and galaxy rotation curves was established in early 1970s [1] after which this missing mass / dark matter problem became a central topic in astrophysics.

This indicates the presence of another mass component to galaxies. This mass component should have a different density profile than the visible, or luminous, mass in the galaxy, so that it could be subdominant in the inner parts of the galaxy, but would become dominant in the outer parts. This dark component appears to extend well beyond the visible parts of galaxies, forming a dark *halo* surrounding the galaxy.

This can be discussed in terms of mass-to-light ratios, M/L, of various objects. It is customarily given in units of $M_\odot/L_\odot$, where $M_\odot$ and $L_\odot$ are the mass and absolute luminosity for the Sun. The luminosity of a star increases with its mass faster than linearly, so that stars with $M > M_\odot$ have $M/L < 1$, and smaller stars have $M/L > 1$. Small stars are more common than large stars, so a typical mass-to-light ratio from the stellar component of galaxies is $M/L \sim$ a few. For stars in our part of the Milky Way galaxy, $M/L \approx 2.2$. Because large stars are more short-lived, M/L increases with the age of the star system, and the typical M/L from stars in the universe is somewhat larger. However, this still does not account for the full masses of galaxies.

The mass-to-light ratio of a galaxy turns out to be difficult to determine; the larger volume around the galaxy you include, the larger M/L you get. But the M is determined from velocities of orbiting bodies and at large distances there may be no such bodies visible. For galaxy clusters you can use the velocities of the galaxies themselves as they orbit the center of the cluster. The mass-to-light ratios of clusters appears to be several hundreds. Presumably isolated galaxies would have similar values if we could measure them to large enough radii.

From galaxy surveys, the luminosity density of the universe is

$$\rho_L = 2.0 \pm 0.7 \times 10^8 h L_\odot \text{ Mpc}^{-3} \,. \tag{7}$$

(Peacock [3], p.368; Efstathiou et al. 1988 [4]). Multiplying this with a typical mass-to-light ratio from galaxy clusters (Peacock, pp. 372–374),

$$M/L \sim 300h\text{–}400h \,, \tag{8}$$

we find an estimate for the density of clustered[1] mass in the universe,

$$\rho_m = (M/L) \cdot \rho_L \sim 0.39\text{–}1.08 \times 10^{11} h^2 M_\odot/\text{Mpc}^3 \tag{9}$$

$$= 2.6\text{–}7.3 \times 10^{-27} h^2 \text{kg/m}^3 \,. \tag{10}$$

(We equate the clustered mass with matter, since gravity causes mass, but not radiation or vacuum energy, to cluster. Implicitly we are assuming that all, or most of, matter clusters form stars, so that we can observe them.) Comparing to the critical density

$$\rho_{\text{cr}0} = h^2 \cdot 1.88 \times 10^{-26} \text{kg/m}^3 \,, \tag{11}$$

we get that

$$\Omega_m = 0.14\text{–}0.39 \,. \tag{12}$$

The estimates for the amount of ordinary matter in the objects we can see on the sky, stars and visible gas and dust clouds, called *luminous matter*, give a much smaller contribution,

$$\Omega_{\text{lum}} \lesssim 0.01 \tag{13}$$

to the density parameter. In Chapter 5 we found that big bang nucleosynthesis leads to an estimate

$$\Omega_b = 0.036\text{–}0.061 \tag{14}$$

for baryonic matter.

Thus we have

$$\Omega_{\text{lum}} < \Omega_b < \Omega_m. \tag{15}$$

[1] "Clustered" here does not refer to just galaxy clusters, but also to isolated galaxies, which are "clusters of matter".

This is consistent, since all luminous matter is baryonic, and all baryonic matter is matter. That we have two inequalities, instead of equalities, tells us that there are two kinds of *dark matter* (as opposed to luminous matter): 1) baryonic dark matter (BDM) and 2) nonbaryonic dark matter. We do not know the precise nature of either kind of dark matter, and therefore this is called the *dark matter problem*. To determine the nature of dark matter is one of the most important problems in astrophysics today. Often the expression "dark matter" is used to refer to the nonbaryonic kind only, as the nature of that is the deeper question.

6.2 Baryonic dark matter

The question of the dominant constituent of BDM is by now probably close to settled [2], so this section and its focus on MACHOs is mainly of historical interest.

Candidates for BDM include compact (e.g. planet-like) objects in interstellar space and thin intergalactic gas (or plasma).

Objects of the former kind have been dubbed MACHOs (Massive Astrophysical Compact Halo Objects) to contrast them with another (nonbaryonic) dark matter candidate, WIMPs, to be discussed later. A way to detect such a dark compact object is *gravitational microlensing*: If such a massive object passes near the line of sight between us and a distant star, its gravity focuses the light of that star towards us, and the star appears to brighten for a while. The brightening has a characteristic time profile, and is independent of wavelength, which clearly distinguishes it from other ways a star may brighten (variable stars).

An observation of a microlensing event gives an estimate of the mass, distance and velocity[2] of the compact object; but tells nothing else about it. Thus in principle we could have nonbaryonic MACHOs. But as we do not know of any such objects (except black holes), the MACHOs are usually thought of as ordinary substellar objects, such as *brown dwarfs* or "jupiters". Ordinary stars can of course also cause a microlensing event, but then we would also see this star. Here we are interested in events where we do not observe light, or any other signal, from the lensing object. Heavier relatively faint objects which could fall into this category, include old white dwarfs, neutron stars, and black holes, but these are expected to be much more rare.

The masses of ordinary black holes are included in the Ω_b estimate from BBN, since they were formed from baryonic matter after BBN. However, if there are *primordial black holes* produced in the big bang before BBN, they would not be included in Ω_b.

A star requires a mass of about $0.07 M_\odot$ to ignite thermonuclear fusion, and to start to shine as a star. Smaller, "failed", stars are called *brown dwarfs*. They are not completely dark; they are warm balls of gas which radiate faint thermal radiation. They were warmed up by the gravitational energy released in their compression to a compact object. Thus brown dwarfs can be, and have been, observed with telescopes if they are quite close by. Smaller[3] such objects are called "jupiters" after the representative such object in the Solar System.

The strategy to observe a microlensing event is to monitor constantly a large number of stars to catch such a brightening when it occurs for one of them. Since the typical time scales of these events are many days, or even months, it is enough to look at each star, say, once every night or so. As most of the dark matter is in the outer parts of the galaxy, further out than we are, it would be best if the stars to be monitored where outside of our galaxy. The Large Magellanic Cloud, a satellite galaxy of our own galaxy, is a good place to look for these events, being at a suitable distance where individual stars are still easy to distinguish. Because of the

[2] Actually we do not get an independent measure of all three quantities, as the observables depend on combinations of these. However, we can make some reasonable assumptions of the expected distance and velocity distributions among such objects, leading to a rough estimate of the mass. Especially from a set of many events, we get an estimate for the typical mass.

[3] That is, objects with smaller mass. Brown dwarfs actually all have roughly the same radius as Jupiter. The increased gravity from the larger mass compresses them to a higher density.

required precise alignment of us, the MACHO, and the distant star, the microlensing events will be rare. But if the BDM in our galaxy consisted mainly of MACHOs (with masses between that of Jupiter and several solar masses), and we monitored constantly millions of stars in the LMC, we should observe many events every year.

Such observing campaigns (MACHO, OGLE, EROS, ...) were begun in the 1990s. Indeed, over a dozen microlensing events towards LMC or SMC were observed. The typical mass of these MACHOs turned out to by $\sim 0.5 M_\odot$ (assuming the lenses were located in the halo of our galaxy), much larger than the brown dwarf mass that had been expected. The most natural faint object with such a mass would be a white dwarf. However, white dwarfs had been expected to be much too rare to explain the number of observed events. On the other hand the number of observed events is too small for these objects to dominate the mass of the BDM in the halo of our galaxy. These mass estimates depend on the assumed distance of the lenses. If one instead assumes that the lens is located in the same Magellanic Cloud as the star, the mass estimates are smaller, $\sim 0.2 M_\odot$. Then the lensing objects could be ordinary red dwarf stars, not visible to us due to this large distance. In any case, these observed lensing objects were too few to explain the BDM of our galaxy.

The present opinion is that the BDM in our universe is dominated by thin intergalactic ionized gas [2]. In fact, in large clusters of galaxies, we can see this gas, as it has been heated by the deep gravitational well of the cluster, and radiates X-rays.

6.3 Nonbaryonic dark matter

The favorite candidates for nonbaryonic dark matter are divided into two main classes, hot dark matter (HDM) and cold dark matter (CDM), based on the typical velocities of the particles making up this matter. These particles are supposed to be at most weakly interacting, so that they decoupled early, or possibly they were never at thermodynamic equilibrium.

The distinction between HDM and CDM comes from their different effect on *structure formation* in the universe. Structure formation refers to how the originally almost homogeneously distributed matter formed galaxies and galaxy clusters under the pull of gravity. For HDM, the velocities of the particles were large when structure formation began, making it difficult to trap them in potential wells of the forming structures. Typically these velocities were then nonrelativistic but larger than the escape velocities of the forming structures. CDM particles, on the other hand, have negligible velocities and they began to form structures early due to their mutual gravity. Structure formation dominated by HDM leads to top-down structure formation, where the largest structures form first, and smaller structures arise from the fragmentation of these larger structures. Structure formation dominated by CDM leads to bottom-up structure formation, where smaller structures form first, and later they cluster or coalesce to form larger structures. The intermediate case is called warm dark matter (WDM).

We shall discuss structure formation in Cosmology II. But we mention already that the observed *large-scale structure* in the universe, i.e., how galaxies are distributed in space, and relating it to the observed anisotropy of the CMB, which shows the primordial inhomogeneity, from which this structure grew, gives today the best way to estimate the relative amounts of BDM, HDM, and CDM in the universe. The result is that dark matter must be dominated by CDM (or possibly it could be somewhat "warm").

A popular class of nonbaryonic dark matter are *thermal relics*, particles and antiparticles that were initially in thermodynamic equilibrium, but decoupled early enough to prevent their annihilation with each other at least to some extent. For thermal relics there is another clear distinction between HDM and CDM: HDM particles decoupled while they were relativistic (the prime example is the neutrinos). They have therefore retained a large number density, and thus their masses must be small, less than 100 eV, for their total mass density not to exceed the

estimated dark matter density. Today, the HDM particles should be nonrelativistic—otherwise we would not classify them as "matter". CDM particles decouple while nonrelativistic and thus a much smaller relic number density is left over. Thus CDM particles must typically be heavy for CDM to form a significant part of dark matter. Since after decoupling the thermal relic CDM temperature falls as a^{-2}, CDM is extremely cold when structure formation begins.

However, there is another possibility for CDM: particles that were never in thermal equilibrium; these can have small velocities and still a large relic number density, requiring their masses to be small (the main such candidate is called the *axion*).

6.4 Hot dark matter

The main candidate for HDM are neutrinos with a small but nonzero rest mass. The cosmic neutrino background would make a significant contribution to the density parameter if the neutrinos had a rest mass of the order of 1 eV.

For massive neutrinos, the number density today is the same as for massless neutrinos, but their energy density today is dominated by their rest masses, giving (factor $\frac{3}{4}$ from their fermionic nature · factor $\frac{4}{11}$ from their lower temperature after electron-positron annihilation = $\frac{3}{11}$)

$$\rho_\nu \;=\; \sum_{\nu=1}^{3} m_\nu n_\nu \;=\; \frac{3}{11} n_\gamma \sum m_\nu \,. \tag{16}$$

For $T_0 = 2.725\,\mathrm{K}$, this gives for the neutrino density parameter

$$\Omega_\nu h^2 = \frac{\sum m_\nu}{94.14\,\mathrm{eV}}\,, \tag{17}$$

which applies if the neutrino masses are well below the neutrino decoupling temperature, $\sim 1\,\mathrm{MeV}$, but well above the present temperature of massless neutrinos, $T_{\nu 0} = 0.168\,\mathrm{meV}$. This counts then as one contribution to Ω_m. As discussed in Chapter 4, neutrino oscillation measurements constrain $\sum m_\nu$ within the range 0.06–6 eV, so that $\Omega_\nu = 0.001$–0.16.

If neutrinos dominated the masses of galaxies there would be a lower limit to their mass called the *Tremaine–Gunn limit* due to the available phase space inside the galaxy volume and below the galaxy escape velocity. This is because neutrinos are fermions and the Pauli exclusion principle prevents two neutrinos from occupying the same quantum state. A similar limit would apply to any fermion candidates for the dominant component of dark matter, but not to bosons.

Exercise: Tremaine–Gunn limit. Suppose neutrinos dominate the mass of galaxies (i.e., ignore other forms of matter). We know the mass of a galaxy (within a certain radius) from its rotation velocity. The mass could come from a smaller number of heavier neutrinos or a larger number of lighter neutrinos, but the available phase space (you don't have to assume a thermal distribution) limits the total number of neutrinos, whose velocity is below the escape velocity. This leads to a lower limit on the neutrino mass m_ν. Let r be the radius of the galaxy, and v its rotation velocity at this distance. Find the minimum m_ν needed for neutrinos to dominate the galaxy mass, assuming all three species have the same mass. (A rough estimate is enough: you can, e.g., assume that the neutrino distribution is spherically symmetric, and that the escape velocity within radius r equals the escape velocity at r). Give the numerical value for the case $v = 220\,\mathrm{km/s}$ and $r = 10\,\mathrm{kpc}$. Repeat the calculation assuming that only one of the three ν species is massive. (We know today that neutrinos are only a small part of dark matter, but a similar limit applies to any fermions.)

Data on large scale structure and CMB combined with structure formation theory requires that a majority of the matter in the universe has to be CDM (or possibly WDM) and the present upper limit to HDM (massive neutrinos) is [5]

$$\omega_\nu \equiv \Omega_\nu h^2 \lesssim 0.0025 \tag{18}$$

requiring that the sum of the three neutrino masses satisfies

$$\sum m_\nu \lesssim 230\,\mathrm{meV}\,. \tag{19}$$

Thus neutrinos make only a small contribution to dark matter,

$$0.0007 \lesssim \Omega_\nu h^2 \lesssim 0.0025\,, \tag{20}$$

where the lower limit comes from the $\sum m_\nu \geq 0.06$ from neutrino oscillations.

6.5 Cold dark matter

Observations require that dark matter is dominated by CDM. No known particle is suitable to act as CDM; therefore this conclusion implies that the standard model of particle physics must be extended with additional particles.

A major class of CDM particle candidates is called WIMPs (Weakly Interacting Massive Particles). We mentioned already that because of the large number density of neutrinos, their masses must be small, in order not to "close the universe" with an energy density $> \rho_{\rm cr}$. However, if the mass of some hypothetical weakly interacting particle species is much larger than the decoupling temperature of weak interactions, these particles will be largely annihilated before this decoupling, leading to a much lower number density, so that again it becomes possible to achieve a total density $< \rho_{\rm cr}$ starting from an initial thermal distribution at very high temperatures. (We calculate this in the next section.) Thus the universe may contain two classes of weakly interacting particles, very light (the neutrinos) and very heavy (the WIMPs), with a cosmologically interesting density parameter value.

The favorite kind of WIMP is provided by the supersymmetric partners of known particles, more specifically, the "lightest supersymmetric partner" (LSP), which could be a stable weakly interacting particle. It should have a mass of the order of 100 GeV. It has been hoped that, if it exists, it could be created and detected at CERN's LHC (Large Hadron Collider) particle accelerator. A measurement of its properties would allow a calculation of its expected number and energy density in the universe. So far (2018) there has been no detection, already considered a disappointment.

A candidate CDM particle should thus be quite heavy, if it was in thermal equilibrium sometime in the early universe. One CDM candidate, the *axion*, is, however, very light; but it was "born cold" and has never been in thermal interaction. It is related to the so-called "strong CP-problem" in particle physics. We shall not go into the details of this, but it can be phrased as the question "why is the neutron electric dipole moment so small?". It is zero to the accuracy of measurement, the upper limit being $d_n < 0.30 \times 10^{-25} e$cm (Particle Data Group 2016 [6]), whereas it has a significant magnetic dipole moment. A proposed solution involves an additional symmetry of particle physics (the Peccei–Quinn symmetry). The axion would then be the "Goldstone boson of the breaking of this symmetry". The important point for us is that these axions would be created in the early universe when the temperature fell below the QCD energy scale (of the order of 100 MeV), and they would be created "cold", i.e., with negligible kinetic energy, and they would never be in thermal interaction. Thus the axions have negligible velocities, and act like CDM.

If these WIMPs or axions make up the CDM, they should be everywhere, also in the Solar System, although they would be very difficult to detect. A *direct detection* is not impossible, however. Sensitive detectors have been built with this purpose. WIMPs and axions require a rather different detection technology.

One kind of an axion detector is a low noise microwave cavity in a strong magnetic field. An axion may interact with the magnetic field and produce a microwave photon. No axions have

so far been detected. On the other hand the detectors have so far not been sensitive enough for us to really expect a detection.

WIMPs interact weakly with ordinary matter. In practice this means that mostly they do not interact at all, so that a WIMP will pass through the Earth easily, without noticing it, but occasionally, very rarely, there will be an interaction. A typical interaction is elastic scattering from a nucleus, with an energy exchange of a few keV.[4] A very sensitive WIMP detector can detect if this much energy is deposited on its target material. The problem is that there are many other "background" events which may cause a similar signal. Thus these WIMP detectors are continuously detecting something.

Therefore the experimentalists are looking for an annual modulation in the signal they observe. The WIMPs should have a particular velocity distribution related to the gravitational well of our galaxy. The Earth is moving with respect to this velocity distribution, and the annual change in the direction of Earth's motion should result in a corresponding variation in the detection rate. One such experiment, DAMA,[5] has already claimed that they detect such an annual variation in the signal they observe, signifying that some of the events they see are due to WIMPs. Other experiments have not been able to confirm this detection.

6.6 Decoupling

An important class of dark matter particle candidates are *thermal relics*, particles that were once in thermal equilibrium and survived because they decoupled before they were annihilated.

Decoupling is the process where a particle species makes a transition from a high interaction rate with other particles to a low, and eventually negligible, interaction rate. While the interaction rate is high, the interactions keep the particles in thermal equilibrium with other species. When the interaction rate becomes low enough the particles decouple from other species. If the decoupled particles are stable (or have very long lifetime, i.e., negligible decay rate), their number will then stay constant so that their number density falls with the expansion as $n \propto a^{-3}$.

Consider the case where the main interaction of particle species x with other species (y, z) is particle-antiparticle annihilation and creation:

$$\mathrm{x} + \bar{\mathrm{x}} \;\leftrightarrow\; \mathrm{y} + \mathrm{z}\,. \tag{21}$$

For simplicity, assume an equal number of particles an antiparticles, $n_\mathrm{x} = n_{\bar{\mathrm{x}}}$, i.e., that $\mu_\mathrm{x} = 0$. (If $\mu_\mathrm{x} \neq 0$ but just very small, we can check after the calculation whether this was a good approximation, i.e., if the thermal relic density we get with the $\mu_\mathrm{x} = 0$ approximation is large compared to the particle-antiparticle excess. If μ_x is important, i.e., particles survive mainly because there were more particles than antiparticles, we wouldn't usually call them thermal relics.) The x particle number density n_x then evolves according to

$$\frac{dn_\mathrm{x}}{dt} + 3Hn_\mathrm{x} \;=\; -\langle\sigma v\rangle n_\mathrm{x} n_{\bar{\mathrm{x}}} + \psi\,, \tag{22}$$

where σ is the annihilation cross section (the effective area the other particle presents as a target), v is the relative velocity of the colliding particles, $\langle\cdot\rangle$ indicates the mean value taken over the particle phase space distribution, and ψ is the rate of creation of x particles.

In equilibrium, as many particles are created as annihilated. Thus

$$\psi \;=\; \langle\sigma v\rangle n_\mathrm{eq}^2\,, \tag{23}$$

[4] Note that weak interactions are "weak" in the sense that they occur rarely, but the energy exchange in such an interaction, when it occurs, does not have to be very small.

[5] http://www.lngs.infn.it/en/dama

where $n_{\rm eq}$ is the equilibrium value of $n_{\rm x}$ and $n_{\bar{\rm x}}$, and we can rewrite (22) as

$$\frac{dn_{\rm x}}{dt} + 3Hn_{\rm x} \;=\; -\langle\sigma v\rangle\left(n_{\rm x}^2 - n_{\rm eq}^2\right) . \tag{24}$$

The Hubble parameter H gives the time scale at which external conditions, and thus also $n_{\rm eq}$, change. Define

$$\Gamma \;\equiv\; n_{\rm eq}\langle\sigma v\rangle , \tag{25}$$

the reaction rate per particle in equilibrium ($\tau \equiv 1/\Gamma$ gives the mean time between interactions for an x particle). We have to compare Γ to H to determine whether the interaction rate is high or low. Defining the *comoving number density*

$$N_{\rm x} \equiv n_{\rm x}a^3 \qquad \text{and} \qquad N_{\rm eq} \equiv n_{\rm eq}a^3 \tag{26}$$

we can rewrite (24) as

$$\frac{1}{N_{\rm eq}}\frac{dN_{\rm x}}{d\ln a} \;=\; -\frac{\Gamma}{H}\left[\left(\frac{N_{\rm x}}{N_{\rm eq}}\right)^2 - 1\right] . \tag{27}$$

(It is often practical to use the logarithm of the scale factor $\ln a$ as time coordinate. It changes by one when the universe expands by a factor e.)

If $\Gamma \gg H$ ($\tau \ll H^{-1}$), the interactions keep $N_{\rm x}$ very close to $N_{\rm eq}$, since a small deviation is enough to make the rhs of (27) large and cause a rapid corrective change in $N_{\rm x}$. On the other hand, if $\Gamma \ll H$ ($\tau \gg H^{-1}$), the rhs stays negligible no matter how much $N_{\rm x}$ deviates from $N_{\rm eq}$ and thus $N_{\rm x}$ stays constant. Typically a particle species is in the $\Gamma \gg H$ regime at first, but may make a transition to the $\Gamma \ll H$ regime (decouple) later. We call the temperature T_d at which $\Gamma = H$, the *decoupling temperature*. (Decoupling is also called "freeze-out".)

The constant comoving number density after decoupling is the comoving *relic density* of the particles. A crude approximation (the *instantaneous decoupling approximation*) is

$$N_{\rm x}({\rm relic}) \approx N_{\rm x}(T_d) \approx N_{\rm eq}(T_d) , \tag{28}$$

which we get if we assume that $N_{\rm x}$ follows $N_{\rm eq}$ until $T = T_d$, and stays constant after that.

There are two distinct situations: 1) Hot thermal relics: particles that were ultrarelativistic ($T_d > m_{\rm x}$) when they decoupled. Their relic density is large, $\sim T^3$. 2) Cold thermal relics: particles that were nonrelativistic ($T_d \ll m_{\rm x}$) when they decoupled. Thus most of them annihilated after T fell below $m_{\rm x}$, but decoupling saved the rest. Thus cold thermal relics survive in much smaller numbers than hot thermal relics. We already discussed neutrinos, which are hot thermal relics, in Chapter 4. Let us now consider cold thermal relics.

Cold thermal relics decouple while they are nonrelativistic, so that their equilibrium number density then is

$$n_{\rm eq}(T_d) \;=\; g_{\rm x}\left(\frac{mT_d}{2\pi}\right)^{3/2} e^{-m/T_d} , \tag{29}$$

where m is the mass of the relic particle. After decoupling $n_{\rm x} \propto a^{-3}$, so that the present (relic) density is

$$n_{\rm x0} \;\approx\; \frac{g_{*S}(T_0)}{g_{*S}(T_d)}\left(\frac{T_0}{T_d}\right)^3 n_{\rm eq}(T_d) . \tag{30}$$

The problem is to find T_d, the decoupling temperature where $\Gamma_d \equiv n_{\rm eq}(T_d)\langle\sigma v\rangle = H$.

The annihilation cross section σ depends on the particle and associated theory, but on general quantum theoretical grounds σv can be expanded in terms of velocity squared, with contributions $\sigma v \propto v^{2q}$, where $q = 0$ is called s-wave annihilation, $q = 1$ p-wave annihilation etc. For the nonrelativistic case, $v \ll 1$, the s-wave annihilation is dominant, unless prohibited for

some reason in which case the p-wave is dominant. For an equilibrium distribution, the mean speed of a nonrelativistic particle is $\langle v\rangle = \sqrt{8/\pi}\sqrt{T/m}$. Since $v \propto (T/m)^{1/2}$, we can write

$$\langle\sigma v\rangle \;=\; \sigma_0 \left(\frac{T}{m}\right)^q , \qquad \text{where} \quad q = 0 \text{ (for s)} \quad \text{or} \quad q = 1 \text{ (for p)} . \tag{31}$$

Thus

$$\Gamma_d \;=\; \sigma_0 \left(\frac{T_d}{m}\right)^q n_{\rm eq}(T_d) \;=\; \sigma_0 \frac{g_{\rm x}}{(2\pi)^{3/2}} m^3 y^{-q-3/2} e^{-y} , \tag{32}$$

where

$$y \equiv \frac{m}{T_d} \gg 1 . \tag{33}$$

In the early universe, the relation between the Hubble parameter and temperature is

$$H^2 \;=\; \frac{8\pi G}{3}\frac{\pi^2}{30} g_*(T) T^4 \tag{34}$$

so that

$$H_d \;=\; \sqrt{\frac{g_*(T_d)}{90}}\frac{\pi m^2}{M_{\rm Pl}} y^{-2} \tag{35}$$

where $M_{\rm Pl} \equiv 1/\sqrt{8\pi G} = 2.436 \times 10^{18}$ GeV is the reduced Planck mass. The decoupling temperature can be solved from the equation

$$\frac{\Gamma_d}{H_d} \;=\; A y^{1/2-q} e^{-y} \;=\; 1 , \tag{36}$$

where

$$A \equiv \sqrt{\frac{45}{4\pi^5 g_*(T_d)}} g_{\rm x} M_{\rm Pl} m \sigma_0 . \tag{37}$$

Given $g_{\rm x}$, m, σ_0, q, and an initial guess for $g_*(T_d)$, we can solve T_d numerically from (36).

The relic number density is, from (30) and (29),

$$\begin{aligned} n_{\rm x0} \;&\approx\; \frac{g_{*S}(T_0)}{g_{*S}(T_d)} \left(\frac{T_0}{T_d}\right)^3 g_{\rm x} \left(\frac{m T_d}{2\pi}\right)^{3/2} e^{-y} \\ &=\; \frac{g_{\rm x}}{(2\pi)^{3/2}} \frac{g_{*S}(T_0)}{g_{*S}(T_d)} y^{3/2} e^{-y} T_0^3 \\ &=\; \frac{g_{\rm x}}{(2\pi)^{3/2}} \frac{g_{*S}(T_0)}{g_{*S}(T_d)} A^{-1} y^{1+q} T_0^3 \\ &=\; \sqrt{\frac{g_*(T_d)}{90}} \frac{g_{*S}(T_0)}{g_{*S}(T_d)} \frac{\pi}{M_{\rm Pl} m \sigma_0} y^{1+q} T_0^3 , \end{aligned} \tag{38}$$

where we used $e^{-y} = A^{-1} y^{q-1/2}$ from (36) to get rid of the exponential dependence on y (this allows us to use an approximate value for y below). We get the relic mass (energy) density by multiplying with m,

$$\rho_{\rm x0} \;=\; \sqrt{\frac{g_*(T_d)}{90}} \frac{g_{*S}(T_0)}{g_{*S}(T_d)} \frac{\pi}{M_{\rm Pl}\sigma_0} y^{1+q} T_0^3 . \tag{39}$$

Relating $n_{\rm x0}$ to the present CMB photon number density $n_{\gamma 0} = (2\zeta(3)/\pi^2) T_0^3$, we have

$$\frac{n_{\rm x0}}{n_{\gamma 0}} \;=\; \frac{\pi^3}{\zeta(3)} \sqrt{\frac{g_*(T_d)}{360}} \frac{g_{*S}(T_0)}{g_{*S}(T_d)} \frac{\pi}{M_{\rm Pl} m \sigma_0} y^{1+q} . \tag{40}$$

Assuming that decoupling happens before electron-positron annihilation and that no particle species is becoming nonrelativistic during the decoupling, we can set $g_*(T_d) = g_{*S}(T_d)$, and using the numerical value $g_{*S}(T_0) \approx 3.91$, this becomes

$$\frac{n_{\mathrm{x}0}}{n_{\gamma 0}} \;=\; \frac{\pi^3}{\zeta(3)} \frac{1}{\sqrt{360}} \frac{g_{*S}(T_0)}{\sqrt{g_*(T_d)}} \frac{\pi}{M_{\mathrm{Pl}} m \sigma_0} y^{1+q} \;\approx\; 5.31 \frac{y^{1+q}}{\sqrt{g_*(T_d)} M_{\mathrm{Pl}} m \sigma_0} \,. \tag{41}$$

For an analytical estimate of y, we can solve Eq. (36) iteratively: Taking the logarithm, it becomes

$$y \;=\; \ln A + (\tfrac{1}{2} - q) \ln y \,. \tag{42}$$

For $y \gg 1$, $\ln y$ is a slowly varying function of y, allowing for rapidly convergent iteration. We make our first guess for y by ignoring the term with $\ln y$:

$$y_0 = \ln A \tag{43}$$

and then iterate

$$y_1 \;=\; \ln A + (\tfrac{1}{2} - q) \ln y_0 \,. \tag{44}$$

For our rough estimate this first iteration is enough, and we take

$$y \;\approx\; y_1 \;=\; \ln A + (\tfrac{1}{2} - q) \ln(\ln A) \,. \tag{45}$$

where $A \propto m\sigma_0$, so that y depends logarithmically on m and σ_0.

The relic density depends mainly on m and σ_0. Assuming a fixed σ_0, we see that the relic number density decreases with increasing m, so that the cold thermal relic mass density $\rho_{\mathrm{x}0}$ depends only logarithmically on the relic particle mass m. However, as we see in the next section, σ_0 may depend on m, changing these conclusions.

6.7 WIMP miracle

Consider now a hypothetical particle with a mass in the GeV range, $g_{\mathrm{x}} = 2$ (spin-$\frac{1}{2}$ fermion), s-wave annihilation ($q = 0$) with a typical weak interaction cross section

$$\sigma_0 \sim G_F^2 E^2 \sim G_F^2 m^2 \tag{46}$$

where $G_F = 1.17 \times 10^{-5}\ \mathrm{GeV}^{-2}$ is the Fermi constant.

We have then $M_{\mathrm{Pl}} m \sigma_0 \approx 3.3 \times 10^8 (m/\,\mathrm{GeV})^3$. Since now $\sigma_0 \propto m^2$ and $m\sigma_0 \propto m^3$, we see that the relic densities depend on the mass as

$$n_{\mathrm{x}0} \propto \frac{1}{m^3} \qquad \text{and} \qquad \rho_{\mathrm{x}0} \propto \frac{1}{m^2} \tag{47}$$

(besides the logarithmic dependence via y). Thus the cold thermal relic mass density decreases with increasing relic particle mass, whereas for hot thermal relics, the number density is independent of m and the mass density increases proportional to m.

The decoupling temperature $T_d = m/y$ depends only logarithmically on $g_*(T_d)$, so the precise value of $g_*(T_d)$ is not important. If we assume that decoupling happened between the electroweak and QCD transitions, $g_*(T_d)$ is between 60 and 100 (in the standard model). Taking $g_*(T_d) = 60$, we get that $A \approx 1.63 \times 10^7 (m/\,\mathrm{GeV})^3$ and $\ln A \approx 16.6 + 3 \ln(m/\,\mathrm{GeV})$. The value of y is very close to this.

For example, for $m = 3$ GeV, $\ln A = 19.9$ and $y \approx \ln A + \frac{1}{2} \ln(\ln A) \approx 21.4$, so that $T_d = m/y \approx 0.14$ GeV. With a higher mass, we get a higher decoupling temperature. For $m = 100$ GeV, $\ln A = 30.4$, $y = 32.1$, and $T_d = 3.1$ GeV.

For the relic number density we get (approximating further $y \approx \ln A$)

$$\frac{n_{\rm x0}}{n_{\gamma 0}} \approx 5.31 \frac{y}{\sqrt{g_*(T_d)} M_{\rm Pl} m \sigma_0} = 2.1 \times 10^{-9} y \left(\frac{m}{\rm GeV}\right)^{-3}$$
$$\approx 3.5 \times 10^{-8} \left(1 + 0.18 \ln \frac{m}{\rm GeV}\right) \left(\frac{m}{\rm GeV}\right)^{-3} . \quad (48)$$

Taking the baryon-to-photon ratio to be $\eta = 6 \times 10^{-10}$, we get for the ratio of cold thermal relics to baryons

$$\frac{n_{\rm x0}}{n_{B0}} \approx 58 \left(1 + 0.18 \ln \frac{m}{\rm GeV}\right) \left(\frac{m}{\rm GeV}\right)^{-3} , \quad (49)$$

and since $m_N \approx 1$ GeV the mass density ratio

$$\frac{\rho_{\rm x0}}{\rho_{b0}} \approx 58 \left(1 + 0.18 \ln \frac{m}{\rm GeV}\right) \left(\frac{m}{\rm GeV}\right)^{-2} . \quad (50)$$

Since $\rho_{b0} \approx 0.05 \rho_{\rm cr0}$, the requirement that such relic particles (with energy density $\rho_{\rm x0} + \rho_{\bar{\rm x}0} = 2\rho_{\rm x0}$) do not close the universe gives the *lower bound* $m \gtrsim 2.6$ GeV for their mass (the *Lee-Weinberg bound*). The corresponding *upper bound* for a massive neutrino species (a hot thermal relic) was $m_\nu \lesssim 50$ eV (from Eq. (17) with $h \sim 0.7$).

We get the observed CDM to baryon density ratio $\rho_{c0}/\rho_{b0} \approx 5.3$ [5] for $m \approx 5.3$ GeV.[6] The fact that we get the right dark matter density for a cold thermal relic with a weak interaction cross section and mass roughly corresponding to the electroweak scale[7], is called the *WIMP miracle*. Such particles appear naturally in extensions to the standard model of particle physics, like supersymmetry (SUSY), which predicts SUSY partners to standard model particles, so this makes them a very natural candidate for dark matter. Such dark matter candidates are called WIMPs (weakly interacting massive particles).

In the minimal supersymmetric extension to the standard model (MSSM) such a particle with a mass of a few GeV would already have been discovered in particle colliders. The lower mass limit for fermionic SUSY partners in MSSM is 15 GeV [7]. Ongoing experiments at LHC are pushing this limit up. With a typical weak interaction σ_0 such heavier WIMPs would make just a small contribution to the dark matter. As the relic density is inversely proportional to σ_0 (besides the logarithmic dependence via y) we can save WIMPs as the main CDM candidate if we assume a smaller interaction cross section. There is enough freedom to adjust parameters in extensions to the standard model that this is possible. Such parameters are constrained by both collider and direct detection experiments.

6.8 Dark Matter vs. Modified Gravity

Since all the evidence for non-standard-model dark matter comes so far from its gravitational effects, it has been suggested that it does not exist, and instead the law of gravity needs to be modified over large distances. While actual proposals for such gravity modifications (MOND, TeVeS) do not appear very convincing and lead to difficulties of their own, it certainly would be comforting to have a direct laboratory detection of a CDM particle.

Evidence for the standard view of dark matter comes from collisions of clusters of galaxies[8], see Fig. 1. According to this standard view the mass of a cluster of galaxies has three main components: 1) the visible galaxies, 2) the intergalactic gas, and 3) cold dark matter. The last component should have the largest mass, and the first one the smallest. When two clusters of galaxies collide, it is unlikely for individual galaxies to collide, since most of the volume, and

[6] Note that ρ_c denotes the CDM energy density and $\rho_{\rm cr}$ the critical density.

[7] The electroweak scale is more like 100 GeV, but this is considered close enough.

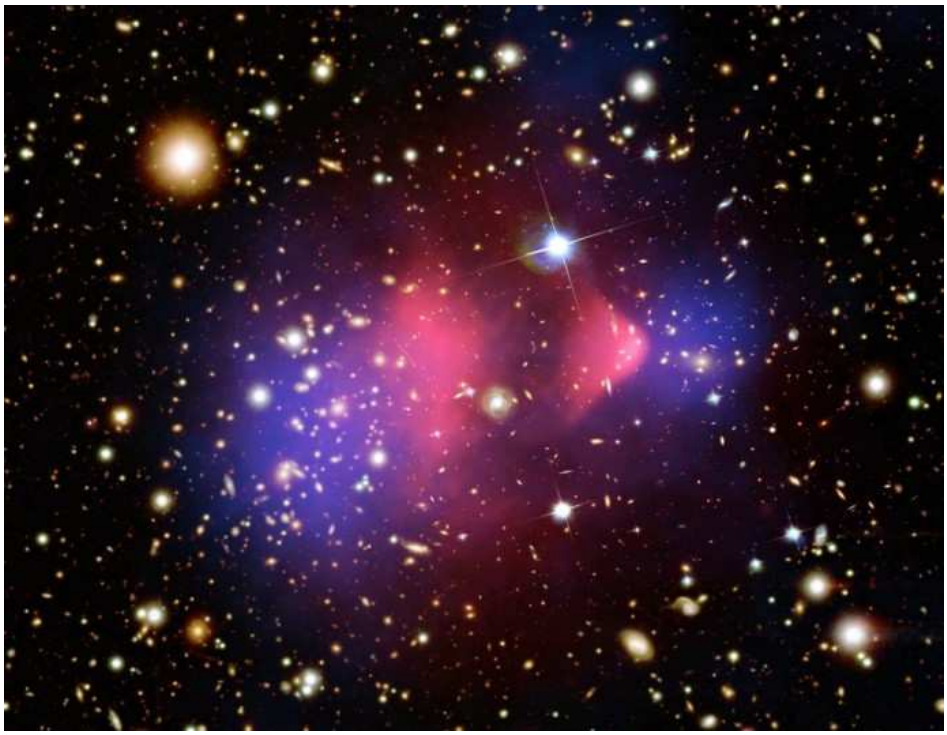


Figure 1: This is a composite image of galaxy cluster 1E 0657-56, also called the Bullet Cluster. It consists of two subclusters, a larger one on the left, and a smaller one on the right. They have recently collided and traveled through each other. One component of the image is an optical image which shows the visible galaxies. Superposed on it, in red, is an X-ray image, which shows the heated intergalactic gas, that has been slowed down by the collision and left behind the galaxy components of the clusters. The blue color is another superposed image, which represent an estimate of the total mass distribution of the cluster, based on gravitational lensing. NASA Astronomy Picture of the Day 2006 August 24. Composite Credit: X-Ray: NASA/CXC/CfA/M. Markevitch et al. Lensing map: NASA/STScI; ESO WFI; Magellan/U. Arizona/D. Clowe et al. Optical: NASA/STScI; Magellan/U. Arizona/D. Clowe et al.

cross section, in a cluster is intergalactic space. The intergalactic gas is too thin to slow down the relatively compact galaxies noticeably. On the other hand, the intergalactic gas components do not travel through each other freely but are slowed down by the collision and heated up. Thus after the clusters have traveled through each other, much of the intergalactic gas is left behind between the receding clusters. Cold dark matter, in turn, should be very weakly interacting, and thus practically collisionless. Thus the CDM components of both clusters should also travel through each other unimpeded.

Figure 1 is a composite image of such a collision of two clusters. We see that the intergalactic gas has been left behind the galaxies in the collision. The mass distribution of the system has been estimated from the gravitational lensing effect on the apparent shapes of galaxies behind the cluster. If there were no cold dark matter, most of the mass would be in the intergalactic gas, whose mass is estimated to be about five times that of the visible galaxies. Even in a modified gravity theory, we would expect most of the lensing effect to be where most of the mass is, even though the total mass estimate would be different. However, the image shows that most of the mass is where the galaxies are. This agrees with the cold dark matter hypothesis, since cold dark matter should move like the galaxies in the collision.

References

[1] K.C. Freeman, Astrophys. J. **160**, 811, Appendix A (1970); M.S. Roberts, A.H. Rots, Astronomy & Astrophysics **26**, 483 (1973); J.P. Ostriker, P.J.E. Peebles, A. Yahil, Astrophys. J. Lett. **193**, L1 (1974); J. Einasto, A. Kaasik, E. Saar, Nature **250**, 309 (1974); V.C. Rubin, W.K. Ford Jr, N. Thonnard, Astrophys. J. Lett. **225**, L107 (1978)

[2] F. Nicastro et al., *Observations of the missing baryons in the warm-hot intergalactic medium*, Nature **558**, 406 (2018)

[3] J.A. Peacock, *Cosmological Physics*, Cambridge University Press 1999

[4] G. Eftstahiou, R.S. Ellis, B.A. Petersin, MNRAS **232**, 431 (1988)

[5] Planck Collaboration, Astronomy & Astrophysics **594**, A13 (2016), arXiv:1502.01589

[6] Particle Data Group, Chinese Physics C **40**, 100001 (2016)

[7] G. Belanger et al., arXiv:1308.3735, Physics Letters B **726**, 773 (2013)

[8] D. Clowe et al., astro-ph/0608407, Astrophys. J. Lett. **648**, L109 (2006)

A More about General Relativity

A.1 Vectors, tensors, and the volume element

The *metric* of spacetime can always be written as

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \equiv \sum_{\mu=0}^{3} \sum_{\nu=0}^{3} g_{\mu\nu} dx^\mu dx^\nu. \tag{1}$$

We introduce Einstein's *summation rule*: there is a sum over repeated indices (that is, we don't bother to write down the summation sign $\sum$ in this case). Greek (spacetime) indices go over the values 0–3, Latin (space) indices over the values 1–3, i.e., $g_{ij} dx^i dx^j \equiv \sum_{i=1}^{3} \sum_{j=1}^{3} g_{ij} dx^i dx^j$. The objects $g_{\mu\nu}$ are the components of the *metric tensor*. They have, in principle, the dimension of distance squared. In practice one often assigns the dimension of distance (or time) to some coordinates, and then the corresponding components of the metric tensor are dimensionless. These *coordinate distances* are then converted to *proper* ("real" or "physical") distances with the metric tensor. The components of the metric tensor form a symmetric 4×4 matrix.

Example 1. The metric tensor for a 2-sphere (discussed in Chapter 2 as an example of a curved 2D space) has the components

$$[g_{ij}] = \begin{bmatrix} a^2 & 0 \\ 0 & a^2 \sin^2 \vartheta \end{bmatrix}. \tag{2}$$

Example 2. The metric tensor for Minkowski space has the components

$$g_{\mu\nu} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

in Cartesian coordinates, and

$$g_{\mu\nu} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2 \vartheta \end{bmatrix} \tag{4}$$

in spherical coordinates.

Example 3. The Robertson-Walker metric, which we discuss in Chapter 3, has components

$$g_{\mu\nu} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & \frac{a^2}{1-Kr^2} & 0 & 0 \\ 0 & 0 & a^2 r^2 & 0 \\ 0 & 0 & 0 & a^2 r^2 \sin^2 \vartheta \end{bmatrix}. \tag{5}$$

Note that the metric tensor components in the above examples always formed a diagonal matrix. This is the case when the coordinate system is orthogonal.

The vectors which occur naturally in relativity are *four-vectors*, with four components, e.g., the four-velocity. The values of the components depend on the basis $\{\mathbf{e}_\alpha\}$ used. Note that the index of the basis vector does not refer to a component, but specifies which one of the four basis vectors is in question. The components of the basis vectors in the basis they define are, of course,

$$(\mathbf{e}_\alpha)^\beta = \delta_\alpha^\beta \,, \tag{6}$$

where δ^β_α is the Kronecker symbol, 1 if $\alpha = \beta$, 0 otherwise.

Given a coordinate system, we have two bases (also called *frames*) naturally associated with it, the *coordinate basis* and the corresponding normalized basis. If the coordinate system is orthogonal, the latter is an *orthonormal basis*. When we use the coordinates to define the components of a vector, like the 4-velocity in Chapter 2, the components naturally come out in the coordinate basis. The basis vectors of a coordinate basis are parallel to coordinate lines, and their length represents the distance from changing the value of the coordinate by one unit. For example, if we move along the coordinate x^1 so that it changes by dx^1, the distance traveled is $ds = \sqrt{g_{11}dx^1dx^1} = \sqrt{g_{11}}dx^1$. The length of the basis vector $\mathbf{e}_1$ is thus $\sqrt{g_{11}}$. Since in the coordinate basis the basis vectors usually are not unit vectors, the numerical values of the components give the wrong impression of the magnitude of the vector. Therefore we may want to convert them to the normalized basis

$$\mathbf{e}_{\hat{\alpha}} \equiv \Big(\frac{1}{\sqrt{|g_{\alpha\alpha}|}}\Big)\mathbf{e}_\alpha \,. \tag{7}$$

(It is customary to denote the normalized basis with a hat over the index, when both bases are used. In the above equation there is no sum over the index α, since it appears only once on the left.) For a four-vector $\mathbf{w}$ we have

$$\mathbf{w} = w^\alpha \mathbf{e}_\alpha = w^{\hat{\alpha}} \mathbf{e}_{\hat{\alpha}} \,, \tag{8}$$

where

$$w^{\hat{\alpha}} \equiv \sqrt{|g_{\alpha\alpha}|}w^\alpha \,. \tag{9}$$

For example, the components of the coordinate velocity of a massive body, $v^i = dx^i/dt$ could be greater than one; the "physical velocity", i.e., the velocity measured by an observer who is at rest in the comoving coordinate system, is [1]

$$v^{\hat{i}} = \sqrt{g_{ii}}dx^i/\sqrt{|g_{00}|}dx^0 \,, \tag{10}$$

with components always smaller than one.

The volume of a region of space (given by some range in the spatial coordinates x^1, x^2, x^3) is given by

$$V = \int_V dV = \int_V \sqrt{\det[g_{ij}]}dx^1dx^2dx^3 \,, \tag{11}$$

where $dV \equiv \sqrt{\det[g_{ij}]}dx^1dx^2dx^3$ is the *volume element*. Here $\det[g_{ij}]$ is the determinant of the 3×3 submatrix of the metric tensor components corresponding to the spatial coordinates. For an orthogonal coordinate system, the volume element is

$$dV = \sqrt{g_{11}}dx^1\sqrt{g_{22}}dx^2\sqrt{g_{33}}dx^3. \tag{12}$$

The metric tensor is used for taking scalar (dot) products of four-vectors,

$$\mathbf{w} \cdot \mathbf{u} \equiv g_{\alpha\beta}u^\alpha w^\beta \,. \tag{13}$$

The (squared) *norm* of a four-vector $\mathbf{w}$ is

$$\mathbf{w} \cdot \mathbf{w} \equiv g_{\alpha\beta}w^\alpha w^\beta. \tag{14}$$

Exercise: Show that the norm of the four-velocity is always -1.

[1] When $g_{00} = -1$, this simplifies to $\sqrt{g_{ii}}dx^i/dt$.

For an *orthonormal* basis we have

$$\begin{aligned} \mathbf{e}_{\hat{0}} \cdot \mathbf{e}_{\hat{0}} &= -1 \\ \mathbf{e}_{\hat{0}} \cdot \mathbf{e}_{\hat{j}} &= 0 \\ \mathbf{e}_{\hat{i}} \cdot \mathbf{e}_{\hat{j}} &= \delta_{ij} \,. \end{aligned} \tag{15}$$

We shall use the short-hand notation

$$\mathbf{e}_{\hat{\alpha}} \cdot \mathbf{e}_{\hat{\beta}} = \eta_{\alpha\beta} \,, \tag{16}$$

where the symbol $\eta_{\alpha\beta}$ is like the Kronecker symbol $\delta_{\alpha\beta}$, except that $\eta_{00} = -1$.

A.2 Contravariant and covariant components

We sometimes write the index as a subscript, sometimes as a superscript. This has a precise meaning in relativity. The component w^α of a four-vector is called a *contravariant* component. We define the corresponding *covariant* component as

$$w_\alpha \equiv g_{\alpha\beta} w^\beta \,. \tag{17}$$

The norm is now simply

$$\mathbf{w} \cdot \mathbf{w} = w_\alpha w^\alpha. \tag{18}$$

In particular, for the 4-velocity we always have

$$u_\mu u^\mu = g_{\mu\nu} u^\mu u^\nu = \frac{ds^2}{d\tau^2} = -1 \,. \tag{19}$$

We defined the metric tensor through its covariant components (Eq. 1). We now define the corresponding covariant components $g^{\alpha\beta}$ as the inverse matrix of the matrix $[g_{\alpha\beta}]$,

$$g_{\alpha\beta} g^{\beta\gamma} = \delta^\gamma_\alpha \,. \tag{20}$$

Now

$$g^{\alpha\beta} w_\beta = g^{\alpha\beta} g_{\beta\gamma} w^\gamma = \delta^\alpha_\gamma w^\gamma = w^\alpha. \tag{21}$$

The metric tensor can be used to lower and raise indices. For tensors,

$$\begin{aligned} A_\alpha{}^\beta &= g_{\alpha\gamma} A^{\gamma\beta} \\ A_{\alpha\beta} &= g_{\alpha\gamma} g_{\beta\delta} A^{\gamma\delta} \\ A^{\alpha\beta} &= g^{\alpha\gamma} g^{\beta\delta} A_{\gamma\delta}. \end{aligned} \tag{22}$$

Note that for the *mixed components* $A_\alpha{}^\beta \neq A^\beta{}_\alpha$, unless the tensor is symmetric, in which case we can write A^β_α. When indices form covariant-contravariant pairs and are summed over, as in $A_{\alpha\beta\gamma} B^{\alpha\beta\gamma}$ the resulting quantity is invariant in coordinate transformations.

For an orthonormal basis,

$$g_{\hat{\alpha}\hat{\beta}} = g^{\hat{\alpha}\hat{\beta}} = \eta_{\alpha\beta}, \tag{23}$$

and the covariant and contravariant components of vectors and tensors have the same values, except that the raising or lowering of the time index 0 changes the sign. These orthonormal components are also called "physical" components, since they have the "right" magnitude.

Note that the symbols $\delta_{\alpha\beta}$ and $\eta_{\alpha\beta}$ are not tensors, and the location of their index carries no meaning.

A.3 Einstein equation

From the first and second partial derivatives of the metric tensor,

$$\partial g_{\mu\nu}/\partial x^\sigma, \qquad \partial^2 g_{\mu\nu}/(\partial x^\sigma \partial x^\tau)\,, \tag{24}$$

one can form various *curvature tensors*. These are the Riemann tensor $R^\mu{}_{\nu\rho\sigma}$, the Ricci tensor $R_{\mu\nu} \equiv R^\alpha{}_{\mu\alpha\nu}$, and the Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R$, where R is the Ricci scalar $g^{\alpha\beta} R_{\beta\alpha}$, also called the "scalar curvature" (not to be confused with the scale factor of the Robertson–Walker metric, which is sometimes denoted $R(t)$). We shall not discuss these curvature tensors in this course. The only purpose of mentioning them here is to be able to show the general form of the Einstein equation, before we go to the much simpler specific case of the Friedmann–Robertson–Walker universe.

In Newton's theory the source of gravity is mass, or, in the case of continuous matter, the mass density ρ. According to Newton, the gravitational field $\mathbf{g}_N$ is given by the equation

$$\nabla^2 \Phi = -\nabla \cdot \mathbf{g}_N = 4\pi G \rho\,. \tag{25}$$

Here Φ is the gravitational potential.

In Einstein's theory, the source of spacetime curvature is the *energy-momentum tensor*, also called the *stress-energy tensor*, or, for short, the "energy tensor" $T^{\mu\nu}$. The energy tensor carries the information on energy density, momentum density, pressure, and stress. The energy tensor of frictionless continuous matter (a *perfect fluid*) is

$$T^{\mu\nu} = (\rho + p) u^\mu u^\nu + p g^{\mu\nu}, \tag{26}$$

where ρ is the energy density and p is the pressure in the *rest frame* of the fluid. In cosmology we can usually assume that the energy tensor has the perfect fluid form. T^{00} is the energy density in the coordinate frame. (T^{i0} gives the momentum density, which is equal to the energy flux T^{0i}. T^{ij} gives the flux of momentum i-component in j-direction.)

We can now give the general form of the Einstein equation,

$$\boxed{G^{\mu\nu} = 8\pi G T^{\mu\nu}\,.} \tag{27}$$

This is the *law of gravity* according to Einstein. Comparing to Newton (Eq. 25) we see that the mass density ρ has been replaced by $T^{\mu\nu}$, and $\nabla^2\Phi$ has been replaced by the Einstein tensor $G^{\mu\nu}$, which is a short way of writing a complicated expression containing first and second derivatives of $g_{\mu\nu}$. Thus the gravitational potential is replaced by the 10 components of $g_{\mu\nu}$ in Einstein's theory.

In the case of a weak gravitational field, the metric is close to the Minkowski metric, and we can write, e.g.,

$$g_{00} = -1 - 2\Phi \tag{28}$$

(in suitable coordinates), where Φ is small. The Einstein equation for g_{00} becomes then

$$\nabla^2 \Phi = 4\pi G(\rho + 3p)\,. \tag{29}$$

Comparing this to Eq. (25) we see that the density ρ has been replaced by $\rho + 3p$. For relativistic matter, where p can be of the same order of magnitude than ρ this is an important modification to the law of gravity. For nonrelativistic matter, where the particle velocities are $v \ll 1$, we have $p \ll \rho$, and we get Newton's equation.

When applied to a homogeneous and isotropic universe filled with ordinary matter, the Einstein equation tells us that the universe cannot be static, it must either expand or contract.[2]

[2]Equation (44) leads to $\ddot{a} < 0$, which does not allow $a(t) = const$. If we momentarily had $\dot{a} = 0$, a would immediately begin to decrease.

When Einstein was developing his theory, he did not believe this was happening in reality. He believed the universe was static. Therefore he modified his equation by adding an extra term,

$$G^{\mu\nu} + \Lambda g^{\mu\nu} = 8\pi G T^{\mu\nu} \,. \tag{30}$$

The constant Λ is called the *cosmological constant*. Without Λ, a universe which was momentarily static, would begin to collapse under its own weight. A positive Λ acts as repulsive gravity. In Einstein's first model for the universe (the *Einstein universe*), Λ had precisely the value needed to perfectly balance the pull of ordinary gravity. This value is so small that we would not notice its effect in small scales, e.g., in the solar system. The Einstein universe is, in fact, unstable to small perturbations.[3] When Einstein heard that the Universe was expanding, he threw away the cosmological constant, calling it "the biggest blunder of my life".[4]

In more recent times the cosmological constant has made a comeback in the form of *vacuum energy*. Considerations in quantum field theory suggest that, due to vacuum fluctuations, the energy density of the vacuum should not be zero, but some constant $\rho_{\rm vac}$.[5] The energy tensor of the vacuum would then have the form $T_{\mu\nu} = -\rho_{\rm vac} g_{\mu\nu}$. Thus vacuum energy has exactly the same effect as a cosmological constant with the value

$$\Lambda = 8\pi G \rho_{\rm vac} . \tag{31}$$

Vacuum energy is observationally indistinguishable from a cosmological constant. This is because in physics, we can usually measure only energy differences. Only gravity responds to absolute energy density, and there a constant energy density has the same effect as the cosmological constant. In principle, however, they represent different ideas. The cosmological constant is an "addition to the left-hand side of the Einstein equation", a *modification of the law of gravity*, whereas vacuum energy is an "addition to the right-hand side", a contribution to the energy tensor, i.e., a form of energy.

A.4 Friedmann equations

We shall now apply the Einstein equation to the homogeneous and isotropic case, which leads to Friedmann–Robertson–Walker (FRW) cosmology. The metric is now the Robertson–Walker (RW) metric,

$$g_{\mu\nu} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & \frac{a^2}{1-Kr^2} & 0 & 0 \\ 0 & 0 & a^2 r^2 & 0 \\ 0 & 0 & 0 & a^2 r^2 \sin^2\vartheta \end{bmatrix} , \tag{32}$$

where K is a constant related to curvature of space and $a(t)$ is a function of time related to expansion of space. Calculating the Einstein tensor from this metric gives

$$G^{\hat{0}\hat{0}} = \frac{3}{a^2}(\dot{a}^2 + K) \tag{33}$$

$$G^{\hat{1}\hat{1}} = -\frac{1}{a^2}(2\ddot{a}a + \dot{a}^2 + K) = G^{\hat{2}\hat{2}} = G^{\hat{3}\hat{3}} . \tag{34}$$

[3] "If you sneeze, the universe will collapse."

[4] This statement does not appear in Einstein's writings, but is reported by Gamow[2].

[5] In field theory, the fundamental physical objects are fields, and particles are just quanta of the field oscillations. *Vacuum* means the ground state of the system, i.e., fields have those values which correspond to minimum energy. This minimum energy is usually assumed to be zero (although this is not necessary). However, in quantum field theory, the fields cannot stay at fixed values, because of quantum fluctuations. Thus even in the ground state the fields fluctuate around their zero-energy value, contributing a positive energy density. This is analogous to the zero-point energy of a harmonic oscillator in quantum mechanics.

We use here the orthonormal basis (signified by the $\hat{}$ over the index).

We assume the perfect fluid form for the energy tensor

$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu + pg^{\mu\nu}. \tag{35}$$

Isotropy implies that the fluid is at rest in the RW coordinates, so that $u^{\hat{\mu}} = (1, 0, 0, 0)$ and (remember, $g^{\hat{\mu}\hat{\nu}} = \eta^{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$)

$$T^{\hat{\mu}\hat{\nu}} = \begin{bmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{bmatrix}. \tag{36}$$

Homogeneity implies that $\rho = \rho(t)$, $p = p(t)$.

The Einstein equation $G^{\hat{\mu}\hat{\nu}} = 8\pi G T^{\hat{\mu}\hat{\nu}}$ becomes now

$$\frac{3}{a^2}(\dot{a}^2 + K) = 8\pi G\rho \tag{37}$$

$$-2\frac{\ddot{a}}{a} - \left(\frac{\dot{a}}{a}\right)^2 - \frac{K}{a^2} = 8\pi G p\,. \tag{38}$$

Let us rearrange this pair of equations to[6]

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{K}{a^2} = \frac{8\pi G}{3}\rho \tag{43}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p)\,. \tag{44}$$

These are the *Friedmann equations*. ("Friedmann equation" in singular refers to Eq. 43.)

References

[1] C.W. Misner, K.S. Thorne, J.A. Wheeler, Gravitation (Freeman 1973)

[2] G. Gamow, My Worldline (Viking Press 1970), p. 44, cited on
`https://blogs.scientificamerican.com/guest-blog/einsteins-greatest-blunder/`

[6] Including the cosmological constant Λ these equations take the form

$$\frac{3}{a^2}(\dot{a}^2 + K) - \Lambda = 8\pi G\rho \tag{39}$$

$$-2\frac{\ddot{a}}{a} - \left(\frac{\dot{a}}{a}\right)^2 - \frac{K}{a^2} + \Lambda = 8\pi G p\,. \tag{40}$$

or, in the rearranged form,

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{K}{a^2} - \frac{\Lambda}{3} = \frac{8\pi G}{3}\rho \tag{41}$$

$$\frac{\ddot{a}}{a} - \frac{\Lambda}{3} = -\frac{4\pi G}{3}(\rho + 3p)\,. \tag{42}$$

We shall not include Λ in these equations. Instead, we allow for the presence of vacuum energy ρ_{vac}, which has the same effect.

Cosmology II

Hannu Kurki-Suonio

Fall 2018

Preface

In Cosmology I we discussed the universe in terms of a homogenous and isotropic approximation to it. In Cosmology II we add the inhomogeneity and anisotropy. The mathematical background required includes Fourier analysis (taught in Fysiikan matemaattiset menetelmät I) and spherical harmonic analysis (taught in Fysiikan matemaattiset menetelmät II). We will take some results from Quantum Field Theory and Cosmological Perturbation Theory, but students are not expected to have them as background – they are more advanced courses. We begin with Inflation, but postpone the discussion of generation of perturbations during it to after we have discussed inhomogeneity in general and its later evolution – the chapter on Structure Formation. Thus in the Inflation chapter we still assume the homogeneous FRW model. We end with the Cosmic Microwave Background Anisotropy, which forms an important part of observational data in cosmology.

7 Inflation

7.1 Motivation

In Cosmology I we discussed how the universe began with a Hot Big Bang. This leaves open the question of initial conditions – how did the Hot Big Bang begin, and why did it begin with such a state of high density and temperature, rapid expansion, and a high level of isotropy and homogeneity. Inflation is a *scenario* to address this question, at least to some extent. Inflation is a period in the very early universe, before the events discussed in Cosmology I, when the expansion of the universe was *accelerating*.

Inflation is not really a specific theory; rather it is a more general idea of a certain kind of behavior (i.e., a "scenario") for the universe. It is not known for sure whether inflation occurred, but it makes a number of predictions that agree with observations. Inflation has been more successful than competing ideas for the very early universe and it has become part of the standard model for cosmology. The most important property of inflation is that it provides a mechanism for generating the initial density fluctuations, the *primordial perturbations*, from which the structure of the universe, stars and galaxies, grew. However, this property was discovered later, and the original motivation for inflation was to explain the initial flatness and homogeneity of the universe and the lack of certain relics that could have been produced at the very high temperatures of the very early universe[1][1]. This chapter discusses inflation in the homogeneous and isotropic approximation. Perturbations are discussed in Chapter 8.

Much of this chapter follows Chapter 3 of the book by Liddle&Lyth.[2]

7.1.1 Flatness problem

The Friedmann equation can be written as

$$\Omega - 1 = \frac{K}{a^2 H^2}\,. \tag{1}$$

If the universe has the critical density, $\Omega = 1$, it stays that way. But if $\Omega \neq 1$, it evolves in time. The difference $|\Omega - 1|$ grows with time during both the radiation-dominated and matter-dominated epochs. If the difference $\Omega - 1$ is small, its time evolution takes the form

$$\text{mat.dom} \qquad a \propto t^{2/3},\ H \propto t^{-1} \Rightarrow \frac{1}{aH} \propto t^{1/3} \qquad \Rightarrow |1-\Omega| \propto t^{2/3} \tag{2}$$

$$\text{rad.dom} \qquad a \propto t^{1/2},\ H \propto t^{-1} \Rightarrow \frac{1}{aH} \propto t^{1/2} \qquad \Rightarrow |1-\Omega| \propto t\,. \tag{3}$$

Since today, and at the end of the matter-dominated epoch, $\Omega_0 = \mathcal{O}(1)$ – and it is not essential here that Ω_0 is very close to 1, it would be enough that, say, $0.1 < \Omega_0 < 10$ – we can calculate backwards in time to, e.g., Big Bang Nucleosynthesis (BBN) and we find that the density parameter must have been extremely close to 1 then:

$$|\Omega(t_{\rm BBN}) - 1| \;=\; |\Omega_k(t_{\rm BBN})| \;\lesssim\; 10^{-16} \tag{4}$$

Thus we get as an initial condition to Big Bang, that Ω must have been initially extremely close to 1. The flatness problem is to explain why it was so. Otherwise, if we start the FRW universe in a radiation-dominated state with some initial value of the density parameter Ω_i not extremely close to 1, one of two things happens:

- $\Omega_i > 1 \quad \Rightarrow$ the universe recollapses almost immediately

[1] Guth[1] was not the first to propose a period of accelerating expansion in the early universe, but it was his proposal that became widely known and made inflation popular.

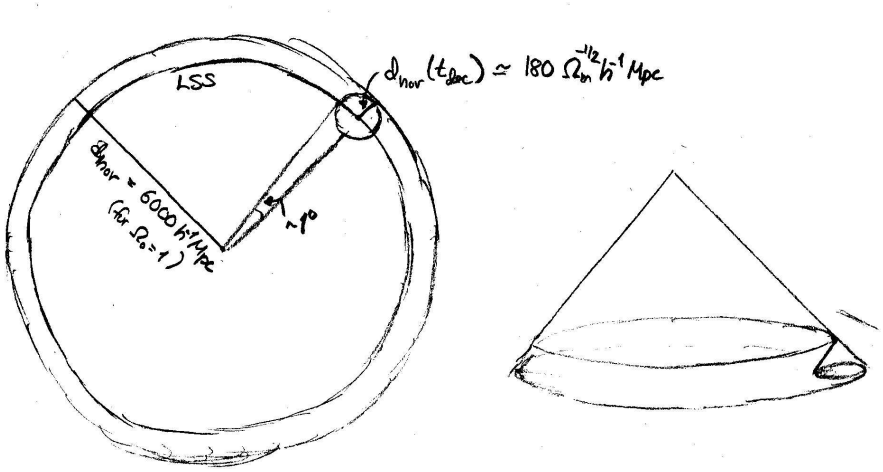


Figure 1: The horizon problem: regions on the CMB sky separated by more than about 1° had not had time to interact, yet their temperature is the same with an accuracy of $\lesssim 10^{-4}$.

- $\Omega_i < 1 \quad \Rightarrow$ the universe expands very fast and cools to $T < 3\,\mathrm{K}$ in a very short time.

Thus the flatness problem can also be called the oldness problem: why did it take so long, 14 billion years, for the universe to cool to $T = 2.7\,\mathrm{K}$.

Exercise: Oldness problem. Assume $\Omega_k(T = 100\,\mathrm{keV}) = 0.1$. Include just curvature and radiation (with $g_* = 3.36$) in the Friedmann equation. How long does it take for the universe to cool to $T = 2.7\,\mathrm{K}$? Why would inclusion of matter (with, say $\eta = 6 \times 10^{-10}$, and $\rho_m = 6\rho_b$) not change the answer?

7.1.2 Horizon problem

The horizon problem can also be called the homogeneity problem. The cosmic microwave background (CMB), which shows the universe at $z = 1090$ (age 380 000 years), is remarkably isotropic, the relative temperature variations being only $\mathcal{O}(10^{-4})$. This implies that density variations at that time must have been also very small, so the early universe was very homogeneous. Calculated according to the standard Hot Big Bang model, the horizon distance at that time was much smaller than the part of the early universe we see in the CMB, corresponding to only about 1° on the sky. Thus there could not have been any process to homogenize conditions over scales larger than this. This implies that this level of homogeneity must have been an initial condition.

Even the small CMB anisotropies show correlations at larger scales that 1°, a fact discovered after inflation was proposed.

7.1.3 Unwanted relics

If the Hot Big Bang begins at very high T it may produce objects surviving to the present, that are ruled out by observations.

- **Gravitino.** The supersymmetric partner of the graviton. $m \sim 100\,\mathrm{GeV}$. They interact very weakly (gravitational strength) $\Rightarrow$ they decay late, after BBN, and ruin the success of BBN.

- **Magnetic monopoles.** If the symmetry of a Grand Unified theory (GUT) is broken in a spontaneous symmetry breaking phase transition, magnetic monopoles are produced. These are point-like *topological defects* that are stable and very massive, $m \sim T_{\rm GUT} \sim 10^{14}\,$GeV. Their expected number density is such that their contribution to the energy density today $\gg$ the critical density.

- **Other topological defects** (cosmic strings, domain walls). These may also be produced in a GUT phase transition, and may also be a problem, but this is model-dependent. On the other hand, *cosmic strings* had been suggested as a possible explanation for the initial density perturbations—but this scenario fell later in trouble with the observational data (especially the anisotropy of the CMB).

These relics are produced very early, at extremely high temperatures, typically $T \gtrsim 10^{14}\,$GeV. From BBN, we only know that we should have standard Hot Big Bang for $T \lesssim 1\,$MeV.

7.1.4 What is needed

The word "problem" in the preceding is not to be taken to imply that the Hot Big Bang theory for the early universe would be in trouble. The theory by itself just does not contain answers to some questions one may pose about its initial conditions, for which we thus need additional ideas. We are perfectly happy if we can produce as an "initial condition" for Big Bang a universe with temperature $1\,{\rm MeV} < T < 10^{14}\,$GeV, which is almost homogeneous and has $\Omega = 1$ with extremely high precision.

7.2 Inflation introduced

7.2.1 Accelerated expansion

Inflation is not a replacement for the Hot Big Bang, but an addition to it, occurring at very early times (e.g., $t \sim 10^{-35}$ s), without disturbing any of its successes. Thus we have first inflation, then Hot Big Bang; so that inflation produces the initial conditions for the Hot Big Bang.

The origin of the flatness problem is that $|\Omega - 1| = |K|/(aH)^2$ grows with time. Now

$$\frac{d}{dt}|\Omega - 1| \;=\; |K|\frac{d}{dt}\left(\frac{1}{a^2H^2}\right) \;=\; |K|\frac{d}{dt}\left(\frac{1}{\dot{a}^2}\right) \;=\; \frac{-2|K|}{\dot{a}^3}\ddot{a}\,. \tag{5}$$

For an expanding universe, $aH = a\left(\frac{\dot{a}}{a}\right) = \dot{a} > 0$. Thus $\dot{a}^3 > 0$, and

$$\frac{d}{dt}|\Omega - 1| > 0 \qquad \Leftrightarrow \qquad \ddot{a} < 0\,. \tag{6}$$

Thus the reason for the flatness problem is that the expansion of the universe is decelerating, i.e., slowing down. If we had an early period in the history of the universe, where the expansion was accelerating, it could make an initially arbitrary value of $|\Omega - 1| = |K|/(aH)^2$ very small.

Definition: Inflation = any epoch when the expansion is accelerating.

$$\text{Inflation} \;\Leftrightarrow\; \ddot{a} > 0 \tag{7}$$

Consider then the horizon problem. The horizon at photon decoupling, $d^p_{\rm hor}(t_{\rm dec})$ is somewhere between the radiation-dominated and matter-dominated values, H^{-1} and $2H^{-1}$. For comparing sizes of regions at different times, we should use there comoving sizes, $d^c \equiv d/a$. We have

$$d^c_{\rm hor}(t_{\rm dec}) \sim \frac{1}{a_{\rm dec}H_{\rm dec}}\,, \tag{8}$$

whereas the size of the observable universe today is of the order of the present Hubble length

$$d^c_{\rm hor}(t_0) \sim H_0^{-1}\,. \tag{9}$$

The horizon problem arises because the first is much smaller than the second,

$$\frac{d^c_{\rm hor}(t_{\rm dec})}{d^c_{\rm hor}(t_0)} \sim \frac{a_0 H_0}{a_{\rm dec} H_{\rm dec}} \ll 1\,. \tag{10}$$

Thus the problem is that aH, whose inverse gives roughly the comoving size of the horizon, decreases with time,

$$\frac{d}{dt}(aH) = \frac{d}{dt}(\dot a) = \ddot a < 0\,. \tag{11}$$

Having a period with $\ddot a > 0$ could solve the problem.

In the preceding we referred to the (comoving) horizon distance at some time t, defined as the comoving distance light has traveled from the beginning of the universe until time t. If there are no surprises at early times, we can calculate or estimate it; like in the preceding where we assumed radiation-dominated or matter-dominated behavior (standard Big Bang). If we now start adding other periods, like accelerating expansion at early times, the calculation of $d_{\rm hor}$ will depend on them. In principle, $d^c_{\rm hor}(t_0) > d^c_{\rm hor}(t_{\rm dec})$ always, since $t_0 > t_{\rm dec}$, so $(0, t_{\rm dec}) \subset (0, t_0)$. But note that in the horizon problem, the relevant present horizon is how far we can *see*: the observable universe is given just by the integrated comoving distance the photon has traveled in the interval $(t_{\rm dec}, t_0)$, which is not affected by what happens before $t_{\rm dec}$. Thus the relevant present horizon is still $\sim H_0^{-1}$.

What is the relation between $d^c_{\rm hor}(t)$ and $1/(aH)$ for arbitrary expansion laws? Introduce the comoving, or conformal, Hubble parameter,

$$\mathcal{H} \equiv aH = \frac{1}{a}\frac{da}{d\eta} \equiv \dot a\,, \tag{12}$$

where η is the conformal time, defined by $d\eta = dt/a$. The Hubble length is

$$l_H \equiv H^{-1}\,, \qquad \text{where } H \equiv \frac{\dot a}{a}\,, \tag{13}$$

and the *comoving Hubble length* is

$$l^c_H \equiv \frac{l_H}{a} = \frac{1}{aH} = \frac{1}{\dot a} = \mathcal{H}^{-1}\,. \tag{14}$$

Roughly speaking, $\mathcal{H}^{-1}$ gives the comoving distance light travels in a "cosmological timescale", i.e., the Hubble time. This statement cannot be exact, since both the comoving Hubble length and the Hubble time change with time. However, if $\mathcal{H}^{-1}$ is increasing with time, the comoving distances traveled at earlier "epochs" are shorter, and thus $\mathcal{H}^{-1}(t)$ is a good estimate for the total comoving distance light has traveled since the beginning of time (the horizon). On the other hand, if $\mathcal{H}^{-1}$ is shrinking, then at earlier epochs light was traveling longer comoving distances, and we expect the horizon at time t to be larger than $\mathcal{H}^{-1}$. In any case

$$d^c_{\rm hor}(t) \gtrsim \mathcal{H}(t)^{-1}\,. \tag{15}$$

Since the Hubble length is more easily "accessible" (less information needed to figure it out) than the horizon distance it has become customary in cosmology to use the word "horizon" also for the Hubble distance. We shall also adopt this practice. The Hubble length gives the distance over which we have causal interaction in cosmological timescales. The comoving Hubble length gives this distance in comoving units.

If aH is decreasing (Eq. 11) then $\mathcal{H}^{-1}$ increases, and vice versa.

$\therefore$ Inflation = any epoch when the comoving Hubble length is shrinking.

$$\text{Inflation} \;\Leftrightarrow\; \frac{d}{dt}\mathcal{H}^{-1} < 0 \tag{16}$$

Thus the comoving distance over which we have causal connection is *decreasing* during inflation: causal contact to other parts of the Universe is being lost.

Inflation can be discussed either 1) in terms of physical distances or 2) in terms of comoving distances.

1) In terms of physical distance, the distance between any two points in the Universe is increasing, with an accelerating rate. The distance over which causal connection can be maintained is increasing (much) more slowly.

2) In terms of comoving distance, i.e., viewed in comoving coordinates, the distance between two points stays fixed; regions of the universe corresponding to present structures maintain fixed size. From this viewpoint (the one normally adopted), the region causally connected to a given location in the Universe is shrinking.

To connect with dynamics, look at the second Friedmann equation,

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p) \tag{17}$$

$$\therefore \text{Inflation} \;\Leftrightarrow\; \rho + 3p < 0 \;\Leftrightarrow\; w < -\tfrac{1}{3} \tag{18}$$

Thus inflation requires negative pressure, $p < -\frac{1}{3}\rho$ (we assume $\rho \geq 0$).

There is a huge class of models to realize the inflation scenario. These models rely on so-far-unknown physics of very high energies. Some models are just "toy models", with a hoped-for resemblance to the actual physics of the early universe. Others are connected to proposed extensions (like supersymmetry) to the standard model of particle physics.

The important point is that inflation makes many *generic*[2] predictions, i.e., predictions that are independent of the particular model of inflation. Present observational data agrees with these predictions. Thus it is widely believed—or considered probable—by cosmologists that inflation indeed took place in the very early universe. There are also numerical predictions of cosmological observables that differ from one model of inflation to another, allowing future observations to rule out classes of such models. (Many inflation models are already ruled out.)

7.2.2 Solving the problems

Inflation can solve[3] all the problems discussed in Sec. 7.1. The idea is that during inflation the universe expands by a large factor (at least by a linear factor of something like $\sim e^{70} \sim 10^{30}$ to solve the problems). This cools the universe to $T \sim 0$ (if the concept of temperature is applicable). When inflation ends, the universe is heated to a high temperature, and the usual Hot Big Bang history follows. This heating at the end of inflation is called *reheating*, since originally the thinking was that inflation started at an earlier hot epoch, but it is actually not clear whether that was the case.

[2] I was once in a conference where a speaker began his talk on inflation by promising not to use the words "generic" or "scenario". He failed in one but not the other.

[3] This is not to be taken too rigorously. The problems are related to the question of initial conditions of the universe at some very early time, whose physics we do not understand. Thus theorists are free to have different views on what kind of initial conditions are "natural". Inflation makes the flatness and horizon problems "exponentially smaller" in some sense, but inflation still places requirements—on the level of homogeneity—for the initial conditions *before* inflation, so that inflation can begin.

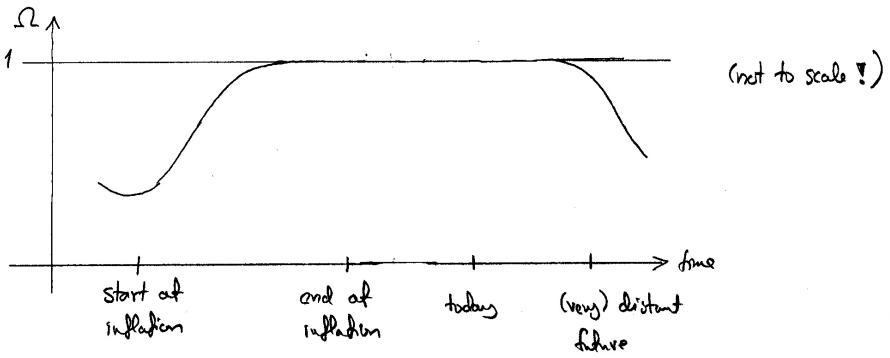


Figure 2: Solving the flatness problem. This figure is for a universe with no dark energy, where the expansion keeps decelerating after inflation ended in the early universe. Present observational evidence indicates that actually the expansion began accelerating again (supposedly due to the mysterious dark energy) a few billion years ago. Thus the universe is, technically speaking, inflating again, and Ω is again being driven towards 1. However, this current epoch of inflation is not enough to solve the flatness problem, or the other problems, since the universe has only expanded by about a factor of 2 during it.

Solving the flatness problem: The flatness problem is solved, since during inflation

$$|1-\Omega| = \frac{|K|}{\mathcal{H}^2} \quad \text{is shrinking.} \tag{19}$$

Thus inflation drives $\Omega \to 1$. Starting with an arbitrary Ω, inflation drives $|1-\Omega|$ so small that, although it has grown all the time from the end of inflation to the recent onset of dark energy domination, it is still very small today. See Fig. 2. In fact, inflation predicts that $\Omega_0 = 1$ to high accuracy, since it would be an unnatural coincidence for inflation to last just the right amount so that Ω would begin to deviate from 1 just at the current epoch.[4]

Solving the horizon problem: The horizon problem is solved, since during inflation the causally connected region is shrinking. It was very large before inflation; much larger than the present horizon. Thus the present observable universe has evolved from a small patch of a much larger causally connected region; and it is natural that the conditions were (or became) homogeneous in that patch then. See Fig. 3.

Getting rid of relics: If unwanted relics are produced before inflation, they are diluted to practically zero density by the huge expansion during inflation. We just have to take care they are not produced after inflation, i.e., the reheating temperature has to be low enough. This is an important constraint on models of inflation.

Did we really solve the problem of initial conditions? Actually solving the flatness and horizon problems is more complicated. We discussed them in terms of a FRW universe, which by assumption is already homogeneous. In fact, for inflation to get started, a sufficiently large region which is not too inhomogeneous and not too curved, is needed. We shall not discuss this in more detail, since the solution of these problems is not the most important aspect of inflation.

If inflation happened, we expect that the early universe after inflation was very homogeneous except for fluctuations generated during inflation and that $\Omega_0 = 1$. Thus inflation leads to predictions that can be tested with observations. More important than flatness and homogeneity

[4]Thus, if it were discovered by observations, that actually $\Omega_0 \neq 1$, this would be a blow to the credibility of inflation. However, there is a version of inflation, called *open inflation*, for which it is natural that $\Omega_0 < 1$. The existence of such models of inflation have led critics of inflation to complain that inflation is "unfalsifiable"—no matter what the observation, there is a model of inflation that agrees with it. Nevertheless, most models of inflation give the same "generic" predictions, including $\Omega_0 = 1$.

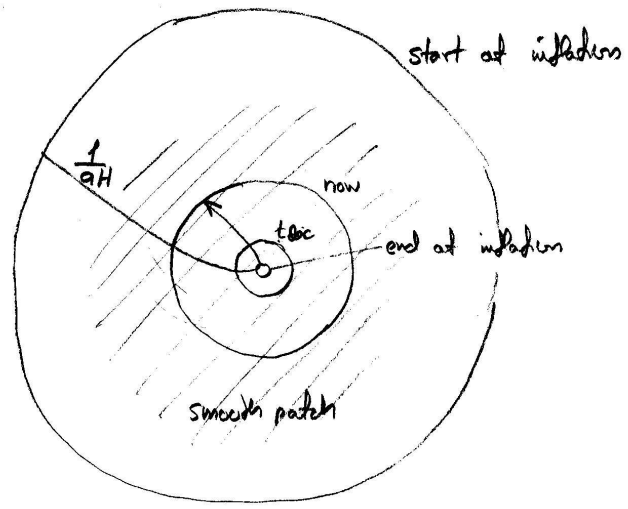


Figure 3: Evolution of the comoving Hubble radius (length, distance) during and after inflation (schematic).

are the predictions inflation makes about primordial perturbations, the "seeds" for structure formation, discussed in the next chapter.

Thus we assume that sufficient inflation has already taken place to make the universe (within a horizon volume) flat and homogeneous, and follow the inflation in detail after that, working in the flat FRW universe.

7.3 Quantum field theory for children

The theories (known and hypothetical) needed to describe the (very) early universe are *quantum field theories* (QFT). The fundamental entities of these theories are *fields*, i.e., functions of space and time. For each particle species, there is a corresponding field, having at least as many (real) components φ_i as the particle has internal degrees of freedom. For example, for the photon, the corresponding field is the vector field $A^\mu = (A^0, A^1, A^2, A^3) = (\phi, \vec{A})$, which you are probably familiar with from electrodynamics. The photon has two internal degrees of freedom. The larger number of components in A^μ is related to the *gauge freedom* of electrodynamics. Since A^μ is a (Lorentz) vector field, it has the same number of components as there are spacetime dimensions, but other types of fields do not have this correspondence.

In classical field theory the evolution of the field is governed by the *field equation*. From the field equation one can identify a field potential, an expression in terms of the field, which helps to understand the field dynamics. Quantizing a field theory gives a quantum field theory. *Particles are quanta of the oscillations of the field around the minimum of its potential.* The state where the field values are constant at the potential minimum is called the *vacuum*. Up to now, we have described the events in the early universe in terms of the *particle picture*. However, the particle picture is not fundamental, and can be used only when the fields are doing small oscillations. For many possible events and objects in the early universe (inflation, topological defects, spontaneous symmetry breaking phase transitions) the field behavior is different, and we need to describe them in terms of field theory. In some of these topics classical field theory is already sufficient for a reasonable and useful description.

In this section we discuss "low-temperature" field theory in Minkowski space, i.e., we forget high-temperature effects and the curvature of spacetime.

The starting point in field theory is the *Lagrangian density*, a function of space and time, which is a scalar quantity constructed from the fields and their derivatives:

$$\mathcal{L}(\varphi_i, \partial_\mu \varphi_i)\,. \tag{20}$$

The Lagrangian density can be expressed as a sum of two parts, the *kinetic term*, which depends on field derivatives, and the *field potential* $V(\varphi_1, \ldots, \varphi_N)$ (for a theory with N fields), which does not. This expression for the Lagrangian density as a function of the fields and their derivatives defines the field theory, and one can derive the field equations (differential equations governing the field evolution) and the energy-momentum tensor (energy density and pressure of the fields) from the Lagrangian density. Usually the kinetic term has a simple form, called the canonical kinetic term, and we assume that here. The remaining freedom in defining the field theory is in defining the potential.

The simplest case is a theory with one scalar field φ, for which

$$\mathcal{L} \;=\; -\tfrac{1}{2}\partial_\mu\varphi\,\partial^\mu\varphi - V(\varphi)\,. \tag{21}$$

(We use the Einstein summation convention, where a repeated index implies summation over it, here $\mu = 0, 1, 2, 3$. Also $\partial_\mu \equiv \partial/\partial x^\mu$, where $x^0 = t$. Here we are in Minkowski space and use Cartesian coordinates, so that $\partial^0 = -\partial_0$ and $\partial^j = \partial_j$ for $j = 1, 2, 3$.) We write

$$V'(\varphi) \equiv \frac{dV}{d\varphi} \qquad \text{and} \qquad V''(\varphi) \equiv \frac{d^2V}{d\varphi^2}\,. \tag{22}$$

The *field equation*, which determines the classical evolution of the field, is obtained from the Lagrangian by minimizing (or extremizing) the action

$$\int \mathcal{L}\, d^4x\,, \tag{23}$$

which leads to the *Euler–Lagrange equation*

$$\frac{\partial \mathcal{L}}{\partial \varphi_i(x)} - \partial_\mu \frac{\partial \mathcal{L}}{\partial[\partial_\mu \varphi_i(x)]} = 0\,. \tag{24}$$

For the above scalar field we get the field equation

$$\partial_\mu \partial^\mu \varphi - V'(\varphi) = 0\,. \tag{25}$$

where $\partial_\mu \partial^\mu \varphi = -\ddot{\varphi} + \nabla^2 \varphi$, so that the field equation is

$$\boxed{\ddot{\varphi} - \nabla^2 \varphi \;=\; -V'(\varphi)\,.} \tag{26}$$

Here we use the overdot to denote partial derivative with respect to time: $\dot{} = \partial_0 = \partial/\partial t$.

The Lagrangian also gives us the energy tensor

$$T^{\mu\nu} = -\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi)} \partial^\nu \varphi + g^{\mu\nu} \mathcal{L}\,. \tag{27}$$

For the scalar field

$$T^{\mu\nu} = \partial^\mu \varphi \partial^\nu \varphi - g^{\mu\nu} \big[\frac{1}{2} \partial_\rho \varphi \partial^\rho \varphi + V(\varphi) \big]\,. \tag{28}$$

In particular, the energy density $\rho = T^{00}$ and pressure $p = \frac{1}{3}(T^{11} + T^{22} + T^{33})$ of a scalar field are

$$\rho \;=\; \frac{1}{2}\dot{\varphi}^2 + \frac{1}{2}(\nabla\varphi)^2 + V(\varphi) \tag{29}$$

$$p \;=\; \frac{1}{2}\dot{\varphi}^2 - \frac{1}{6}(\nabla\varphi)^2 - V(\varphi)\,. \tag{30}$$

(We are in Minkowski space, so that $g^{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$). We see that the pressure due to a scalar field may be negative. The minimum value of $V(\varphi)$ is the *vacuum energy*. In principle it could be negative, acting like a negative cosmological constant. Any other contribution to ρ is positive. Since there is no evidence for a negative vacuum energy or cosmological constant, let us assume that $V(\varphi) \geq 0$.

If $V(\phi) = 0$ the field equation becomes the wave equation

$$\ddot{\varphi} \;=\; \nabla^2 \varphi\,, \tag{31}$$

whose solutions are waves propagating at the speed of light.

For the corresponding quantum theory, the potential gives information about the masses and interactions of the particles that are the quanta of the field oscillations. The particles corresponding to scalar fields are spin-0 bosons. Spin-$\frac{1}{2}$ particles correspond to spinor fields and spin-1 particles to vector fields. The case $V(\varphi) = 0$ corresponds to massless noninteracting particles. If the potential has the form

$$V(\varphi) = \frac{1}{2} m^2 \varphi^2\,, \tag{32}$$

the particle corresponding to the field φ will have mass m and it will have no interactions. In general, the mass of the particle is given by $m^2 = V''(\varphi)$.

Interactions between particles of two different species are due to terms in the Lagrangian which involve both fields. For example, in the Lagrangian of quantum electrodynamics (QED) the term

$$-ie\psi^\dagger \gamma^0 \gamma^\mu A_\mu \psi \tag{33}$$

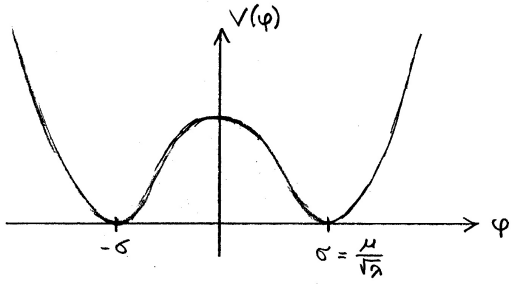
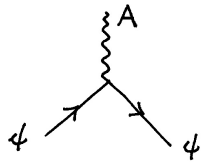


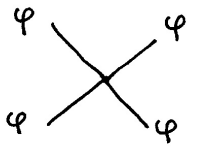
Figure 4: Potential giving rise to spontaneous symmetry breaking.

is responsible for the interaction between photons (A^μ) and electrons (ψ). (The γ^μ are Dirac matrices). A graphical representation of this interaction is the Feynman diagram



A higher power (third or fourth) of a field, e.g.,

$$V(\varphi) = \frac{1}{4}\lambda\varphi^4\,, \tag{34}$$



represents self-interaction, i.e., φ particles interacting with each other directly (as opposed to, e.g., electrons, who interact with each other indirectly, via photons). In QCD, gluons have this property.

Some theories exhibit *spontaneous symmetry breaking* (SSB). For example, the potential

$$V(\varphi) = V_0 - \frac{1}{2}\mu^2\varphi^2 + \frac{1}{4}\lambda\varphi^4 \tag{35}$$

has two minima, at $\varphi = \pm\sigma$, where $\sigma = \mu/\sqrt{\lambda}$. At low temperatures, the field is doing small oscillations around one of these two minima (see Fig. 4). Thus the vacuum value of the field is nonzero. If the Lagrangian has interaction terms, $c\varphi\psi^2$, with other fields ψ, these can now be separated into a mass term, $c\sigma\psi^2$, for ψ and an interaction term, by redefining the field φ as

$$\varphi = \sigma + \tilde{\varphi} \quad \Rightarrow \quad c\varphi\psi^2 = c\sigma\psi^2 + c\tilde{\varphi}\psi^2\,. \tag{36}$$

Thus spontaneous symmetry breaking gives the ψ particles a mass $\sqrt{2c\sigma}$. This kind of a field φ is called a *Higgs field*. In electroweak theory the fermion masses are due to a Higgs field.

7.4 Inflaton field

As we saw in Sec. 7.2, inflation requires negative pressure. In Chapter 4 we considered systems of particles where interaction energies can be neglected (ideal gas approximation). For such systems the pressure is always nonnegative. However, negative pressure is possible in systems with attractive interactions. In the field picture, negative pressure comes from the potential term. In many models of inflation, the inflation is caused by a scalar field. This scalar field (and the corresponding spin-0 particle) is called the *inflaton*.

Historical note. The idea of scalar fields playing an important role in the very early universe was very natural at the time inflation was proposed by Guth[1]. We already mentioned how a scalar field, the Higgs field, is responsible for the electroweak phase transition at $T \sim 100\,\mathrm{GeV}$. It is thought that at a much higher temperature, $T \sim 10^{14}\,\mathrm{GeV}$, another spontaneous symmetry breaking phase transition occurred, the GUT (Grand Unified Theory) phase transition, so that above this temperature the strong and electroweak forces were unified. This GUT phase transition gives rise to the monopole problem. Guth realized that the Higgs field associated with the GUT transition might lead the universe to "inflate" (the term was coined by Guth), solving this monopole problem. It was soon found out, however, that inflation based on the GUT Higgs field is not a viable inflation model, since in this model too strong inhomogeneities were created. So the inflaton field must be some other scalar field. The supersymmetric extensions of the standard model contain many inflaton field candidates.

During inflation the inflaton field is *almost* homogeneous.[5] The energy density and pressure of the inflaton are thus those of a homogeneous scalar field,

$$\begin{aligned} \rho &= \frac{1}{2}\dot{\varphi}^2 + V(\varphi) \\ p &= \frac{1}{2}\dot{\varphi}^2 - V(\varphi)\,, \end{aligned} \tag{37}$$

where $V(\varphi) \geq 0$. For the equation-of-state parameter $w \equiv p/\rho$ we have

$$w = \frac{\dot{\varphi}^2 - 2V(\varphi)}{\dot{\varphi}^2 + 2V(\dot{\varphi})} = \frac{1 - (2V/\dot{\varphi}^2)}{1 + (2V/\dot{\varphi}^2)}\,, \tag{38}$$

so that

$$-1 \leq w \leq 1\,. \tag{39}$$

If the kinetic term $\dot{\varphi}^2$ dominates, $w \approx 1$; if the potential term $V(\varphi)$ dominates, $w \approx -1$.

For the present discussion, the potential $V(\varphi)$ is some arbitrary non-negative function. Different inflaton models correspond to different $V(\varphi)$. From Eq. (37), we get the useful combinations

$$\begin{aligned} \rho + p &= \dot{\varphi}^2 \\ \rho + 3p &= 2\left[\dot{\varphi}^2 - V(\varphi)\right]\,. \end{aligned} \tag{40}$$

We already had the field equation for a scalar field in Minkowski space,

$$\ddot{\varphi} - \nabla^2\varphi = -V'(\varphi)\,. \tag{41}$$

For the homogenous case it is just

$$\ddot{\varphi} = -V'(\varphi)\,. \tag{42}$$

We get a working mental picture of the evolution of a homogeneous field by comparing it to the classical mechanics equation for a particle in a gravitational potential $V(\mathbf{r})$ whose accelaration

[5] Inflation makes the inflaton field homogenous. Again, a sufficient level of initial homogeneity of the field is required to get inflation started. We start our discussion when a sufficient level of inflation has already taken place to make the gradients negligible.

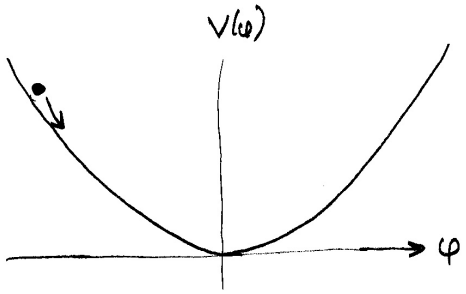


Figure 5: The inflaton and its potential.

is given by $\ddot{\mathbf{r}} = -\nabla V(\mathbf{r})$. Thus we can think of the field "rolling down" its potential like a stone rolling down a hillside, see Fig. 5; this motion is governed by Eq. (42).

We need to modify (42) for the expanding universe. We do not need to go to the GR formulation of field theory, since the modification for the present case can be simply obtained by sticking the ρ and p from Eq. (37) into the energy continuity equation

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

This gives

$$\dot{\varphi}\ddot{\varphi} + V'(\varphi)\dot{\varphi} = -3H\dot{\varphi}^2 \;\Rightarrow\; \boxed{\ddot{\varphi} + 3H\dot{\varphi} = -V'(\varphi)}, \tag{44}$$

the field equation for a homogeneous φ in an expanding (FRW) universe. We see that the effect of expansion is to add the term $3H\dot{\varphi}$, which acts like a *friction term*, slowing down the evolution of φ.

The condition for inflation, $\rho + 3p = 2\dot{\varphi}^2 - 2V(\dot{\varphi}) < 0$, is satisfied, if

$$\dot{\varphi}^2 < V(\varphi)\,. \tag{45}$$

The idea of inflation is that φ is initially far from the minimum of $V(\varphi)$. The potential then pulls φ towards the minimum. See Fig. 5. If the potential has a suitable (sufficiently flat) shape, the friction term soon makes $\dot{\varphi}$ small enough to satisfy Eq. (45), even if it was not satisfied initially.

We shall also need the Friedmann equation for the flat universe,

$$H^2 = \frac{8\pi G}{3}\rho = \frac{1}{3M_{\rm Pl}^2}\rho\,. \tag{46}$$

where we have introduced the *reduced Planck mass*

$$M_{\rm Pl} \equiv \frac{1}{\sqrt{8\pi}} m_{\rm Pl} \equiv \frac{1}{\sqrt{8\pi G}} = 2.436 \times 10^{18}\,{\rm GeV}\,. \tag{47}$$

Inserting Eq. (37), this becomes

$$\boxed{H^2 = \frac{1}{3M_{\rm Pl}^2}\left[\frac{1}{2}\dot{\varphi}^2 + V(\varphi)\right]\,.} \tag{48}$$

We have ignored other components to energy density and pressure besides the inflaton. During inflation, the inflaton φ moves slowly, so that the inflaton energy density, which is dominated by $V(\varphi)$ also changes slowly. If there are matter and radiation components to the energy density, they decrease fast, $\rho \propto a^{-3}$ or $\propto a^{-4}$ and soon become negligible. Again, this puts some initial conditions for inflation to get started, for the inflaton to become dominant. But once inflation gets started, we can soon forget the other components to the universe besides the inflaton.

7.5 Slow-roll inflation

The friction (expansion) term tends to slow down the evolution of φ, so that we may easily reach a situation where:

$$\dot{\varphi}^2 \ll V(\varphi) \tag{49}$$

$$|\ddot{\varphi}| \ll |3H\dot{\varphi}| \tag{50}$$

These are the *slow-roll conditions*.

If the slow-roll conditions are satisfied, we may approximate (the *slow-roll approximation*) Eqs. (48) and (44) by the *slow-roll equations*:

$$H^2 = \frac{V(\varphi)}{3M_{\rm Pl}^2} \tag{51}$$

$$3H\dot{\varphi} = -V'(\varphi) \tag{52}$$

The shape of the potential $V(\varphi)$ determines the *slow-roll parameters*:

$$\begin{aligned} \varepsilon(\varphi) &\equiv \frac{1}{2}M_{\rm Pl}^2\left(\frac{V'}{V}\right)^2 \\ \eta(\varphi) &\equiv M_{\rm Pl}^2\frac{V''}{V} \end{aligned} \tag{53}$$

Exercise: Show that

$$\varepsilon \ll 1 \quad \text{and} \quad |\eta| \ll 1 \qquad \Leftarrow \qquad \text{Eqs. (49) and (50)} \tag{54}$$

Note that the implication goes only in this direction. The conditions $\varepsilon \ll 1$ and $|\eta| \ll 1$ are necessary, but not sufficient for the slow-roll approximation to be valid (i.e., the slow-roll conditions to be satisfied).

The conditions $\varepsilon \ll 1$ and $|\eta| \ll 1$ are just *conditions on the shape of the potential*, and identify from the potential a *slow-roll section*, where the slow-roll approximation *may* be valid. Since the initial field equation, Eq. (44) was second order, it accepts arbitrary φ and $\dot{\varphi}$ as initial conditions. Thus Eqs. (49) and (50) may not hold initially, even if φ is in the slow-roll section. However, it turns out that the *slow-roll solution*, the solution of the slow-roll equations (51) and (52), is an *attractor* of the full equations, (48) and (44). This means that the solution of the full equations rapidly approaches it, starting from arbitrary initial conditions. Well, not fully arbitrary, the initial conditions need to lie in the *basin of attraction*, from which they are then attracted into the attractor. To be in the basin of attraction, means that φ must be in the slow-roll section, and that if $\dot{\varphi}$ is very large, φ needs to be deeper in the slow-roll section.

Once we have reached the attractor, where Eqs. (51) and (52) hold, $\dot{\varphi}$ is determined by φ (since we replaced the second-order differential equation with a first-order one). In fact everything is determined by φ (assuming a known form of $V(\varphi)$). The value of φ is the *single parameter describing the state of the universe*, and φ evolves down the potential $V(\varphi)$ as specified by the slow-roll equations.

This language of "attractor" and "basin of attraction" can be taken further. **If** the universe (or a region of it) finds itself initially (or enters) the basin of attraction of slow-roll inflation, meaning that: there is a sufficiently large region, where the curvature is sufficiently small, the inflaton makes a sufficient contribution to the total energy density, the inflaton is sufficiently homogeneous, and lies sufficiently deep in the slow-roll section, **then** this region begins inflating,

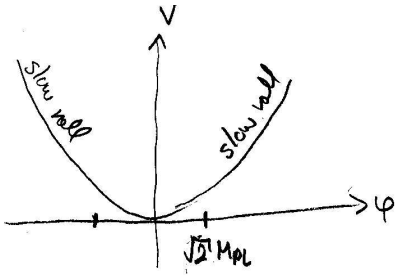


Figure 6: The potential $V(\varphi) = \frac{1}{2}m^2\varphi^2$ and its two slow-roll sections.

it becomes rapidly very homogeneous and flat, all other contributions to the energy density besides the inflaton become negligible, and the inflaton begins to follow the slow-roll solution.

Thus inflation *erases all memory of the initial conditions*, and we can predict the later history of the universe just from the shape of $V(\varphi)$ and the assumption that φ started out far enough in the slow-roll part of it.

Example: The simplest model of inflation (see Fig. 6) is the one where

$$V(\varphi) = \frac{1}{2}m^2\varphi^2 \quad \Rightarrow \quad V'(\varphi) = m^2\varphi\,, \quad V''(\varphi) = m^2\,. \tag{55}$$

The slow-roll parameters are

$$\left.\begin{aligned} \varepsilon(\varphi) &= \frac{1}{2}M_{\rm Pl}^2\left(\frac{2}{\varphi}\right)^2 \\ \eta(\varphi) &= M_{\rm Pl}^2\frac{2}{\varphi^2} \end{aligned}\right\} \quad \Rightarrow \quad \varepsilon = \eta = 2\left(\frac{M_{\rm Pl}}{\varphi}\right)^2 \tag{56}$$

and the slow-roll section of the potential is given by the condition

$$\varepsilon, \eta \ll 1 \quad \Rightarrow \quad \varphi^2 \gg 2M_{\rm Pl}^2\,. \tag{57}$$

7.5.1 Relation between inflation and slow roll

From the definition of the Hubble parameter,

$$H = \frac{\dot a}{a} \quad \Rightarrow \quad \dot H = \frac{\ddot a}{a} - \frac{\dot a^2}{a^2} \quad \Rightarrow \quad \boxed{\frac{\ddot a}{a} = \dot H + H^2} \tag{58}$$

Thus the condition for inflation is $\dot H + H^2 > 0$. This would be satisfied, if $\dot H > 0$, but this is not possible here, since it would require $p < -\rho$, i.e., $w \equiv p/\rho < -1$, which is not allowed by Eq. (37).[6] Thus

$$\boxed{\dot H \le 0} \tag{59}$$

and

$$\text{Inflation} \;\Leftrightarrow\; -\frac{\dot H}{H^2} < 1 \tag{60}$$

[6] From the Friedmann eqs.,

$$\left.\begin{aligned} \left(\frac{\dot a}{a}\right)^2 &= \frac{8\pi G}{3}\rho - \frac{K}{a^2} \\ \frac{\ddot a}{a} &= -\frac{4\pi G}{3}(\rho + 3p) \end{aligned}\right\} \quad \Rightarrow \quad \dot H \;=\; \frac{\ddot a}{a} - \frac{\dot a^2}{a^2} \;=\; -4\pi G(\rho + p) + \frac{K}{a^2}$$

In the above, we are assuming space is already flat, i.e., $K = 0$. Then $\dot H > 0 \;\Rightarrow\; \rho + p < 0$.

If the slow-roll approximation is valid,

$$H^2 = \frac{V}{3M_{\rm Pl}^2} \quad \Rightarrow \quad 2H\dot{H} = \frac{V'\dot{\varphi}}{3M_{\rm Pl}^2} \quad \Rightarrow \quad H^2\dot{H} = \frac{V'H\dot{\varphi}}{6M_{\rm Pl}^2} \overset{3H\dot{\varphi}=-V'}{=} -\frac{V'^2}{18M_{\rm Pl}^2}$$
$$\Rightarrow \quad -\frac{\dot{H}}{H^2} = \frac{V'^2}{18M_{\rm Pl}^2}\frac{9M_{\rm Pl}^4}{V^2} = \frac{1}{2}M_{\rm Pl}^2\left(\frac{V'}{V}\right)^2 = \varepsilon \ll 1$$

Therefore, *if the slow-roll approximation is valid, inflation is guaranteed*. This is a sufficient, not necessary condition. The above result for slow-roll inflation, $-\dot{H}/H^2 \ll 1$ can also be written as

$$\left|\frac{\dot{H}}{H}\right| \ll \frac{\dot{a}}{a}\,. \tag{61}$$

During slow-roll inflation, the Hubble parameter H changes much more slowly than the scale factor a. For a constant H, the universe expands exponentially, since

$$\frac{\dot{a}}{a} = \frac{d\ln a}{dt} = H = const \quad \Rightarrow \quad \ln\frac{a}{a_1} = H(t-t_1) \quad \Rightarrow \quad a \propto e^{Ht}\,. \tag{62}$$

Thus, in slow-roll inflation, the universe expands "almost exponentially".

Note that accelerated expansion, which is defined to mean that $\ddot{a} > 0$, does not mean that the *expansion rate*, as given by H, would increase. Even during inflation, $\dot{H} < 0$, so the expansion rate decreases. (There may be some ambiguity in what is meant by an increasing/decreasing expansion rate. The Hubble parameter is a better quantity to be called the expansion rate than $\dot{a}$, since the value of the latter depends on the normalization of a_0. With the normalization $a_0 = 1$, $H = \dot{a}$ "today".)

Sometimes it is carelessly said that inflation was a period of very rapid expansion. Rapid compared to what? Certainly the expansion rate was larger than today, or indeed larger than during any period after inflation (since $\dot{H} < 0$ always). But note that in the original Hot Big Bang picture $H \to \infty$ (and also $\dot{a} \to \infty$) as $t \to 0$. When we replace the earliest part of Hot Big Bang with inflation, we replace it with *slower* expansion, H almost constant (and $\dot{a}$ becoming smaller towards earlier times—this is what acceleration means), instead of $H \to \infty$.

It is possible to have inflation without the slow-roll parameters being small (fast-roll inflation), but we will see that slow-roll inflation produces the observed primordial perturbation spectrum naturally (unlike fast-roll inflation).

7.5.2 Models of inflation

A model of inflation[7] consists of

1. a potential $V(\varphi)$
2. a way of ending inflation

There are two ways of ending inflation:

1. Slow-roll approximation is no more valid, as φ approaches the minimum of the potential with $V(\varphi_{\rm min}) = 0$ or very small. For a reasonable approximation we can assume inflation ends, when $\varepsilon(\varphi) = 1$ or $|\eta(\varphi)| = 1$. Denote this value of the inflaton field by $\varphi_{\rm end}$.
2. Extra physics intervenes to end inflation (e.g., *hybrid* inflation). In this case inflation may end while the slow-roll approximation is valid.

[7] There are also models of inflation which are not based on a scalar field.

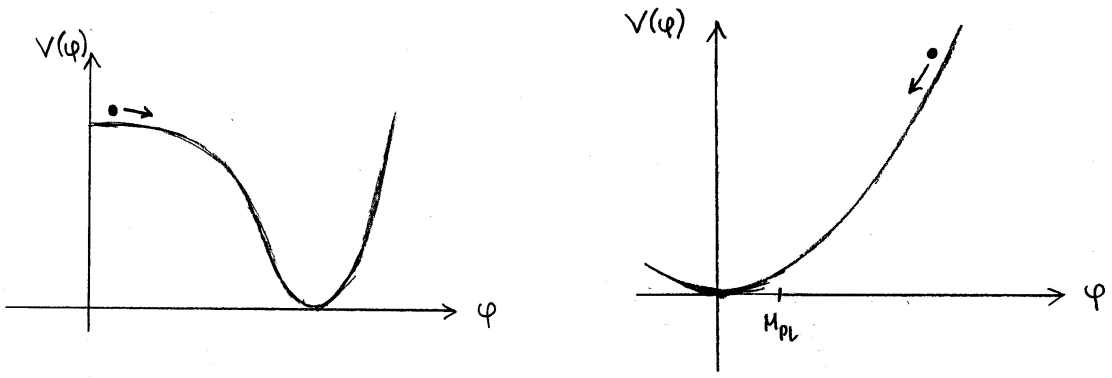


Figure 7: Potential for small-field (a) and large-field (b) inflation. For a typical small-field model, the entire range of φ shown is $\ll M_{\rm Pl}$.

Inflation models can be divided into two classes:

1. small-field inflation, $\Delta\varphi < M_{\rm Pl}$ in the slow-roll section
2. large-field inflation, $\Delta\varphi > M_{\rm Pl}$ in the slow-roll section

Here $\Delta\varphi$ is the range in which φ varies during (the observationally relevant part of) inflation. See Fig. 7 for typical shapes of potentials for large-field and small-field models.

Example of small-field inflation:

$$V(\varphi) = V_0 \left[1 - \frac{\lambda}{4}\left(\frac{\varphi}{M_{\rm Pl}}\right)^4 + \dots\right], \tag{63}$$

where the omitted terms, responsible for keeping $V \geq 0$ for larger φ are assumed negligible in the region of interest. We assume further that the second term is small in the slow-roll section, so that we can approximate $V(\varphi) \approx V_0$ except for its derivatives. The slow-roll parameters are then

$$\varepsilon = \frac{1}{2}\lambda^2 \left(\frac{\varphi}{M_{\rm Pl}}\right)^6 \qquad \text{and} \qquad \eta = -3\lambda \left(\frac{\varphi}{M_{\rm Pl}}\right)^2, \tag{64}$$

so that

$$\frac{\varepsilon}{|\eta|} = \frac{1}{6}\lambda \left(\frac{\varphi}{M_{\rm Pl}}\right)^4 \ll 1\,. \tag{65}$$

Thus $\eta < 0$ and $\varepsilon \ll |\eta|$, which is typical for small-field inflation, and inflation ends when

$$|\eta| = 1 \quad \Rightarrow \quad \varphi_{\rm end} = \frac{M_{\rm Pl}}{\sqrt{3\lambda}}\,. \tag{66}$$

The assumption that the second term in the potential is still small at $\varphi_{\rm end}$, requires that $\lambda \gtrsim 1$, and thus $|\eta| \ll 1$ requires $\varphi \ll M_{\rm Pl}/\sqrt{3}$, so this is indeed a small-field model.

Example of large-field inflation: A simple monomial potential of the form

$$V(\varphi) = A\varphi^n \qquad (n > 1)\,. \tag{67}$$

The slow-roll parameters are

$$\varepsilon = \frac{n^2}{2}\left(\frac{M_{\rm Pl}}{\varphi}\right)^2 \qquad \text{and} \qquad \eta = n(n-1)\left(\frac{M_{\rm Pl}}{\varphi}\right)^2, \tag{68}$$

so that $\eta > 0$ and ε and η are of similar size, typical for large-field inflation. This is a large-field model, since $\varepsilon \ll 1$ requires $\varphi^2 \gg \frac{1}{2}n^2 M_{\rm Pl}^2$.

For the special case of $V(\varphi) = \frac{1}{2}m^2\varphi^2$, $\varepsilon = \eta$, and inflation ends at $\varphi_{\rm end} = \sqrt{2}M_{\rm Pl}$. To get inflation to end, e.g., at energy scale $V(\varphi_{\rm end}) \equiv m^2 M_{\rm Pl}^2 = \left(10^{14}\,{\rm GeV}\right)^4$, we need $m = \left(10^{14}\,{\rm GeV}\right)^2/M_{\rm Pl} \approx 4\times 10^9\,{\rm GeV}$.

7.5.3 Exact solutions

Usually the slow-roll approximation is sufficient. It fails near the end of inflation, but this just affects slightly our estimate of the total amount of inflation. It is much easier to solve the slow-roll equations, (51) and (52), than the full equations, (44) and (48). However, it is useful to have some exact solutions to the full equations, for comparison. For some special cases, exact analytical solutions exist.

One such case is *power-law inflation*, where the potential is

$$V(\varphi) = V_0 \exp\left(-\sqrt{\frac{2}{p}}\frac{\varphi}{M_{\rm Pl}}\right) , \qquad p > 1 , \tag{69}$$

where V_0 and p are constants.

An exact solution for the full equations, (44) and (48), is (**exercise**)

$$a(t) \propto t^p \tag{70}$$

$$\varphi(t) = \sqrt{2p} M_{\rm Pl} \ln\left(\sqrt{\frac{V_0}{p(3p-1)}}\frac{t}{M_{\rm Pl}}\right) . \tag{71}$$

The general solution approaches rapidly this particular solution (i.e., it is an attractor). You can see that the expansion, $a(t)$, is power-law, giving the model its name.

The slow-roll parameters for this model are

$$\varepsilon = \frac{1}{2}\eta = \frac{1}{p} , \tag{72}$$

independent of φ. In this model inflation never ends, unless other physics intervenes.

7.6 Reheating

During inflation, practically all the energy in the universe is in the inflaton potential $V(\varphi)$, since the slow-roll condition says $\frac{1}{2}\dot\varphi^2 \ll V(\varphi)$. When inflation ends, this energy is transferred in the reheating process to a thermal bath of particles produced in the reheating. Thus reheating creates, from $V(\varphi)$, all the stuff there is in the later universe!

Note that *re*heating may be a misnomer, since we don't know whether the universe was in a thermodynamical equilibrium ever before.

In single-field models of inflation, reheating does not affect the primordial density perturbations,[8] except that it affects the relation of φ_k and k/H_0 given in (85), i.e., how much the distance scale of the perturbations is stretched between inflation and today (these will be discussed later).

Reheating is important for the question of whether unwanted—or wanted—relics are produced after inflation. The reheating temperature must be high enough so that we get standard Big Bang Nucleosynthesis (BBN) after reheating, but sufficiently low so that we do not produce unwanted relics. The latter constraint depends on the extended theory, but it should at least be below the GUT scale. Thus we can take that

$$1\,{\rm MeV} < T_{\rm reh} < 10^{14}\,{\rm GeV} . \tag{73}$$

7.6.1 Scalar field oscillations

After inflation, the inflaton field φ begins to oscillate at the bottom of the potential $V(\varphi)$, see Fig. 8. The inflaton field is still homogeneous, $\varphi(t,\vec{x}) = \varphi(t)$, so it oscillates in the same phase

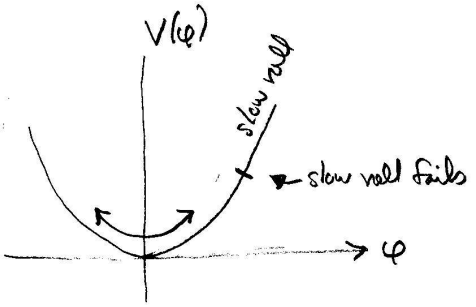


Figure 8: After inflation, the inflaton field is left oscillating at the bottom.

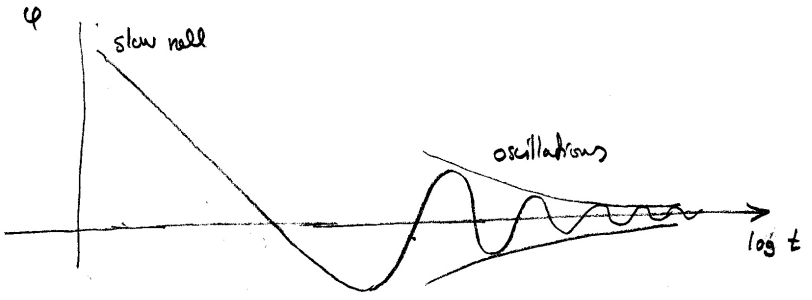


Figure 9: The time evolution of φ as inflation ends.

everywhere (we say the oscillation is *coherent*). The expansion time scale H^{-1} soon becomes much longer than the oscillation period.

Assume the potential can be approximated as $\propto \varphi^2$ near the minimum of $V(\varphi)$, so that we have a harmonic oscillator. Write $V(\varphi) = \frac{1}{2}m^2\varphi^2$:

$$\left.\begin{aligned} \ddot{\varphi} + 3H\dot{\varphi} &= -V'(\varphi) \\ \rho &= \frac{1}{2}\dot{\varphi}^2 + V(\varphi) \end{aligned}\right\} \quad \text{become} \quad \begin{cases} \ddot{\varphi} + 3H\dot{\varphi} = -m^2\varphi \\ \rho = \frac{1}{2}\left(\dot{\varphi}^2 + m^2\varphi^2\right) \end{cases}$$

What is $\rho(t)$?

$$\dot{\rho} + 3H\rho \;=\; \dot{\varphi}\overbrace{\left(\ddot{\varphi} + m^2\varphi\right)}^{-3H\dot{\varphi}} + 3H\cdot\frac{1}{2}\left(\dot{\varphi}^2 + m^2\varphi^2\right) \;=\; \frac{3}{2}H\overbrace{\left(m^2\varphi^2 - \dot{\varphi}^2\right)}^{\text{oscillates}}$$

The oscillating factor on the right hand side averages to zero over one oscillation period (in the limit where the period is $\ll H^{-1}$).

Averaging over the oscillations, we get that the long-time behavior of the energy density is

$$\dot{\rho} + 3H\rho = 0 \quad \Rightarrow \quad \rho \propto a^{-3}\,, \tag{74}$$

just like in a matter-dominated universe (we use this result in Sec. 7.7.2). The fall in the energy density shows as a decrease of the oscillation amplitude, see Fig. 9.

[8] In more complicated models of inflation, involving several fields, reheating may also change the nature of primordial density perturbations.

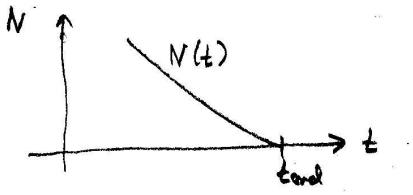


Figure 10: Remaining number of e-foldings $N(t)$ as a function of time.

7.6.2 Inflaton decays

Now that the inflaton field is doing small oscillations around the potential minimum, the particle picture becomes appropriate, and we can consider the energy density ρ_φ to be due to inflaton particles. These inflatons decay into other particles, once the Hubble time ($\sim$ the time after inflation ended) reaches the inflaton decay time.

If the decay is slow (which is the case if the inflaton can only decay into fermions) the inflaton energy density follows the equation

$$\dot{\rho}_\varphi + 3H\rho_\varphi \;=\; -\Gamma_\varphi \rho_\varphi \,, \tag{75}$$

where $\Gamma_\varphi = 1/\tau_\varphi$, the *decay width*, is the inverse of the inflaton decay time τ_φ, and the term $-\Gamma_\varphi\rho_\varphi$ represents energy transfer to other particles.

If the inflaton can decay into bosons, the decay may be very rapid, involving a mechanism called *parametric resonance*. This kind of rapid decay is called *preheating*, since the bosons thus created are far from thermal equilibrium (occupation numbers of states are huge—not possible for fermions).

7.6.3 Thermalization

The particles produced from the inflatons will interact, create other particles through particle reactions, and the resulting particle soup will eventually reach thermal equilibrium with some temperature $T_{\rm reh}$. This *reheating temperature* is determined by the energy density $\rho_{\rm reh}$ at the end of the reheating epoch:

$$\rho_{\rm reh} = \frac{\pi^2}{30} g_*(T_{\rm reh}) T_{\rm reh}^4 \,. \tag{76}$$

Necessarily $\rho_{\rm reh} < \rho_{\rm end}$ (end = end of inflation). If reheating takes a long time, we may have $\rho_{\rm reh} \ll \rho_{\rm end}$. After reheating, we enter the standard Hot Big Bang history of the universe.

7.7 Scales of inflation

7.7.1 Amount of inflation

During inflation, the scale factor $a(t)$ grows by a huge factor. We define the *number of e-foldings* from time t to end of inflation ($t_{\rm end}$) by

$$N(t) \equiv \ln \frac{a(t_{\rm end})}{a(t)} \tag{77}$$

See Fig. 10.

As we saw in Sec. 7.5.1, $a(t)$ changes much faster than $H(t)$ (when the slow-roll approximation is valid), so that the comoving Hubble length $\mathcal{H}^{-1} = 1/aH$ shrinks by almost as many e-foldings. ($a(t)$ grows fast, $H(t)$ decreases slowly.)

We can calculate $N(t) \equiv N(\varphi(t)) \equiv N(\varphi)$ from the shape of the potential $V(\varphi)$ and the value of φ at time t:

$$N(\varphi) \equiv \ln \frac{a(t_{\rm end})}{a(t)} = \int_t^{t_{\rm end}} H(t)dt = \int_\varphi^{\varphi_{\rm end}} \frac{H}{\dot\varphi} d\varphi \overset{\text{slow roll}}{\approx} \boxed{\frac{1}{M_{\rm Pl}^2} \int_{\varphi_{\rm end}}^{\varphi} \frac{V}{V'} d\varphi} . \tag{78}$$

where we used

$$d\ln a = \frac{da}{a} = Hdt = H\frac{d\varphi}{\dot\varphi} . \tag{79}$$

Example: For the simple inflation model $V(\varphi) = \frac{1}{2}m^2\varphi^2$,

$$N(\varphi) = \frac{1}{M_{\rm Pl}^2} \int_{\varphi_{\rm end}}^{\varphi} \frac{V}{V'} d\varphi = \frac{1}{M_{\rm Pl}^2} \int_{\varphi_{\rm end}}^{\varphi} \frac{\varphi}{2} = \frac{1}{4M_{\rm Pl}^2} \left(\varphi^2 - {\varphi_{\rm end}}^2\right) = \frac{1}{4}\left(\frac{\varphi}{M_{\rm Pl}}\right)^2 - \frac{1}{2} . \tag{80}$$

The largest initial value of φ we may contemplate is that which gives the Planck density, $V(\varphi) = M_{\rm Pl}^4 \quad \Rightarrow \quad \varphi = \sqrt{2}M_{\rm Pl}^2/m$. Starting from this value we get $\frac{1}{2}\left[(M_{\rm Pl}/m)^2 - 1\right]$ e-foldings of inflation. With $m = 4 \times 10^9\,\text{GeV}$ (see the earlier example with this model), this gives 1.85×10^{17} e-foldings, i.e., expansion by a factor $e^{1.85\times10^{17}} \sim 10^{8\times10^{16}} = 10^{80\,000\,000\,000\,000\,000}$. That's quite a lot!

7.7.2 Evolution of scales

When discussing (next chapter) evolution of density perturbations and formation of structure in the universe, we will be interested in the history of each comoving distance scale, or each *comoving wave number* k (from a Fourier expansion in comoving coordinates).

$$k = \frac{2\pi}{\lambda} , \qquad k^{-1} = \frac{\lambda}{2\pi}$$

An important question is, whether a distance scale is larger or smaller than the Hubble length at a given time.

We define a scale to be

- superhorizon, when $k < \mathcal{H} \quad \left(k^{-1} > \mathcal{H}^{-1}\right)$
- at horizon (exiting or entering horizon), when $k = \mathcal{H}$
- subhorizon, when $k > \mathcal{H} \quad \left(k^{-1} < \mathcal{H}^{-1}\right)$

Note that *large* scales (large k^{-1}) correspond to *low* k, and *vice versa*, although we often talk about "scale k". This can easily cause confusion, so watch for this, and be careful with wording! To avoid confusion, use the words *high*/*low* instead of large/small for k. Notice also that we are here using the word "horizon" to refer to the Hubble length.[9] Recall that:

$$\text{Inflation} \Rightarrow \mathcal{H}^{-1} \text{ shrinking}$$
$$\text{All other times} \Rightarrow \mathcal{H}^{-1} \text{ growing}$$

See Fig. 11.

[9] As discussed in Cosmology I, there are (at least) three different usages for the word "horizon":

1. particle horizon
2. event horizon (not used in Cosmology I/II)
3. Hubble length

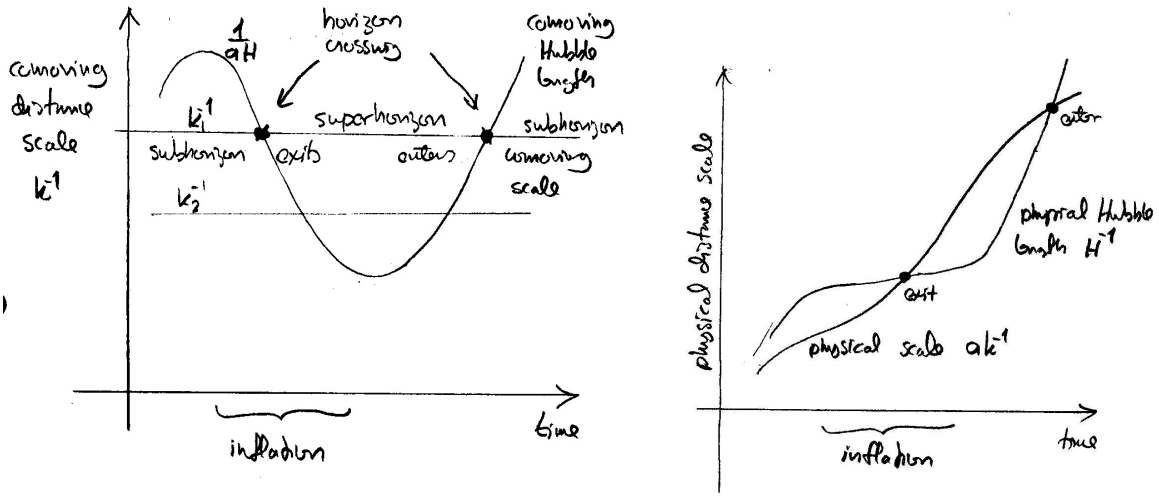


Figure 11: The evolution of the Hubble length, and two scales, k_1^{-1} and k_2^{-1}, seen in comoving coordinates (left) and in terms of physical distance (right).

We shall later find that the amplitude of primordial density perturbations at a given comoving scale is determined when this scale exits the horizon during inflation. The largest observable scales, $k \approx \mathcal{H}_0 = H_0$, are "at horizon" today. (Since the universe has recently began accelerating again, these scales have just barely entered, and are actually now exiting again.)

To identify the distance scales *during inflation* with the corresponding distance scales in the *present universe*, we need a complete history from inflation to the present. We divide it into the following periods:

1. **From** the time the scale k of interest exits the horizon during inflation **to** the end of inflation (t_k to $t_{\rm end}$).

2. **From** the end of inflation **to** reheating. We assume (as discussed in Sec. 7.6.1) that the universe behaves as if matter-dominated, $\rho \propto a^{-3}$, during this period ($t_{\rm end}$ to $t_{\rm reh}$).

3. **From** reheating **to** the present time ($t_{\rm reh}$ to t_0).

Consider now some scale k, which exits at $t = t_k$, when $a = a_k$ and $H = H_k$

$$\Rightarrow \quad k = \mathcal{H}_k = a_k H_k \, .$$

To find out how large this scale is today, we relate it to the present "horizon", i.e., the Hubble scale:

$$\frac{k}{H_0} \;=\; \frac{a_k H_k}{a_0 H_0} \;=\; \frac{a_k}{a_{\rm end}} \frac{a_{\rm end}}{a_{\rm reh}} \frac{a_{\rm reh}}{a_0} \frac{H_k}{H_0} \;=\; e^{-N(k)} \left(\frac{\rho_{\rm reh}}{\rho_{\rm end}}\right)^{\frac{1}{3}} \left(\frac{\rho_{r0}}{\rho_{\rm reh}}\right)^{\frac{1}{4}} \left(\frac{\rho_k}{\rho_{\rm cr0}}\right)^{\frac{1}{2}} , \qquad (81)$$

where $\rho_k \approx V(\varphi_k) \equiv V_k$ (since $\frac{1}{2}\dot{\varphi}^2 \ll V(\varphi)$ during slow roll) is the energy density when scale k exited and $N(k) \equiv$ number of e-foldings of inflation after that. Eq. (78) allows us to relate φ_k to $N(k)$. The factor $a_{\rm reh}/a_0 = a_{\rm reh}$ is related to the change in energy density from $t_{\rm reh} \to t_0$. The behavior of the total energy density as a function of a changed from the radiation-dominated to the matter-dominated to the dark-energy-dominated era, but we can keep things simpler by considering just the radiation component ρ_r, which was equal to the total energy density $\rho_{\rm reh}$ at end of reheating and behaves after that as $\rho_r \propto a^{-4}$. This is slightly inaccurate, since $\rho_r \propto a^{-4}$

does not take into account the change in g_*. However, the $\propto a^{-4}$ approximation is good enough[10] for us—we are making other comparable approximations also. From end of inflation to reheating

$$\rho \propto a^{-3} \quad \Rightarrow \quad \frac{a_{\rm end}}{a_{\rm reh}} = \left(\frac{\rho_{\rm reh}}{\rho_{\rm end}}\right)^{\frac{1}{3}} ,$$

and the ratio H_k/H_0 we got from

$$H_k = \sqrt{\frac{8\pi G}{3}\rho_k}\,, \; H_0 = \sqrt{\frac{8\pi G}{3}\rho_{\rm cr0}} \quad \Rightarrow \quad \frac{H_k}{H_0} = \left(\frac{\rho_k}{\rho_{\rm cr0}}\right)^{\frac{1}{2}} .$$

Thus we get that

$$\frac{k}{H_0} = e^{-N(k)} \left(\frac{\rho_{\rm reh}}{\rho_{\rm end}}\right)^{1/12} \left(\frac{V_k}{\rho_{\rm end}}\right)^{1/4} \frac{V_k^{1/4}\rho_{r0}^{1/4}}{\rho_{\rm cr0}^{1/2}} .$$

We can now relate $N(k)$ to k/H_0 as

$$N(k) \;=\; -\ln\frac{k}{H_0} - \frac{1}{3}\ln\frac{\rho_{\rm end}^{1/4}}{\rho_{\rm reh}^{1/4}} + \ln\frac{V_k^{1/4}}{\rho_{\rm end}^{1/4}} + \ln\frac{V_k^{1/4}}{10^{16}\,{\rm GeV}} + \ln\frac{10^{16}\,{\rm GeV}\cdot\rho_{r0}^{1/4}}{\rho_{\rm cr0}^{1/2}} , \tag{84}$$

where $10^{16}\,{\rm GeV}$ serves as a reference scale for V_k. This is roughly an upper limit to V_k due to lack of observation of primordial gravitational waves (discussed in the next chapter). Sticking in the known values of $\rho_{r0}^{1/4} = 2.4\times10^{-13}\,{\rm GeV}$ (assuming massless neutrinos; however, neutrino masses will not change the result for $N(\varphi_k)$) and $\rho_{\rm cr0}^{1/4} = 3.000\times10^{-12}\,{\rm GeV}\cdot h^{1/2}$, the last term becomes $60.85 - \ln h \approx 61$.

The final result is

$$N(\varphi_k) \;=\; -\ln\frac{k}{H_0} + 61 + \ln\frac{V_k^{1/4}}{\rho_{\rm end}^{1/4}} - \frac{1}{3}\ln\frac{\rho_{\rm end}^{1/4}}{\rho_{\rm reh}^{1/4}} - \ln\frac{10^{16}\,{\rm GeV}}{V_k^{1/4}} , \tag{85}$$

where the terms have been arranged so that they are all positive (when the sign in front of them is not included). Since the potential V_k changes slowly during slow roll, the k-dependence is dominated by the first term and the third term is small. The fourth term depends on how fast the reheating was. If it was instantaneous, this term is zero. The last term can be large if the inflation scale is much lower than $10^{16}\,{\rm GeV}$.

[10] Accurately this would go as:

$$g_{*s}a^3T^3 = {\rm const.} \quad \Rightarrow \quad \frac{a_{\rm reh}}{a_0} = \left[\frac{g_{*s}(T_0)}{g_{*s}(T_{\rm reh})}\right]^{\frac{1}{3}} \frac{T_0}{T_{\rm reh}} \tag{82}$$

Eq. (81) approximates this with

$$\left(\frac{\rho_{r0}}{\rho_{\rm reh}}\right)^{\frac{1}{4}} = \left[\frac{g_*(T_0)}{g_*(T_{\rm reh})}\right]^{\frac{1}{4}} \frac{T_0}{T_{\rm reh}} \tag{83}$$

Taking $g_{*s}(T_{\rm reh}) = g_*(T_{\rm reh}) \sim 100$, the ratio of these two becomes

$$\frac{(82)}{(83)} = \frac{g_{*s}(T_0)^{\frac{1}{3}}}{g_*(T_0)^{\frac{1}{4}} g_*(T_{\rm reh})^{\frac{1}{12}}} \approx \frac{3.909^{\frac{1}{3}}}{3.363^{\frac{1}{4}} 100^{\frac{1}{12}}} = 0.79 \sim 1$$

Note that $a \propto \rho_r^{-1/4}$ is a better approximation than $a \propto T^{-1}$, since these two differ by

$$\left[\frac{g_*(T_{\rm reh})}{g_*(T_0)}\right]^{\frac{1}{4}} \sim \left(\frac{100}{3.363}\right)^{\frac{1}{4}} \sim 2.33\,.$$

For any given present scale, given as a fraction of the present Hubble distance,[11] Eq. (85) identifies the value φ_k the inflaton had, when this scale exited the horizon during inflation. The last three terms give the dependence on the energy scales connected with inflation and reheating. In typical inflation models, they are relatively small. Usually, the precise value of N is not that important; we are more interested in the derivative dN/dk, or rather $d\varphi_k/dk$. We can see that typically (for high-energy-scale inflation) about 60 e-foldings of inflation occur *after* the largest observable scales exit the horizon. The number of e-foldings *before* that can be very large (e.g., billions or much more), depending on the inflation model and how the inflation is assumed to begin.

7.8 Initial conditions for inflation

Inflation provides the initial conditions for the Hot Big Bang. What about initial conditions for inflation? As we discussed earlier, inflation erases all memory of these initial conditions, removing this question from the reach of observational verification. However, a complete picture of the history of the universe should also include some idea about the conditions before inflation. To weigh how plausible inflation is as an explanation we may contemplate how easy it is for the universe to begin inflating.

Although inflation differs radically from the other periods of the history of the universe we have discussed, two qualitative features still hold true also during inflation: 1) the universe is expanding and 2) the energy density is decreasing (although slowly during inflation).

Thus the energy density should be higher before inflation than during it or after it. Often it is assumed that inflation begins right at the Planck scale, $\rho \sim M_{\rm Pl}^4$, which is the limit to how high energy densities we can extend our discussion, which is based on classical GR. Consider one such scenario:

When $\rho > M_{\rm Pl}^4$, quantum gravitational effects should be important. We can imagine that the universe at that time, the *Planck era*, is some kind of "spacetime foam", where the fabric of spacetime itself is subject to large quantum fluctuations. When the energy density of some region, larger than H^{-1}, falls below $M_{\rm Pl}^4$, spacetime in that region begins to behave in a classical manner. See Fig. 12. The initial conditions, i.e, conditions at the time when "our universe" (referring to one such region) emerges from the spacetime foam, are usually assumed *chaotic* (term due to Linde, does not refer to chaos theory), i.e., φ takes different, random, values at different regions. Since $\rho \geq \rho_\varphi$, and

$$\rho_\varphi = \frac{1}{2}\dot{\varphi}^2 + \frac{1}{2}\nabla\varphi^2 + V(\varphi)\,, \tag{86}$$

we must have

$$\dot{\varphi}^2 \lesssim M_{\rm Pl}^4\,, \quad \nabla\varphi^2 \lesssim M_{\rm Pl}^4\,, \quad V(\varphi) \lesssim M_{\rm Pl}^4 \tag{87}$$

in a region for it to emerge from the spacetime foam. If the conditions are suitable such a region may then begin to inflate. Thus inflation may begin at many different parts of the spacetime foam. Our observable universe would be just one small part of one such region which has inflated to a huge size.

It is also possible that during inflation, for some part of the potential, quantum fluctuations of the inflaton (not of the spacetime) dominate over the classical evolution, pushing φ higher in some regions. These regions will then expand faster, and dominate the volume. This gives rise to *eternal inflation*, where, at any given time, most of the volume of the universe is inflating. (This possibility depends on the shape of the potential.) But our observable universe would be a

[11] For example, $k/H_0 = 10$ means that we are talking about a scale corresponding to a wavelength λ, where $\lambda/2\pi$ is one tenth of the Hubble distance.

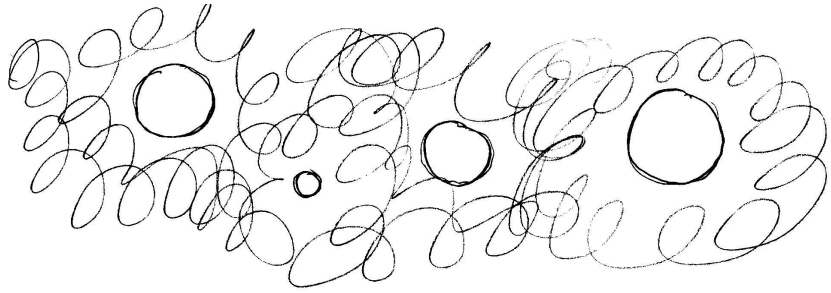


Figure 12: Spacetime foam and some regions emerging from it.

part of a region where φ came down to a region of the potential, where the quantum fluctuations of φ were small and the classical behavior began to dominate and eventually inflation ended.

Thus we see that the very, very, very large scale structure of the universe may be very complicated. But we will never discover this, since our entire observable universe is just a small homogeneous part of a patch which inflated, and then the inflation ended in that patch. All the observable features of the Universe can be explained in terms of what happened in this patch during and after inflation.

These ideas of spacetime foam and eternal inflation are rather speculative and there are also other suggestions for the initial stages of the universe.

References

[1] A.H. Guth, *Inflationary universe: A possible solution to the horizon and flatness problems*, Phys. Rev. D **23**, 347 (1981).

[2] A.R. Liddle and D.H. Lyth: Cosmological Inflation and Large-Scale Structure (Cambridge University Press 2000).

8 Structure Formation

Up to this point we have discussed the universe in terms of a homogeneous and isotropic model (which we shall now refer to as the "unperturbed" or the "background" universe). Clearly the universe is today rather inhomogeneous. By *structure formation* we mean the generation and evolution of this inhomogeneity. We are here interested in distance scales from galaxy size to the size of the whole observable universe. The structure is manifested in the existence of galaxies and in their uneven distribution, their *clustering*. This is the obvious inhomogeneity, but we understand it reflects a density inhomogeneity also in other, nonluminous, components of the universe, especially the *cold dark matter*. The structure has formed by gravitational amplification of a small primordial inhomogeneity. There are thus two parts to the theory of structure formation:

1) The generation of this primordial inhomogeneity, "the seeds of galaxies". This is the more speculative part of structure formation theory. We cannot claim that we know how this primordial inhomogeneity came about, but we have a good candidate scenario, *inflation*, whose predictions agree with the present observational data, and can be tested more thoroughly by future observations. In inflation, the structure originates from *quantum fluctuations* of the inflaton field φ near the time the scale in question exits the horizon.

2) The growth of this small inhomogeneity into the present observable structure of the universe. This part is less speculative, since we have a well established theory of gravity, *general relativity*. However, there is uncertainty in this part too, since we do not know the precise nature of the dominant components to the energy density of the universe, the *dark matter* and the *dark energy*. The gravitational growth depends on the equations of state and the streaming lengths (particle mean free path between interactions) of these density components. Besides gravity, the growth is affected by pressure forces.

We shall do the second part first. But before that we discuss statistical measures of inhomogeneity: correlation functions and power spectra.

8.1 Inhomogeneity

We write all our inhomogeneous quantities as a sum of a homogeneous background value, and a perturbation, the deviation from the background value. For example, for energy density and pressure we write

$$\begin{aligned} \rho(t,\mathbf{x}) &= \bar{\rho}(t) + \delta\rho(t,\mathbf{x}) \\ p(t,\mathbf{x}) &= \bar{p}(t) + \delta p(t,\mathbf{x})\,, \end{aligned} \tag{1}$$

where $\bar{\rho}$ and $\bar{p}$ are the background density and pressure, $\mathbf{x}$ is the *comoving* 3D space coordinate, and $\delta\rho$ and δp are the density and pressure perturbations. We further define the relative density perturbation

$$\delta(t,\mathbf{x}) \equiv \frac{\delta\rho(t,\mathbf{x})}{\bar{\rho}(t)}\,. \tag{2}$$

Since $\rho \geq 0$, necessarily $\delta \geq -1$. These quantities can be defined separately for different components to the energy density, e.g., matter, radiation, and dark energy. Perturbations in dark energy are expected to be small, and if it is just vacuum energy, it has no perturbations. When we discuss the later history of the universe, the main interest is in the matter density perturbation,

$$\delta_m(t,\mathbf{x}) \equiv \frac{\delta\rho_m(t,\mathbf{x})}{\bar{\rho}_m(t)}\,, \tag{3}$$

and then we will often write just δ for δ_m.

We do the split into the background and perturbation so that the background is equal to the mean (volume average) of the full quantity. An important question is, whether the $\bar{\rho}(t)$ and $\bar{p}(t)$ defined this way correspond to a (homogeneous and isotropic) solution of General Relativity, i.e., an FRW universe. We expect the exact answer to be negative, since GR is a nonlinear theory, so that perturbations affect the evolution of the mean. This effect is called *backreaction*.

However, if the perturbations are small, we can make an approximation, where we drop from our equations all those terms which contain a product of two or more perturbations, as these are "higher-order" small. The resulting approximate theory is called *first-order perturbation theory* or *linear* perturbation theory. As the second name implies, the theory is now linear in the perturbations, meaning that the effect of overdensities cancel the effect of underdensities on, e.g., the average expansion rate. In this case the mean values evolve just like they would in the absence of perturbations.

While the perturbations at large scales have remained small, during the later history of the universe the perturbations have grown large at smaller scales. How big is the effect of backreaction, is an open research question in cosmology, since the calculations are difficult, but a common view is that the effect is small compared to the present accuracy of observations. For this course, we adopt this view, and assume that the background universe simultaneously represents an FRW universe ("the universe we would have if we did not have the perturbations") and the mean values of the quantities in the true universe at each time t.

Moreover, in Cosmology II we shall (mostly) assume that the background universe is flat ($K = 0$).

8.1.1 Statistical homogeneity and isotropy

We assume that the origin of the perturbations is some random process in the early universe. Thus over- ($\delta > 0$) and underdensities ($\delta < 0$) occur at randomly determined locations and we cannot expect to theoretically predict the values of $\delta(t, \mathbf{x})$ for particular locations $\mathbf{x}$. Instead, we can expect theory to predict statistical properties of the inhomogeneity field $\delta(t, \mathbf{x})$. The statistical properties are typically defined as averages of some quantities. We will deal with two kind of averages: *volume average* and *ensemble average*; the ensemble average is a theoretical concept, whereas the volume average is more observationally oriented.

We denote the volume average of some quantity $f(\mathbf{x})$ with the overbar, $\bar{f}$, and it is defined as

$$\bar{f} \equiv \frac{1}{V} \int_V d^3x f(\mathbf{x}) \,. \tag{4}$$

The integration volume V in question will depend on the situation.

For the ensemble average we assume that our universe is just one of an *ensemble* of an infinite number of possible universes (*realizations* of the random process) that could have resulted from the random process producing the initial perturbations. To know the random process, means to know the probability distribution $\mathrm{Prob}(\gamma)$ of the quantities γ produced by it. (At this stage we use the abstract notation of γ to denote the infinite number of these quantities. They could be the generated initial density perturbations at all locations, $\delta(\mathbf{x})$, or the corresponding Fourier coefficients $\delta_{\mathbf{k}}$. We will be more explicit later.) The ensemble average of a quantity f depending on these quantities γ as $f(\gamma)$ is denoted by $\langle f \rangle$ and defined as the (possibly infinite-dimensional) integral

$$\langle f \rangle \equiv \int d\gamma \mathrm{Prob}(\gamma) f(\gamma) \,. \tag{5}$$

Here f could be, e.g., the value of $\rho(\mathbf{x})$ at some location $\mathbf{x}$. The ensemble average is also called the *expectation value*. Thus the ensemble represents a probability distribution of universes. A

cosmological theory predicts such a probability distribution, but it does not predict in which realization from this distribution we live in. Thus the theoretical properties of the universe we will discuss (e.g., statistical homogeneity and isotropy, and ergodicity, see below) will be properties of this ensemble.

We now make the assumption that, although the universe is inhomogeneous, it is *statistically homogeneous and isotropic*. This is the second version of the *Cosmological Principle*. Statistical homogeneity means that the expectation value $\langle f(\mathbf{x})\rangle$ must be the same at all $\mathbf{x}$, and thus we can write it as $\langle f\rangle$. Statistical isotropy means that for quantities which involve a direction, the statistical properties are independent of the direction. For example, for vector quantities $\mathbf{v}$, all directions must be equally probable. This implies that $\langle \mathbf{v}\rangle = 0$. The assumption of statistical homogeneity and isotropy is justified by inflation: inflation makes the background universe homogeneous and isotropic so that the external conditions for quantum fluctuations are everywhere the same.

If the theoretical properties of the universe are those of an ensemble, and we can only observe one universe from that ensemble, how can we compare theory and observation? It seems reasonable that the statistics we get by comparing different parts of the universe should be similar to the statistics of a given part of the universe over different realizations, i.e., that they provide a *fair sample* of the probability distribution. This is called *ergodicity*. Fields $f(\mathbf{x})$ that satisfy

$$\bar{f} = \langle f\rangle \tag{6}$$

for an infinite volume V (for $\bar{f}$) and an arbitrary location $\mathbf{x}$ (for $\langle f\rangle$) are called *ergodic*. We assume that cosmological perturbations are ergodic. The equality does not hold for a finite volume V; the difference is called *sample variance* or *cosmic variance*. The larger the volume, the smaller is the difference. Since cosmological theory predicts $\langle f\rangle$, whereas observations probe $\bar{f}$ for a limited volume, cosmic variance limits how accurately we can compare theory with observations.[1]

8.1.2 Density autocorrelation function

From ergodicity,

$$\langle\rho\rangle = \bar{\rho} \quad \Rightarrow \quad \langle\delta\rho\rangle = 0 \quad \text{and} \quad \langle\delta\rangle = 0\,. \tag{7}$$

Thus we cannot use $\langle\delta\rangle$ as a measure of the inhomogeneity. Instead we can use the square of δ, which is necessarily nonnegative everywhere, so it cannot average out like δ did. Its expectation value

$$\langle\delta^2\rangle = \frac{\langle\delta\rho^2\rangle}{\bar{\rho}^2} \tag{8}$$

is the *variance* of the density perturbation, and the square root of the variance,

$$\delta_{\rm rms} \equiv \sqrt{\langle\delta^2\rangle} \tag{9}$$

the *root-mean-square* (rms) density perturbation, is a typical expected absolute value of δ at an arbitrary location.[2] It tells us about how strong the inhomogeneity is, but nothing about the shapes or sizes of the inhomogeneities. To get more information, we introduce the correlation function ξ.

We define the *density (2-point) autocorrelation function* (often called just *correlation function*) as

$$\xi(\mathbf{x}_1,\mathbf{x}_2) \equiv \langle\delta(\mathbf{x}_1)\delta(\mathbf{x}_2)\rangle\,. \tag{10}$$

[1] Another notation I will use for volume average is $\widehat{f}$, for smaller volumes, e.g., the volume observed in a galaxy survey. I try to reserve $\bar{f}$ for situations where we can assume $\bar{f} = \langle f\rangle$, whereas cosmic variance is the difference between $\widehat{f}$ and $\langle f\rangle$.

[2] In other words, $\delta_{\rm rms}$ is the standard deviation of $\rho/\bar{\rho}$.

It is positive if the density perturbation is expected to have the same sign at both $\mathbf{x}_1$ and $\mathbf{x}_2$, and negative for an overdensity at one and underdensity at the other. Thus it probes how density perturbations at different locations are correlated with each other. Due to statistical homogeneity, $\xi(\mathbf{x}_1, \mathbf{x}_2)$ can only depend on the difference $\mathbf{r} \equiv \mathbf{x}_2 - \mathbf{x}_1$, so we redefine ξ as

$$\xi(\mathbf{r}) \equiv \langle \delta(\mathbf{x}) \delta(\mathbf{x}+\mathbf{r}) \rangle \,. \tag{11}$$

From statistical isotropy, $\xi(\mathbf{r})$ is independent of direction, i.e., spherically symmetric (isotropic),

$$\xi(\mathbf{r}) = \xi(r) \,. \tag{12}$$

We will have use for both the 3D, $\xi(\mathbf{r})$, and 1D, $\xi(r)$, versions. The correlation function is large and positive for r smaller than the size of a typical over- or underdense region, and becomes small for larger distances.

The correlation function at zero separation gives the variance of the density perturbation,

$$\langle \delta^2 \rangle \equiv \langle \delta(\mathbf{x}) \delta(\mathbf{x}) \rangle \equiv \xi(0) \,. \tag{13}$$

We can also define a correlation function $\widehat{\xi}(\mathbf{r})$ for a single realization as a volume average,

$$\widehat{\xi}(\mathbf{r}) \equiv \frac{1}{V} \int d^3x \, \delta(\mathbf{x}) \delta(\mathbf{x}+\mathbf{r}) \,. \tag{14}$$

Integrating over $\mathbf{r}$ and assuming periodic boundary conditions[3] we get the *integral constraint*

$$\int d^3r \, \widehat{\xi}(\mathbf{r}) \;=\; \frac{1}{V} \int d^3r d^3x \, \delta(\mathbf{x}) \delta(\mathbf{x}+\mathbf{r}) \;=\; \frac{1}{V} \int d^3x \, \delta(\mathbf{x}) \int d^3r \, \delta(\mathbf{x}+\mathbf{r}) = 0 \,, \tag{16}$$

since the latter integral is $\bar{\delta} = 0$. Since $\xi(\mathbf{r}) = \langle \widehat{\xi}(\mathbf{r}) \rangle$ the integral constraint applies to it likewise. Therefore $\xi(r)$ must become negative at some point, so that at such a distance from an overdense region we are more likely to find an underdense region. Going to ever larger distances, ξ as a function of r may oscillate around zero, the oscillation becoming ever smaller in amplitude. Most of the interest in $\xi(r)$ is for the small r within the initial positive region.

8.1.3 Fourier space

The evolution of perturbations is best discussed in Fourier space. For mathematical convenience, we assume the observable part of the universe lies within a fiducial cubic box, volume $V = L^3$, with periodic boundary conditions. This box is assumed to be much larger than the region of interest, so that these boundary conditions should have no effect. Since the infinite universe is now periodic, the volume average over the infinite universe will be equal to the volume average over the fiducial box. Thus also the ergodicity assumption requires the fiducial volume to be large, so that it can provide a fair sample of the ensemble. We can now expand any function of space $f(\mathbf{x})$ as a Fourier series

$$f(\mathbf{x}) = \sum_{\mathbf{k}} f_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} \,, \tag{17}$$

[3]The other option is not to use periodic boundary conditions but to understand the integral in (14) to go over only those $\mathbf{x}$, for which both $\mathbf{x}$ and $\mathbf{x}+\mathbf{r} \in V$. This is what we have to do when V refers to an actual survey. The double integral in (16) then goes over all pairs $(\mathbf{x}, \mathbf{y})$ in V and can be written as

$$\frac{1}{V} \int_V d^3x \, \delta(\mathbf{x}) \int_V d^3y \, \delta(\mathbf{y}) = 0 \cdot 0 \,. \tag{15}$$

where the wave vectors $\mathbf{k} = (k_1, k_2, k_3)$ take values

$$k_i = n_i \frac{2\pi}{L}, \qquad n_i = 0, \pm 1, \pm 2, \ldots \tag{18}$$

The Fourier coefficients $f_\mathbf{k}$ are obtained as

$$f_\mathbf{k} = \frac{1}{V} \int_V f(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} d^3x \,. \tag{19}$$

If $f(\mathbf{x})$ is a perturbation so that its mean value vanishes, then the term $\mathbf{k} = 0$ does not occur. The Fourier coefficients are complex numbers even though we are dealing with real quantities $f(\mathbf{x})$. From the reality of $f(\mathbf{x})$ follows that

$$f_{-\mathbf{k}} = f_\mathbf{k}^* \,. \tag{20}$$

The Fourier expansion works only if the background universe is flat, although it can be used as an approximation in open and closed universes,[4] if the region of interest is much smaller than the curvature radius.

The separation of neighboring k_i values is $\Delta k_i = 2\pi/L$, so we can write

$$f(\mathbf{x}) \;=\; \sum_\mathbf{k} f_\mathbf{k} e^{i\mathbf{k}\cdot\mathbf{x}} \left(\frac{L}{2\pi}\right)^3 \Delta k_1 \Delta k_2 \Delta k_3 \;\approx\; \frac{1}{(2\pi)^3} \int f(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} d^3k \,, \tag{21}$$

where

$$f(\mathbf{k}) \equiv L^3 f_\mathbf{k} \,. \tag{22}$$

replacing the Fourier series with the Fourier integral. The size of the Fourier coefficients depends on the fiducial volume V – increasing V tends to make the $f_\mathbf{k}$ smaller to compensate for the denser sampling of $\mathbf{k}$ in Fourier space.

In the limit $V \to \infty$, the approximation in (21) becomes exact, and we have the *Fourier transform* pair

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{(2\pi)^3} \int f(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} d^3k \\ f(\mathbf{k}) &= \int f(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} d^3x \,. \end{aligned} \tag{23}$$

Note that this assumes that the integrals converge, which requires that $f(\mathbf{x}) \to 0$ for $|\mathbf{x}| \to \infty$. Thus we use only the Fourier series for, e.g., $\delta(\mathbf{x})$, but for, e.g., the correlation function $\xi(\mathbf{x})$ the Fourier transform is appropriate.

Even with a finite V we can use the Fourier integral as an approximation. Often it is conceptually simpler to work first with the Fourier series (so that one can, e.g., use the Kronecker delta $\delta_{\mathbf{k}\mathbf{k}'}$ instead of the Dirac delta function $\delta_D(\mathbf{k} - \mathbf{k}')$), replacing it with the integral in the end, when it needs to be calculated. The recipe for going from the series to the integral is

$$\begin{aligned} \left(\frac{2\pi}{L}\right)^3 \sum_\mathbf{k} &\;\to\; \int d^3k \\ L^3 f_\mathbf{k} &\;\to\; f(\mathbf{k}) \\ \left(\frac{L}{2\pi}\right)^3 \delta_{\mathbf{k}\mathbf{k}'} &\;\to\; \delta_D^3(\mathbf{k} - \mathbf{k}') \,. \end{aligned} \tag{24}$$

so that, e.g.,

$$\sum_\mathbf{k} f_\mathbf{k} \, e^{i\mathbf{k}\cdot\mathbf{x}} \;\to\; \frac{1}{(2\pi)^3} \int f(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} d^3k \,. \tag{25}$$

[4] An exact treatment in open and closed universes requires expansion in terms of suitable other functions instead of the plane waves $e^{i\mathbf{k}\cdot\mathbf{x}}$.

8.1.4 Power spectrum

We now expand the density perturbation as a Fourier series

$$\delta(\mathbf{x}) = \sum_{\mathbf{k}} \delta_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}}\,, \tag{26}$$

with

$$\delta_{\mathbf{k}} = \frac{1}{V}\int_V \delta(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} d^3x \tag{27}$$

and $\delta_{-\mathbf{k}} = \delta^*_{\mathbf{k}}$. Note that

$$\langle\delta(\mathbf{x})\rangle = 0 \quad \Rightarrow \quad \langle\delta_{\mathbf{k}}\rangle = 0\,. \tag{28}$$

In analogy with the correlation function $\xi(\mathbf{x},\mathbf{x}')$, we may ask what is the corresponding correlation in Fourier space, $\langle\delta^*_{\mathbf{k}}\delta_{\mathbf{k}'}\rangle$. Note that due to the mathematics of complex numbers, correlations of Fourier coefficients are defined with the complex conjugate *. This way the correlation of $\delta_{\mathbf{k}}$ with itself, $\langle\delta^*_{\mathbf{k}}\delta_{\mathbf{k}}\rangle = \langle|\delta_{\mathbf{k}}|^2\rangle$ is a real (and nonnegative) quantity, the expectation value of the absolute value (modulus) of $\delta_{\mathbf{k}}$ squared, i.e., the variance of $\delta_{\mathbf{k}}$. Calculating

$$\begin{aligned}
\langle\delta^*_{\mathbf{k}}\delta_{\mathbf{k}'}\rangle &= \frac{1}{V^2}\int d^3x e^{i\mathbf{k}\cdot\mathbf{x}}\int d^3x' e^{-i\mathbf{k}'\cdot\mathbf{x}'}\langle\delta(\mathbf{x})\delta(\mathbf{x}')\rangle \\
&= \frac{1}{V^2}\int d^3x e^{i\mathbf{k}\cdot\mathbf{x}}\int d^3r e^{-i\mathbf{k}'\cdot(\mathbf{x}+\mathbf{r})}\langle\delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle \\
&= \frac{1}{V^2}\int d^3r e^{-i\mathbf{k}'\cdot\mathbf{r}}\xi(\mathbf{r})\int d^3x e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{x}} \\
&= \frac{1}{V}\delta_{\mathbf{k}\mathbf{k}'}\int d^3r e^{-i\mathbf{k}\cdot\mathbf{r}}\xi(\mathbf{r}) \;\equiv\; \frac{1}{V}\delta_{\mathbf{k}\mathbf{k}'}P(\mathbf{k})\,,
\end{aligned} \tag{29}$$

where we used $\langle\delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle = \xi(\mathbf{r})$, i.e., independent of $\mathbf{x}$, which results from statistical homogeneity, and the orthogonality of plane waves

$$\boxed{\int d^3x e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{x}} \;=\; V\delta_{\mathbf{k}\mathbf{k}'} \;\to\; (2\pi)^3\delta_D^3(\mathbf{k}-\mathbf{k}')\,.} \tag{30}$$

Note that here $\delta_{\mathbf{k}\mathbf{k}'}$ is the Kronecker delta, 1 for $\mathbf{k}=\mathbf{k}'$, 0 otherwise – nothing to do with the density perturbation! In the limit $V\to\infty$ we get the Dirac delta function $\delta_D^3(\mathbf{k}-\mathbf{k}')$.

Written in terms of $\delta(\mathbf{k}) = V\delta_{\mathbf{k}}$, the result (29) reads as

$$\langle\delta(\mathbf{k})^*\delta(\mathbf{k}')\rangle \;=\; V\delta_{\mathbf{k}\mathbf{k}'}P(\mathbf{k}) \;\to\; (2\pi)^3\delta_D^3(\mathbf{k}-\mathbf{k}')P(\mathbf{k})\,, \tag{31}$$

Thus, *from statistical homogeneity follows that the Fourier coefficients $\delta_{\mathbf{k}}$ are uncorrelated.* The quantity

$$P(\mathbf{k}) \;\equiv\; V\langle|\delta_{\mathbf{k}}|^2\rangle \;=\; \int d^3r\, e^{-i\mathbf{k}\cdot\mathbf{r}}\xi(\mathbf{r})\,, \tag{32}$$

which gives the *variance* of $\delta_{\mathbf{k}}$, is called the *power spectrum* of $\delta(\mathbf{x})$. Since the correlation function $\to 0$ for large distances, we can replace the integration volume V in (32) with an infinite volume. We see that the power spectrum is the 3D Fourier transform of $\xi(\mathbf{r})$, and therefore also

$$\xi(\mathbf{r}) \;=\; \frac{1}{(2\pi)^3}\int d^3k\, e^{i\mathbf{k}\cdot\mathbf{r}}P(\mathbf{k})\,. \tag{33}$$

Unlike the correlation function, the power spectrum $P(\mathbf{k})$ is positive everywhere. Perturbations at large distance scales are more commonly discussed in terms of $P(\mathbf{k})$ than $\xi(\mathbf{r})$.

From statistical isotropy

$$\xi(\mathbf{r}) = \xi(r) \quad \Rightarrow \quad P(\mathbf{k}) = P(k) \tag{34}$$

(the 3D Fourier transform of a spherically symmetric function is also spherically symmetric), so that the variance of $\delta_{\mathbf{k}}$ depends only on the magnitude k of the wave vector $\mathbf{k}$, i.e., on the corresponding distance scale. Using spherical coordinates and doing the angular integrals we obtain (**exercise**) the relation between the 1D correlation function $\xi(r)$ and the 1D power spectrum $P(k)$,

$$\begin{aligned} P(k) &= \int_0^\infty \xi(r) \frac{\sin kr}{kr} 4\pi r^2 dr \\ \xi(r) &= \frac{1}{(2\pi)^3} \int_0^\infty P(k) \frac{\sin kr}{kr} 4\pi k^2 dk \,, \end{aligned} \tag{35}$$

For the density variance we get

$$\langle \delta^2 \rangle \equiv \xi(0) = \frac{1}{(2\pi)^3} \int_0^\infty P(k) 4\pi k^2 dk = \frac{1}{2\pi^2} \int_0^\infty k^3 P(k) d\ln k \equiv \int_{-\infty}^\infty \mathcal{P}(k) d\ln k \,. \tag{36}$$

where we have defined

$$\mathcal{P}(k) \equiv \frac{k^3}{2\pi^2} P(k) \,. \tag{37}$$

Another common notation for $\mathcal{P}(k)$ is $\Delta^2(k)$. The word "power spectrum" is used to refer to both $P(k)$ and $\mathcal{P}(k)$. Of these two, $\mathcal{P}(k)$ has the more obvious physical meaning: it gives the contribution of a logarithmic interval of scales, i.e., from k to ek, to the density variance. $\mathcal{P}(k)$ is dimensionless, whereas $P(k)$ has the dimension of Mpc^3 (when discussing observed values, it is usually given in units of $h^{-3}\mathrm{Mpc}^3$ as distance determinations are proportional to the Hubble constant).

Exercise: Define $\widehat{\xi}(\mathbf{r})$ as the volume average of $\delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})$, i.e., integrate $\mathbf{x}$ over the box V with periodic boundary conditions, and show that

$$\widehat{\xi}(\mathbf{r}) = \frac{V}{(2\pi)^3} \int d^3k \, |\delta_{\mathbf{k}}|^2 e^{i\mathbf{k}\cdot\mathbf{r}} \,, \tag{38}$$

for a single realization. Note that here we do not need any statistical assumptions (like statistical homogeneity or ergodicity). Contrast this result with (33).

8.1.5 Scales of interest and window functions

In (36) we integrated over all scales, from the infinitely large ($k = 0$ and $\ln k = -\infty$) to the infinitely small ($k = \infty$ and $\ln k = \infty$) to get the density variance. Perhaps this is not really what we want. The average matter density today is $3 \times 10^{-27}\,\mathrm{kg/m^3}$. The density of the Earth is $5.5 \times 10^3\,\mathrm{kg/m^3}$ and that of an atomic nucleus $2 \times 10^{17}\,\mathrm{kg/m^3}$, corresponding to $\delta \approx 2 \times 10^{30}$ and $\delta \approx 10^{44}$. Probing the density of the universe at such small scales finds a huge variance in it, but this is no longer the topic of cosmology - we are not interested here in planetary science or nuclear physics.

Even the study of the structure of individual galaxies is not considered to belong to cosmology, so the smallest (comoving) scale of cosmological interest, at least when we discuss the present universe,[5] is that of a typical separation between neighboring galaxies, of the order of 1 Mpc.

[5] In early universe cosmology we may study events, or possible events, related to also smaller comoving scales.

To exclude scales smaller than R ($r < R$ or $k > R^{-1}$) we *filter* the density field with a *window function*. This can be done in **k**-space or **x**-space.

The filtering in **x**-space is done by convolution. We introduce a (usually spherically symmetric) window function $W(\mathbf{r})$ such that

$$\int d^3r\, W(\mathbf{r}) = 1 \tag{39}$$

(normalization) and $W \sim 0$ for $|\mathbf{r}| \gg R$ and define the filtered density field

$$\delta(\mathbf{x}, R) \;\equiv\; (\delta * W)(\mathbf{x}) \;\equiv\; \int d^3\mathbf{x}'\, \delta(\mathbf{x}')W(\mathbf{x}' - \mathbf{x})\,. \tag{40}$$

The simplest window function is the top-hat window function

$$W_T(\mathbf{r}) \;\equiv\; \left(\frac{4\pi}{3}R^3\right)^{-1} \qquad \text{for} \quad |\mathbf{r}| \le R \tag{41}$$

and $W_T(\mathbf{r}) = 0$ elsewhere, i.e., $\delta(\mathbf{x})$ is filtered by replacing it with its mean value within the distance R. Mathematically more convenient is the Gaussian window function

$$W_G(\mathbf{r}) \;\equiv\; \frac{1}{(2\pi)^{3/2}R^3} e^{-\frac{1}{2}|\mathbf{r}|^2/R^2}\,. \tag{42}$$

By the convolution theorem, the filtering in Fourier space becomes just multiplication:

$$\delta(\mathbf{k}, R) \;=\; \delta(\mathbf{k})W(\mathbf{k})\,, \tag{43}$$

where $W(\mathbf{k})$ is the Fourier transform of the window function. For W_T and W_G we have (**exercise**)

$$\begin{aligned} W_T(\mathbf{k}) &= \frac{3(\sin kR - kR\cos kR)}{(kR)^3} \\ W_G(\mathbf{k}) &= e^{-\frac{1}{2}(kR)^2}\,. \end{aligned} \tag{44}$$

We can also define the **k**-space top-hat window function

$$W_k(\mathbf{k}) \;\equiv\; 1 \qquad \text{for} \quad k \le 1/R \tag{45}$$

and $W_k(\mathbf{k}) = 0$ elsewhere. In **x**-space this becomes (**exercise**)

$$W_k(\mathbf{r}) \;=\; \frac{1}{2\pi^2R^3}\frac{\sin y - y\cos y}{y^3}\,, \qquad \text{where} \quad y \equiv |\mathbf{r}|/R\,. \tag{46}$$

The variance of the filtered density field (**Exercise:** derive the second equalities of both expressions)

$$\begin{aligned} \sigma^2(R) &\equiv \langle\delta(\mathbf{x}, R)^2\rangle \;=\; \frac{1}{(2\pi)^3}\int d^3k\, P(k)|W(\mathbf{k})|^2 \\ \widehat{\sigma}^2(R) &\equiv \frac{1}{V}\int d^3x\, \delta(\mathbf{x}, R)^2 \;=\; \frac{V}{(2\pi)^3}\int d^3k\, |\delta_{\mathbf{k}}|^2|W(\mathbf{k})|^2\,. \end{aligned} \tag{47}$$

is a measure of the inhomogeneity at scale R. For the **k**-space top-hat window this becomes simply

$$\sigma^2(R) \;=\; \frac{1}{(2\pi)^3}\int_0^{R^{-1}} 4\pi k^2 P(k)dk \;=\; \int_{-\infty}^{-\ln R} \mathcal{P}(k)d\ln k\,. \tag{48}$$

One may also ask, whether scales larger than the observed universe (the lower limit $k = 0$ or $\ln k = -\infty$ in the k integrals) are relevant, since we cannot observe the inhomogeneity at such scales. Due to such very-large-scale inhomogeneities, the average density in the observed universe may deviate from the average density of the entire universe. Inhomogeneities at scales somewhat larger than the observed universe could appear as an anisotropy in the observed universe. The importance of such large scales depends on how strong the inhomogeneities at these scales are, i.e., how the power spectrum behaves as $k \to 0$. The present understanding, supported by observations, is that the contribution of such large scales is small.

8.1.6 Power-law spectra

We have observational information and theoretical predictions for $\xi(r)$ and $P(k)$ for a wide range of scales. (We will discuss the theory in detail later.) For certain intervals, they can be approximated by a power-law form,

$$\xi(r) \propto r^{-\gamma} \qquad \text{or} \qquad P(k) \propto k^n \,. \tag{49}$$

When plotted on a log-log scale, such functions appear as straight lines with slope $-\gamma$ and n. The proportionality constant can be given in terms of a reference scale. For $\xi(r)$ we usually choose the scale r_0 where $\xi(r_0) = 1$, so that

$$\xi(r) \;=\; \left(\frac{r}{r_0}\right)^{-\gamma} . \tag{50}$$

For $P(k)$ we may write

$$P(k) \;=\; A^2 \left(\frac{k}{k_p}\right)^n \qquad \text{or} \qquad \mathcal{P}(k) \;=\; A^2 \left(\frac{k}{k_p}\right)^{n+3} , \tag{51}$$

where k_p is called a *pivot scale* (whose choice depends on the application) and $A \equiv \sqrt{P(k_p)}$ or $\sqrt{\mathcal{P}(k_p)}$ is the amplitude of the power spectrum at the pivot scale.

We define the spectral index $n(k)$ as

$$n(k) \;\equiv\; \frac{d\ln P}{d\ln k} \,. \tag{52}$$

It gives the slope of $P(k)$ on a log-log plot. For a power-law $P(k)$, $n(k) = const = n$. We can study power-law $\xi(r)$ and $P(k)$ as a playground to get a feeling what different values of the spectral index mean, and, e.g., how γ and n are related.[6]

The Fourier transform of a power law is a power law. For the correlation function of (50) we get (**exercise**)

$$\begin{aligned} P(k) &= \frac{4\pi}{k^3}\Gamma(2-\gamma)\sin\frac{(2-\gamma)\pi}{2}\,(kr_0)^\gamma \\ \mathcal{P}(k) &= \frac{2}{\pi}\Gamma(2-\gamma)\sin\frac{(2-\gamma)\pi}{2}\,(kr_0)^\gamma \end{aligned} \tag{53}$$

for $1 < \gamma < 2$ or $2 < \gamma < 3$, and

$$\begin{aligned} P(k) &= \frac{2\pi^2}{k^3}(kr_0)^2 \\ \mathcal{P}(k) &= (kr_0)^2 \end{aligned} \tag{54}$$

[6] In reality the spectral index is very different at small scales than at large scales. Observationally, for small scales, $\gamma \sim 1.8$, and for large scales, $n \sim 1$. We discuss this later.

for $\gamma = 2$. Thus

$$n \;=\; \gamma - 3 \qquad \text{for } 1 < \gamma < 3, \text{ i.e., } -2 < n < 0\,. \tag{55}$$

The variance

$$\langle \delta^2 \rangle = \xi(0) = \int_0^\infty \mathcal{P}(k) \frac{dk}{k} \propto \int_0^\infty k^{n+2} dk = \frac{1}{n+3} \left[k^{n+3} \right]_0^\infty \qquad \text{for} \quad n \neq -3 \tag{56}$$

diverges at small scales (high k) for $n \geq -3$ and at large scales (low k) for $n \leq -3$. We cure the small scale divergence with filtering as discussed in Sec. 8.1.5.

Exercise: For a power-law spectrum and a Gaussian window function, show that

$$\sigma^2(R) \;=\; \frac{1}{2} \Gamma \left(\frac{n+3}{2} \right) \mathcal{P}(R^{-1})\,. \tag{57}$$

8.1.7 Galaxy 2-point correlation function

The most obvious way to try to measure the cosmological density perturbations is to observe the spatial distribution of galaxies. We treat individual galaxies as mathematical points, so that each galaxy has a comoving coordinate value $\mathbf{x}$. We define the *galaxy 2-point correlation function* $\xi_g(\mathbf{r})$ as the *excess probability* of finding a galaxy at separation $\mathbf{r}$ from another galaxy:

$$dP \equiv \bar{n} \left[1 + \xi_g(\mathbf{r}) \right] dV \tag{58}$$

where $\bar{n}$ is the mean galaxy number density, dV is a volume element that is a separation $\mathbf{r}$ away from a chosen reference galaxy, and dP is the probability that there is a galaxy within dV. (Here dV is assumed so small that there is at most one galaxy in it.)

If the galaxy number density $n(\mathbf{x})$ faithfully traces the underlying matter density, so that

$$\delta_g \equiv \frac{\delta n}{\bar{n}} = \delta \equiv \frac{\delta \rho_m}{\bar{\rho}_m}\,, \tag{59}$$

then ξ_g becomes equal to the matter density autocorrelation function ξ: The probability of finding a galaxy in volume dV_1 at a random location $\mathbf{x}$ is

$$dP_1 \;=\; \langle n(\mathbf{x}) \rangle dV_1 \;=\; \langle \bar{n} + \delta n(\mathbf{x}) \rangle dV_1 \;=\; \bar{n} dV_1\,. \tag{60}$$

The probability of finding a galaxy pair at $\mathbf{x}$ and $\mathbf{x} + \mathbf{r}$ is

$$\begin{aligned} dP_{12} \;&=\; \langle n(\mathbf{x}) n(\mathbf{x}+\mathbf{r}) \rangle dV_1 dV_2 \;=\; \bar{n}^2 \langle [1 + \delta(\mathbf{x})][1 + \delta(\mathbf{x}+\mathbf{r})] \rangle dV_1 dV_2 \\ &=\; \bar{n}^2 \left[1 + \langle \delta(\mathbf{x}) \rangle + \langle \delta(\mathbf{x}+\mathbf{r}) \rangle + \langle \delta(\mathbf{x}) \delta(\mathbf{x}+\mathbf{r}) \rangle \right] dV_1 dV_2 \\ &=\; \bar{n}^2 \left[1 + \langle \delta(\mathbf{x}) \delta(\mathbf{x}+\mathbf{r}) \rangle \right] dV_1 dV_2\,, \end{aligned} \tag{61}$$

since $\langle \delta(\mathbf{x}) \rangle = \langle \delta(\mathbf{x}+\mathbf{r}) \rangle = 0$. Dividing dP_{12} with dP_1 we get the probability dP_2 of finding the second galaxy once we have found the first one

$$dP_2 \;=\; \bar{n} \left[1 + \langle \delta(\mathbf{x}) \delta(\mathbf{x}+\mathbf{r}) \rangle \right] dV_2 \;=\; \bar{n} \left[1 + \xi(\mathbf{r}) \right] dV_2\,. \tag{62}$$

Thus $\xi_g = \xi$.

It is probable that the galaxy number density does not trace the matter density faithfully, since galaxy formation is likely to be more efficient in high-density regions. This is called *bias*. Specifically the bias, or *galaxy bias* b_g, is defined as the ratio

$$b_g \equiv \frac{\delta_g}{\delta_m} \quad \Rightarrow \quad \xi_g = b_g^2 \xi\,, \tag{63}$$

where the expectation is that $b_g > 1$. In principle the bias could depend on the scale k, the time t (or redshift z), and/or the strength of the density perturbation δ_m. The simplest treatment of bias is to assume b_g is a constant over the observationally relevant ranges of these quantities.

The bias will depend on the type of tracer (all galaxies, specific types of galaxies, galaxy clusters) and is typically larger for more massive objects.

8.2 Newtonian perturbation theory

We shall now study the evolution of perturbations during the history of the universe. Initially the perturbations were small and we restrict the quantitative treatment to that part of the evolution when they remained small (for large scales, this extends to the present time and the future). This allows us to use *first-order perturbation theory*, where we drop from our equations all those terms which contain a product of two or more perturbations (as these products are even smaller). The remaining equations will then contain only terms which are either *zeroth order*, i.e., contain only background quantities, or *first order*, i.e., contain exactly one perturbation. If we kept only the zeroth order parts, we would be back to the equations of the homogenous and isotropic universe. Subtracting these from our equations we arrive at the *perturbation equations* where every term is first-order in the perturbation quantities, i.e., it is a *linear equation* for them. This makes the equations easy to handle, we can, e.g., Fourier transform them.

As we discovered in our discussion of inflation, the different cosmological distance scales first exit the horizon during inflation, then enter the horizon during various epochs of the later history. Matter perturbations at subhorizon scales, i.e., after horizon entry, can be treated with *Newtonian perturbation theory*, but scales which are close to horizon size or superhorizon require *relativistic perturbation theory*, which is based on general relativity.

The Newtonian equations for (perfect gas)[7] fluid dynamics with gravity are

$$\frac{\partial \rho}{\partial t'} + \nabla_{\mathbf{r}} \cdot (\rho \mathbf{u}) = 0 \tag{64}$$

$$\frac{\partial \mathbf{u}}{\partial t'} + (\mathbf{u} \cdot \nabla_{\mathbf{r}})\mathbf{u} + \frac{1}{\rho}\nabla_{\mathbf{r}} p + \nabla_{\mathbf{r}} \tilde{\Phi} = 0 \tag{65}$$

$$\nabla^2_{\mathbf{r}} \tilde{\Phi} = 4\pi G \rho \tag{66}$$

Here ρ is the mass density, p is the pressure, and $\mathbf{u}$ is the flow velocity of the fluid. We write $\tilde{\Phi}$ for the Newtonian gravitational potential, since we want to reserve Φ for its perturbation. The subscript $\mathbf{r}$ in $\nabla_{\mathbf{r}}$ emphasizes that the space derivatives are taken with respect to the Newtonian space coordinate $\mathbf{r}$ (instead of a comoving coordinate). Although the Newtonian time coordinate t' is equal to the cosmic time coordinate t, we need to make a distinction between t' and t in partial derivatives as will become clear soon.

The first equation is the law of mass conservation. The second equation is called the *Euler equation*, and it is just "$F = ma$" for a fluid element, whose mass is ρdV. Here the acceleration of a fluid element is not given by $\partial \mathbf{u}/\partial t'$ which just tells how the velocity field changes at a given position, but by $d\mathbf{u}/dt'$, where

$$\frac{d}{dt'} \equiv \frac{\partial}{\partial t'} + (\mathbf{u} \cdot \nabla_{\mathbf{r}}) \tag{67}$$

is the *convective time derivative*, which follows the fluid element as it moves. The two other terms give the forces due to pressure gradient and gravitational field.

We can apply Newtonian physics if:

1) Distance scales considered are $\ll$ the scale of curvature of spacetime (given by the Hubble length in cosmology[8])

2) The fluid flow is nonrelativistic, $u \ll c \equiv 1$.

3) We are considering nonrelativistic matter, $|p| \ll \rho$

[7] perfect gas = no internal friction $\Rightarrow$ pressure is isotropic

[8] As discussed in Chapter 3, the spacetime curvature has two distance scales, the Hubble length H^{-1} and the curvature radius $R_{\rm curv} \equiv a|K|^{-1/2}$. From observations we know that the curvature radius is larger than the Hubble length (at all times of interest), possibly infinite.

The last condition corresponds to particle velocities being nonrelativistic, if the matter is made out of particles. Although the pressure is small compared to mass density, the pressure gradient can be important if the pressure varies at small scales.

Note: Energy density and mass density. In Newtonian gravity, the source of gravity is mass density ρ_m, not energy density ρ. For nonrelativistic matter, the kinetic energies of particles are negligible compared to their masses, and thus so is the energy density compared to mass density, if we don't count the rest energy in it. The Newtonian equations for mass density and energy density are

$$\frac{\partial \rho_m}{\partial t'} + \nabla_{\mathbf{r}} \cdot (\rho_m \mathbf{u}) = 0 \tag{68}$$

$$\frac{\partial \rho_u}{\partial t'} + \nabla_{\mathbf{r}} \cdot (\rho_u \mathbf{u}) + p \nabla_{\mathbf{r}} \cdot \mathbf{u} = 0\,, \tag{69}$$

where $\nabla_{\mathbf{r}} \cdot \mathbf{u}$ gives the rate of change in the volume of the fluid element and $p\nabla_{\mathbf{r}} \cdot \mathbf{u}$ is the work done by pressure. In Newtonian physics, rest energy (mass) is not included in the energy density. Eq. (69) applies whether we include it or not. Define total energy density as

$$\rho \equiv \rho_m + \rho_u\,,$$

where ρ_u is the Newtonian energy density and ρ_m is the mass density. Adding Eqs. (68) and (69) gives

$$\frac{\partial \rho}{\partial t'} + \nabla_{\mathbf{r}} \cdot (\rho \mathbf{u}) + p \nabla_{\mathbf{r}} \cdot \mathbf{u} = 0\,. \tag{70}$$

For nonrelativistic matter $\rho_u \ll \rho_m$ and $p \ll \rho_m$. We can thus drop the last term in (70) and ignore the distinction between mass density and total energy density.

A homogeneously expanding fluid,

$$\rho = \rho(t_0) a^{-3} \tag{71}$$

$$\mathbf{u} = \frac{\dot{a}}{a}\mathbf{r} \tag{72}$$

$$\tilde{\Phi} = \frac{2\pi G}{3}\rho r^2 \tag{73}$$

is a solution to these equations (**exercise**), with a condition to the function $a(t)$ giving the expansion law. It is the Newtonian version of the matter-dominated Friedmann model. Writing $H(t) \equiv \dot{a}/a$ we find that the homogeneous solution satisfies

$$\dot{\rho} + 3H\rho = 0\,, \tag{74}$$

and the condition for $a(t)$ (from the exercise) can be written as

$$\frac{\ddot{a}}{a} = \dot{H} + H^2 = -\frac{4\pi G}{3}\rho\,. \tag{75}$$

You should recognize these equations as the energy-continuity equation and the second Friedmann equation for a matter-dominated FRW universe.[9] The result for $\tilde{\Phi}$, Eq. (73), has no relativistic counterpart, the whole concept of gravitational potential does not exist in relativity (except in special cases; like here in perturbation theory, where we introduce potentials related to perturbations).

[9] The freedom of choosing the initial value of the expansion rate leaves the connection between H and ρ open up to a constant. This constant has the same effect on the time evolution of $a(t)$ as the curvature constant K in the first Friedmann equation, but of course in the Newtonian treatment it is not interpreted as curvature, and it does not otherwise have the same physical effects. We shall (unless otherwise noted) choose this constant so that the background solution matches the *flat* FRW universe. Then we have

$$H^2 = \frac{8\pi G}{3}\rho \qquad \text{or} \qquad 4\pi G\rho = \frac{3}{2}H^2\,. \tag{76}$$

8.2.1 Comoving coordinates

Introduce now a new (comoving) coordinate system $(t, \mathbf{x})$ which is related to the Newtonian coordinate system $(t', \mathbf{r})$ by

$$t' = t \qquad \mathbf{r} = a(t)\mathbf{x}\,. \tag{77}$$

Thus the time coordinate is the same in both coordinate systems, but we need to distinguish between the partial derivatives $\partial/\partial t$ and $\partial/\partial t'$, since in the first $\mathbf{x}$ is kept constant and in the second $\mathbf{r}$ is kept constant. Relate now the partial derivatives:

$$\begin{aligned}
\frac{\partial}{\partial t} &= \frac{\partial t'}{\partial t}\frac{\partial}{\partial t'} + \sum_i \frac{\partial r_i}{\partial t}\frac{\partial}{\partial r_i} = \frac{\partial}{\partial t'} + \sum \dot{a} x_i \frac{\partial}{\partial r_i} = \frac{\partial}{\partial t'} + H\mathbf{r}\cdot\nabla_{\mathbf{r}} \\
\frac{\partial}{\partial x_i} &= \frac{\partial t'}{\partial x_i}\frac{\partial}{\partial t'} + \sum_j \frac{\partial r_j}{\partial x_i}\frac{\partial}{\partial r_j} = \sum_j \delta_{ij} a \frac{\partial}{\partial r_j} = a\frac{\partial}{\partial r_i} \quad \Rightarrow \quad \nabla_{\mathbf{x}} = a\nabla_{\mathbf{r}}\,.
\end{aligned} \tag{78}$$

Thus

$$\frac{\partial}{\partial t'} = \frac{\partial}{\partial t} - H\mathbf{x}\cdot\nabla_{\mathbf{x}} \qquad \text{and} \qquad \nabla_{\mathbf{r}} = \frac{1}{a}\nabla_{\mathbf{x}}\,. \tag{79}$$

(Later we will work exclusively in the comoving coordinates and write just ∇ for $\nabla_{\mathbf{x}}$. The "original" coordinates $\mathbf{r}$ are just an artifact of the Newtonian approach and do not appear in relativistic perturbation theory.)

8.2.2 The perturbation

Now, consider a small perturbation, so that

$$\rho(t', \mathbf{r}) = \bar{\rho}(t) + \delta\rho(t', \mathbf{r}) \tag{80}$$

$$p(t', \mathbf{r}) = \bar{p}(t) + \delta p(t', \mathbf{r}) \tag{81}$$

$$\mathbf{u}(t', \mathbf{r}) = H(t)\mathbf{r} + \mathbf{v}(t', \mathbf{r}) \tag{82}$$

$$\tilde{\Phi}(t', \mathbf{r}) = \frac{2\pi G}{3}\bar{\rho}r^2 + \Phi(t', \mathbf{r})\,, \tag{83}$$

where $\bar{\rho}$, $\bar{p}$, and H denote homogeneous background quantities (solutions of the background, or zeroth-order, equations) and $\delta\rho$, δp, $\mathbf{v}$, Φ are small inhomogeneous perturbations.

Inserting these into the Eqs. (64,65,66) and subtracting the homogeneous equations (73,74,75) we get (**exercise**) the *perturbation equations*

$$\frac{\partial\delta\rho}{\partial t'} + 3H\delta\rho + H\mathbf{r}\cdot\nabla_{\mathbf{r}}\delta\rho + \bar{\rho}\nabla_{\mathbf{r}}\cdot\mathbf{v} = 0 \tag{84}$$

$$\frac{\partial\mathbf{v}}{\partial t'} + H\mathbf{v} + H\mathbf{r}\cdot\nabla_{\mathbf{r}}\mathbf{v} + \frac{1}{\bar{\rho}}\nabla_{\mathbf{r}}\delta p + \nabla_{\mathbf{r}}\Phi = 0 \tag{85}$$

$$\nabla^2_{\mathbf{r}}\Phi = 4\pi G\delta\rho\,. \tag{86}$$

In terms of the comoving coordinates these become (**exercise**):

$$\frac{\partial\delta\rho}{\partial t} + 3H\delta\rho + \frac{\bar{\rho}}{a}\nabla_{\mathbf{x}}\cdot\mathbf{v} = 0 \tag{87}$$

$$\frac{\partial\mathbf{v}}{\partial t} + H\mathbf{v} + \frac{1}{a\bar{\rho}}\nabla_{\mathbf{x}}\delta p + \frac{1}{a}\nabla_{\mathbf{x}}\Phi = 0 \tag{88}$$

$$\nabla^2_{\mathbf{x}}\Phi = 4\pi G a^2\delta\rho\,. \tag{89}$$

In terms of the relative density perturbation $\delta \equiv \delta\rho/\bar{\rho}$ we have $\delta\rho = \bar{\rho}\cdot\delta$ and

$$\frac{\partial\delta\rho}{\partial t} = \dot{\bar{\rho}}\cdot\delta + \bar{\rho}\frac{\partial\delta}{\partial t} \qquad \text{where} \qquad \dot{\bar{\rho}}\cdot\delta = -3H\bar{\rho}\,\delta\,, \tag{90}$$

and we can write

$$\frac{\partial \mathbf{v}}{\partial t} + H\mathbf{v} = \frac{1}{a}\frac{\partial}{\partial t}(a\mathbf{v}) \tag{91}$$

so that the set of perturbation equations becomes

$$\frac{\partial \delta}{\partial t} + \frac{1}{a}\nabla_{\mathbf{x}} \cdot \mathbf{v} = 0 \tag{92}$$

$$\frac{\partial}{\partial t}(a\mathbf{v}) + \frac{1}{\bar{\rho}}\nabla_{\mathbf{x}}\delta\rho + \nabla_{\mathbf{x}}\Phi = 0 \tag{93}$$

$$\nabla^2_{\mathbf{x}}\Phi = 4\pi G a^2 \bar{\rho}\delta \tag{94}$$

Finally, we Fourier expand the perturbations,

$$\delta(t, \mathbf{x}) = \sum_{\mathbf{k}} \delta_{\mathbf{k}}(t) e^{i\mathbf{k}\cdot\mathbf{x}} \qquad \text{etc.} \tag{95}$$

In Fourier space the perturbation equations become

$$\dot{\delta}_{\mathbf{k}} + \frac{i\mathbf{k}\cdot\mathbf{v}_{\mathbf{k}}}{a} = 0 \tag{96}$$

$$\frac{d}{dt}(a\mathbf{v}_{\mathbf{k}}) + i\mathbf{k}\frac{\delta p_{\mathbf{k}}}{\bar{\rho}} + i\mathbf{k}\Phi_{\mathbf{k}} = 0 \tag{97}$$

$$\Phi_{\mathbf{k}} = -4\pi G\left(\frac{a}{k}\right)^2 \bar{\rho}\delta_{\mathbf{k}}\,. \tag{98}$$

Solving the evolution of the perturbations is a two-step process:

1) Solve the background equations to obtain the functions $a(t)$, $H(t)$, and $\bar{\rho}(t)$. After this, these are *known functions* in the perturbation equations.

2) Solve the perturbation equations.

8.2.3 Vector and scalar perturbations

We now divide the velocity perturbation field $\mathbf{v}(t, \mathbf{r})$ into its rotational (solenoidal, divergence-free) and irrotational (curl-free) parts,

$$\mathbf{v} = \mathbf{v}_\perp + \mathbf{v}_\parallel\,, \tag{99}$$

where $\nabla \cdot \mathbf{v}_\perp = 0$ and $\nabla \times \mathbf{v}_\parallel = 0$. For Fourier components this simply means that $\mathbf{k} \cdot \mathbf{v}_{\perp\mathbf{k}} = 0$ and $\mathbf{k} \times \mathbf{v}_{\parallel\mathbf{k}} = 0$. That is, we divide $\mathbf{v}_{\mathbf{k}}$ into the components perpendicular and parallel to the wave vector $\mathbf{k}$. The parallel part we can write in terms of a scalar function v, whose Fourier components $v_{\mathbf{k}}$ are given by

$$\mathbf{v}_{\parallel\mathbf{k}} \equiv v_{\mathbf{k}}\hat{\mathbf{k}}\,, \tag{100}$$

where $\hat{\mathbf{k}}$ denotes the unit vector in the $\mathbf{k}$ direction.

We can now take the perpendicular and parallel parts of Eq. (97),

$$\frac{d}{dt}(a\mathbf{v}_{\perp\mathbf{k}}) = 0 \tag{101}$$

$$\frac{d}{dt}(av_{\mathbf{k}}) + ik\frac{\delta p_{\mathbf{k}}}{\bar{\rho}} + ik\Phi_{\mathbf{k}} = 0\,. \tag{102}$$

We see that the rotational part of the velocity perturbation has a simple time evolution,

$$\mathbf{v}_\perp \propto a^{-1}\,, \tag{103}$$

i.e., it *decays* from whatever initial value it had, inversely proportional to the scale factor.

The other perturbation equations involve only the irrotational part of the velocity perturbation. Thus we can divide the total perturbation into two parts, commonly called the *vector* and *scalar* perturbations, which evolve independent of each other:

1) The vector perturbation: $\mathbf{v}_\perp$.

2) The scalar perturbation: δ, δp, v, Φ, which are all coupled to each other.

The vector perturbations are thus not related to the density perturbations, or the structure of the universe. Also, any primordial vector perturbation should become rather small as the universe expands, at least while first-order perturbation theory applies.[10] They are thus not very important, and we shall have no more to say about them. The rest of our discussion focuses on the scalar perturbations.

8.2.4 The equations for scalar perturbations

We summarize here the equations for scalar perturbations:

$$\dot{\delta}_{\mathbf{k}} + \frac{ikv_{\mathbf{k}}}{a} = 0 \qquad \Rightarrow \qquad v_{\mathbf{k}} = i\frac{a}{k}\dot{\delta}_{\mathbf{k}} \tag{104}$$

$$\frac{d}{dt}(av_{\mathbf{k}}) + ik\frac{\delta p_{\mathbf{k}}}{\bar{\rho}} + ik\Phi_{\mathbf{k}} = 0 \tag{105}$$

$$\Phi_{\mathbf{k}} = -4\pi G\left(\frac{a}{k}\right)^2 \bar{\rho}\delta_{\mathbf{k}}\,. \tag{106}$$

Inserting $v_{\mathbf{k}}$ from (104) and $\Phi_{\mathbf{k}}$ from (106) into (105) we get

$$\boxed{\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} = -\frac{k^2}{a^2}\frac{\delta p_{\mathbf{k}}}{\bar{\rho}} + 4\pi G\bar{\rho}\delta_{\mathbf{k}}\,.} \tag{107}$$

8.2.5 Adiabatic and entropy perturbations

Suppose the equation of state is *barotropic*,

$$p = p(\rho) \tag{108}$$

i.e., pressure is uniquely determined by the energy density. Then the perturbations δp and $\delta\rho$ are necessarily related by the derivative $dp/d\rho$ of this function $p(\rho)$,

$$p = \bar{p} + \delta p = \bar{p}(\bar{\rho}) + \frac{dp}{d\rho}(\bar{\rho})\delta\rho \quad \Rightarrow \quad \delta p = \frac{dp}{d\rho}\delta\rho\,.$$

The time derivatives of the background quantities $\bar{p}$ and $\bar{\rho}$ are related by this same derivative,

$$\dot{\bar{p}} = \frac{d\bar{p}}{dt} = \frac{dp}{d\rho}(\bar{\rho})\frac{d\bar{\rho}}{dt} = \frac{dp}{d\rho}\dot{\bar{\rho}}\,.$$

Assuming this derivative $dp/d\rho$ is nonnegative, we call its square root the *speed of sound*

$$c_s \equiv \sqrt{\frac{dp}{d\rho}}\,. \tag{109}$$

[10] Thus we end up with an irrotational velocity field. The rotational motion (e.g., rotation of galaxies) which is common in the present universe at small scales has arisen from higher-order effects from the primordial scalar perturbations, not from the primordial vector perturbations.

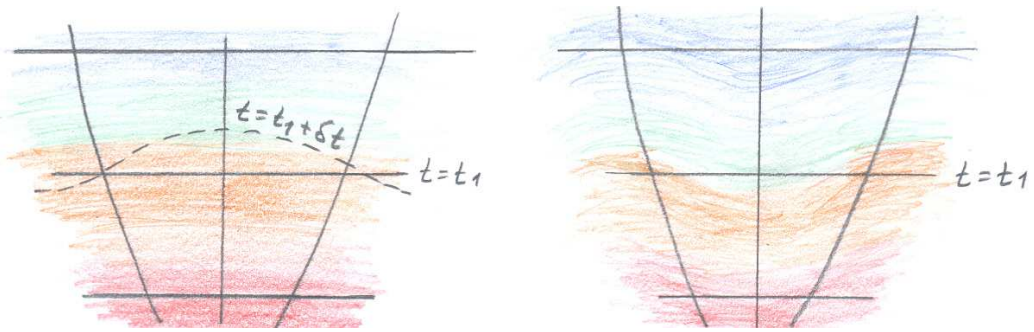


Figure 1: For adiabatic perturbations, the conditions in the perturbed universe (right) at $(t_1, \mathbf{x})$ equal conditions in the (homogeneous) background universe (left) at some time $t_1 + \delta t(\mathbf{x})$.

(We shall indeed find that sound waves propagate at this speed.) We thus have the relation

$$\frac{\delta p}{\delta \rho} = \frac{\dot{\bar{p}}}{\dot{\bar{\rho}}} = c_s^2 \,.$$

In general, when p may depend on other variables besides ρ, the speed of sound in a fluid is given by

$$c_s^2 = \left(\frac{\partial p}{\partial \rho}\right)_S \tag{110}$$

where the subscript S indicates that the derivative is taken so that the entropy of the fluid element is kept constant. Since the background universe expands adiabatically (meaning that there is no entropy production), we have that

$$\frac{\dot{\bar{p}}}{\dot{\bar{\rho}}} = \left(\frac{\partial p}{\partial \rho}\right)_S = c_s^2 \,. \tag{111}$$

Perturbations with the property

$$\frac{\delta p}{\delta \rho} = \frac{\dot{\bar{p}}}{\dot{\bar{\rho}}} \tag{112}$$

are called *adiabatic perturbations* in cosmology.

If $p = p(\rho)$, perturbations are necessarily adiabatic. In the general case the perturbations may or may not be adiabatic. In the latter case, the perturbation can be divided into an adiabatic component and an *entropy perturbation*. An entropy perturbation is a perturbation in the entropy-per-particle ratio.

For adiabatic perturbations we thus have

$$\delta p = c_s^2 \delta \rho = \frac{\dot{\bar{p}}}{\dot{\bar{\rho}}} \delta \rho \,. \tag{113}$$

Adiabatic perturbations have the property that the local state of matter (determined here by the quantities p and ρ) at some spacetime point $(t, \mathbf{x})$ of the perturbed universe is the same as in the background universe at some slightly different time $t + \delta t$, this time difference being different for different locations $\mathbf{x}$. See Fig. 1.

Thus we can view adiabatic perturbations as some parts of the universe being "ahead" and others "behind" in the evolution.

Adiabatic perturbations are the simplest kind of perturbations. Single-field inflation produces adiabatic perturbations, since perturbations in all quantities are proportional to a perturbation $\delta\varphi$ in a single scalar quantity, the inflaton field.

Adiabatic perturbations stay adiabatic while they are outside horizon, but may develop entropy perturbations when they enter the horizon. This happens for many-component fluids (discussed a little later).

Present observational data is consistent with the primordial (i.e., before horizon entry) perturbations being adiabatic.

8.2.6 Adiabatic perturbations in matter

Consider now adiabatic perturbations of a non-relativistic single-component fluid. The equation for the density perturbation is now

$$\boxed{\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} + \left[\frac{c_s^2 k^2}{a^2} - 4\pi G\bar{\rho}\right]\delta_{\mathbf{k}} = 0\,.} \tag{114}$$

I shall call this the *Jeans equation*[11] (although Jeans considered a static, not an expanding fluid).

This is a second-order differential equation from which we can solve the time evolution of the Fourier amplitudes $\delta_{\mathbf{k}}(t)$ of the perturbation. Before solving this equation we need to first find the background solution which gives the functions $a(t)$, $H(t) = \dot{a}/a$, and $\bar{\rho}(t)$.

The nature of the solution to Eq. (114) depends on the sign of the factor in the brackets. The first term in the brackets is due to pressure gradients. Pressure tries to resist compression, so if this term dominates, we get an oscillating solution, standing density (sound) waves. The second term in the brackets is due to gravity. If this term dominates, the perturbations grow. The wavenumber for which the terms are equal,

$$k_J \;=\; \frac{a\sqrt{4\pi G\bar{\rho}}}{c_s} \;=\; \sqrt{\frac{3}{2}}\frac{1}{c_s}\mathcal{H}\,, \tag{115}$$

is called the *Jeans wave number*, and the corresponding wavelength

$$\lambda_J \;=\; \frac{2\pi}{k_J} \;=\; 2\pi c_s\sqrt{\frac{2}{3}}\mathcal{H}^{-1} \tag{116}$$

the *Jeans length*. In the latter equalities we assumed that the background solution is the flat FRW universe, so that

$$4\pi G\bar{\rho} = \tfrac{3}{2}H^2\,. \tag{117}$$

For nonrelativistic matter $c_s \ll 1$, so that the Jeans length is much smaller than the Hubble length, $k_J \gg \mathcal{H}$. Thus we can apply Newtonian theory for scales both larger and smaller than the Jeans length.

For **scales much smaller that the Jeans length**, $k \gg k_J$, we can approximate the Jeans equation by

$$\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} + \frac{c_s^2 k^2}{a^2}\delta_{\mathbf{k}} = 0\,. \tag{118}$$

The solutions are oscillating, i.e., we get sound waves. The exact solutions of (118) are Bessel functions, but for small scales we can make a further approximation by first ignoring the middle term (which is smaller than the other two) and the time-dependence of a and c_s to get that $\delta_{\mathbf{k}}(t) \sim e^{\pm i\omega t}$, where $\omega = c_s k/a$. These oscillations are damped by the $2H\dot{\delta}_{\mathbf{k}}$ term, so the amplitude of the oscillations decreases with time. There is no growth of structure for sub-Jeans scales.

Exercise: Sound waves. For short-wavelength modes $k \gg k_J$, density perturbations in the matter-dominated universe satisfy (118). Switch to conformal time, $d\eta = dt/a$, and solve $\delta_{\mathbf{k}}(\eta)$ for the $\Omega_m = 1$,

[11] In the literature, there is usually no name given to this equation, but the terms *Jeans length* etc. are standard.

$\Omega_\Lambda = 0$ cosmology, assuming $c_s = const$. How does the amplitude and frequency of the oscillations change with time and scale factor? (Hint: The solutions are spherical Bessel functions.)

For **scales much longer than the Jeans length** (but still subhorizon), $\mathcal{H} \ll k \ll k_J$, we can approximate the Jeans equation by

$$\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} - 4\pi G\bar{\rho}\delta_{\mathbf{k}} = 0\,. \tag{119}$$

We dropped the pressure gradient term, which means that this equation applies also to nonadiabatic perturbations for scales where pressure gradients can be ignored. Note that Eq. (119) is the same for all $\mathbf{k}$, i.e., there is no k-dependence in the coefficients. This means that the equation applies also in coordinate space, i.e. for $\delta(\mathbf{x})$, as long as we ignore contributions from scales that do not satisfy $\mathcal{H} \ll k \ll k_J$.

For a matter-dominated universe, the background solution is $a \propto t^{2/3}$, so that

$$H = \frac{\dot{a}}{a} = \frac{2}{3t} \tag{120}$$

and

$$\frac{8\pi G}{3}\bar{\rho} = H^2 = \frac{4}{9t^2} \qquad \Rightarrow \qquad \bar{\rho} = \frac{1}{6\pi G t^2}\,, \tag{121}$$

so the Jeans equation becomes

$$\ddot{\delta}_{\mathbf{k}} + \frac{4}{3t}\dot{\delta}_{\mathbf{k}} - \frac{2}{3t^2}\delta_{\mathbf{k}} = 0\,. \tag{122}$$

The general solution is

$$\delta_{\mathbf{k}}(t) = b_1 t^{2/3} + b_2 t^{-1}\,. \tag{123}$$

The first term is the *growing mode* and the second term the *decaying mode*. After some time the decaying mode has died out, and the perturbation grows

$$\boxed{\delta \propto t^{2/3} \propto a\,.} \tag{124}$$

Thus *density perturbations in matter grow proportional to the scale factor.*

From Eq. (98) we have that

$$\Phi \propto a^2\bar{\rho}\delta \propto a^2 a^{-3} a = \text{const.}$$

The gravitational potential perturbation is constant in time during the matter-dominated era.

8.2.7 Many fluid components

Assume now that the "cosmic fluid" contains several components i (different types of matter or energy) which *do not interact with each other*, except gravitationally. This means that each component sees only its own pressure[12], and that the components can have different flow velocities. Then the Newtonian equations for each component i are

$$\frac{\partial \rho_i}{\partial t'} + \nabla_{\mathbf{r}} \cdot (\rho_i \mathbf{u}_i) = 0 \tag{125}$$

$$\frac{\partial \mathbf{u}_i}{\partial t'} + (\mathbf{u}_i \cdot \nabla_{\mathbf{r}})\mathbf{u}_i + \frac{1}{\rho_i}\nabla_{\mathbf{r}} p_i + \nabla_{\mathbf{r}}\tilde{\Phi} = 0 \tag{126}$$

$$\nabla^2_{\mathbf{r}}\tilde{\Phi} = 4\pi G\rho\,, \tag{127}$$

[12] In standard cosmology, we actually have just one component, the baryon-photon fluid, which sees its own pressure, and the other components do not see even *their own* pressure (neutrinos after decoupling) or do not even *have* pressure (cold dark matter). But we shall first do this general treatment, and do the application to standard cosmology later.

where $\rho = \sum \rho_i$. Note that there is only one gravitational potential $\tilde{\Phi}$, due to the total density, and this way the different components do interact gravitationally.

We again have the homogeneous solution, where now each component has to satisfy

$$\dot{\rho}_i + 3H\rho_i = 0\,, \tag{128}$$

and the expansion law

$$\dot{H} + H^2 = -\frac{4\pi G}{3}\rho\,, \tag{129}$$

is determined by the total density.

We can now introduce the density, pressure, and velocity perturbations for each component separately,

$$\begin{aligned}
\rho_i(t', \mathbf{r}) &= \bar{\rho}_i(t) + \delta\rho_i(t', \mathbf{r}) && (130)\\
p_i(t', \mathbf{r}) &= \bar{p}_i(t) + \delta p_i(t', \mathbf{r}) && (131)\\
\mathbf{u}_i(t', \mathbf{r}) &= H(t)\mathbf{r} + \mathbf{v}_i(t', \mathbf{r})\,, && (132)
\end{aligned}$$

but there is only one gravitational potential perturbation,

$$\tilde{\Phi}(t', \mathbf{r}) = \frac{2\pi G}{3}\bar{\rho}r^2 + \Phi(t', \mathbf{r})\,. \tag{133}$$

Following the earlier procedure, we obtain the perturbation equations for the fluid components,

$$\begin{aligned}
\frac{\partial}{\partial t'}\delta\rho_i + 3H\delta\rho_i + H\mathbf{r}\cdot\nabla_{\mathbf{r}}\delta\rho_i + \bar{\rho}_i\nabla\cdot\mathbf{v}_i &= 0 && (134)\\
\frac{\partial}{\partial t'}\mathbf{v}_i + H\mathbf{v}_i + H\mathbf{r}\cdot\nabla_{\mathbf{r}}\mathbf{v}_i + \frac{1}{\bar{\rho}_i}\nabla_{\mathbf{r}}\delta p_i + \nabla_{\mathbf{r}}\Phi &= 0 && (135)\\
\nabla^2_{\mathbf{r}}\Phi &= 4\pi G\delta\rho && (136)
\end{aligned}$$

in Newtonian coordinate space, and

$$\begin{aligned}
\dot{\delta}_{i\mathbf{k}} + \frac{i\mathbf{k}\cdot\mathbf{v}_{i\mathbf{k}}}{a} &= 0 && (137)\\
\frac{d}{dt}\left(a\mathbf{v}_{i\mathbf{k}}\right) + i\mathbf{k}\frac{\delta p_{i\mathbf{k}}}{\bar{\rho}_i} + i\mathbf{k}\Phi_{\mathbf{k}} &= 0 && (138)\\
\Phi_{\mathbf{k}} &= -4\pi G\frac{a^2}{k^2}\sum\bar{\rho}_i\delta_{i\mathbf{k}} && (139)
\end{aligned}$$

in comoving Fourier space. Here $\delta\rho = \sum\delta\rho_i$ and

$$\delta_i \equiv \frac{\delta\rho_i}{\bar{\rho}_i}\,. \tag{140}$$

Separating out the scalar perturbations we finally get

$$\ddot{\delta}_{i\mathbf{k}} + 2H\dot{\delta}_{i\mathbf{k}} = -\frac{k^2}{a^2}\frac{\delta p_{i\mathbf{k}}}{\bar{\rho}_i} + 4\pi G\delta\rho_{\mathbf{k}}\,, \tag{141}$$

where

$$\delta\rho_{\mathbf{k}} = \sum_j \bar{\rho}_j\delta_{j\mathbf{k}}\,. \tag{142}$$

8.2.8 Radiation

Since radiation is a relativistic form of energy, we cannot apply the preceding Newtonian discussion to perturbations in radiation. However, the qualitative results are similar.

The equation of state for radiation is $p = \rho/3$, and the speed of sound in a radiation fluid is given by

$$c_s^2 = \frac{dp}{d\rho} = \frac{1}{3}\,.$$

Thus the Jeans length for radiation is comparable to the Hubble length, and the subhorizon scales are also sub-Jeans scales for radiation. Thus for subhorizon radiation perturbations we only get oscillatory solutions. During the radiation-dominated epoch they are not damped by expansion, but the oscillation amplitude stays roughly constant.

Relativistic perturbations in non-expanding space. While the full treatment of relativistic perturbations is beyond the level of this course, we can obtain the limit where we ignore the effect of expansion by combining special relativity and the Newtonian limit of general relativity. Special relativistic fluid dynamics follows from the energy-momentum continuity equation

$$\frac{\partial T^{\mu\nu}}{\partial x^\nu} \equiv \partial_\nu T^{\mu\nu} \equiv T^{\mu\nu}{}_{,\nu} \;=\; 0\,. \tag{143}$$

For a perfect fluid

$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu + p g^{\mu\nu}\,, \tag{144}$$

where the metric is now that of Minkowski space, $g^{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$. The 4-velocity u^μ is related to the 3-velocity $\vec{v} = v^i$ by

$$u^\mu = (\gamma, \gamma\mathbf{v})\,, \tag{145}$$

where $\gamma = 1/\sqrt{1 - v^2}$.

By contracting the energy tensor $T^{\mu\nu}$ with the 4-velocity u_μ we obtain $u_\nu T^{\mu\nu}{}_{,\nu} = 0$, which gives

$$(\rho u^\mu)_{,\mu} + p u^\mu{}_{,\mu} \;=\; 0\,, \tag{146}$$

the energy continuity equation. Subtracting u^ν times this from (143) we get the special relativistic Euler equation

$$(\rho + p)u^\mu u^\nu{}_{,\mu} + (g^{\mu\nu} + u^\mu u^\nu)p_{,\mu} \;=\; 0\,, \tag{147}$$

where

$$u^\mu u^\nu{}_{,\mu} \equiv a^\nu \tag{148}$$

is the 4-acceleration.

For small velocities, $v \ll 1$, we can approximate $\gamma \approx 1$, so that

$$u^\mu \approx (1, \mathbf{v}) \tag{149}$$

and (146),(147) become

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) &= -p\nabla\cdot\mathbf{v} \\ (\rho + p)\left(\frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla\right)\vec{v} &= -\nabla p - \mathbf{v}(\mathbf{v}\cdot\nabla p) \;\approx\; -\nabla p\,. \end{aligned} \tag{150}$$

In the Newtonian limit of general relativity, but without the assumption $p \ll \rho$, the passive gravitational mass density is given by $\rho + p$, so that the gravitational force on a volume element of fluid is given by $-(\rho + p)\nabla\Phi$ and the active gravitational mass density by $\rho + 3p$, so that the gravitational potential is given by

$$\nabla^2\Phi = 4\pi G(\rho + 3p)\,. \tag{151}$$

Thus the Euler equation with gravity becomes

$$(\rho + p)\left(\frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla\right)\mathbf{v} \;=\; -\nabla p - (\rho + p)\nabla\Phi\,. \tag{152}$$

For several fluid components, not interacting with each other except gravitationally, the fluid equations become thus

$$\begin{aligned} \frac{\partial \rho_i}{\partial t} + \nabla \cdot (\rho_i \mathbf{v}_i) &= -p_i \nabla \cdot \mathbf{v}_i \\ (\rho_i + p_i)\left(\frac{\partial}{\partial t} + \mathbf{v}_i \cdot \nabla\right)\mathbf{v}_i &= -\nabla p_i - (\rho_i + p_i)\nabla\Phi \\ \nabla^2\Phi &= 4\pi G \sum_i (\rho_i + 3p_i)\,. \end{aligned} \tag{153}$$

For perturbations $\rho_i = \bar{\rho}_i + \delta\rho_i = \bar{\rho}_i(1+\delta_i)$, $p_i = \bar{p}_i + \delta p_i$, where the background density and pressure are now constant both in space and time, we get to first order in perturbations

$$\begin{aligned} \frac{\partial \delta_i}{\partial t} &= -(1+w_i)\nabla \cdot \mathbf{v}_i \\ (\bar{\rho}_i + \bar{p}_i)\frac{\partial \mathbf{v}_i}{\partial t} &= -\nabla \delta p_i - (\bar{\rho}_i + \bar{p}_i)\nabla\Phi \\ \nabla^2\Phi &= 4\pi G \sum_i (\bar{\rho}_i \delta_i + 3\delta p_i)\,. \end{aligned} \tag{154}$$

For Fourier components this becomes

$$\begin{aligned} \dot{\delta}_{i\mathbf{k}} &= -ik(1+w_i)\mathbf{k} \cdot \mathbf{v}_{i\mathbf{k}} \\ (\bar{\rho}_i + \bar{p}_i)\dot{\mathbf{v}}_{i\mathbf{k}} &= -i\mathbf{k}\delta p_{i\mathbf{k}} - i\mathbf{k}(\bar{\rho}_i + \bar{p}_i)\Phi_{\mathbf{k}} \\ \Phi_{\mathbf{k}} &= \frac{-4\pi G}{k^2} \sum_i (\bar{\rho}_i \delta_{i\mathbf{k}} + 3\delta p_{i\mathbf{k}})\,. \end{aligned} \tag{155}$$

For vector perturbations the second equation gives

$$\dot{\mathbf{v}}_{i\perp\mathbf{k}} = 0 \quad \Rightarrow \quad \mathbf{v}_{i\perp\mathbf{k}} = const\,, \tag{156}$$

and for scalar perturbations the first and second equations become

$$\begin{aligned} \dot{\delta}_{i\mathbf{k}} &= -i(1+w_i)k v_{i\mathbf{k}} \\ \dot{v}_{i\mathbf{k}} &= -ik\frac{\delta p_{i\mathbf{k}}}{\bar{\rho}_i + \bar{p}_i} - ik\Phi_{\mathbf{k}}\,, \end{aligned} \tag{157}$$

from which we get (note that $\dot{w}_i = 0$, since $w_i \equiv \bar{\rho}_i/\bar{p}_i$) the Jeans equation as

$$\ddot{\delta}_{i\mathbf{k}} + k^2\frac{\delta p_{i\mathbf{k}}}{\bar{\rho}_i} + k^2(1+w_i)\Phi_{\mathbf{k}} = 0\,. \tag{158}$$

8.2.9 Adiabatic and entropy perturbations again

The simplest inflation models predict that the primordial perturbations are adiabatic. This means that locally the perturbed universe at some $(t, \mathbf{x})$ looks like the background universe at some time $t + \delta t(\mathbf{x})$. See Sec. 8.2.5.

$$\left.\begin{aligned} \delta\rho_i(\mathbf{x}) &= \dot{\bar{\rho}}_i \delta t(\mathbf{x}) \\ \delta p_i(\mathbf{x}) &= \dot{\bar{p}}_i \delta t(\mathbf{x}) \end{aligned}\right\} \quad \Rightarrow \quad \begin{cases} \dfrac{\delta p_i}{\delta\rho_i} = \dfrac{\dot{\bar{p}}_i}{\dot{\bar{\rho}}_i} \\ \dfrac{\delta\rho_i}{\delta\rho_j} = \dfrac{\dot{\bar{\rho}}_i}{\dot{\bar{\rho}}_j} \end{cases} \quad \Rightarrow \quad \frac{\delta_i}{\delta_j} = \frac{\dot{\bar{\rho}}_i}{\bar{\rho}_i}\frac{\bar{\rho}_j}{\dot{\bar{\rho}}_j} \tag{159}$$

If there is no energy transfer between the fluid components at the background level, the energy continuity equation is satisfied by them separately,

$$\dot{\bar{\rho}}_i = -3H(\bar{\rho}_i + \bar{p}_i) \equiv -3H(1+w_i)\bar{\rho}_i\,, \tag{160}$$

where $w_i \equiv \bar{p}_i/\bar{\rho}_i$. Thus for adiabatic perturbations,

$$\frac{\delta_i}{1+w_i} = \frac{\delta_j}{1+w_j} \tag{161}$$

(which is thus related to $\bar{\rho}_i \propto a^{-(1+w_i)}$). For matter components $w_i \approx 0$, and for radiation components $w_i = \frac{1}{3}$. Thus, for adiabatic perturbations, all matter components have the same perturbation

$$\delta_i = \delta_m$$

and all radiation perturbations have likewise

$$\delta_i = \delta_r = \frac{4}{3}\delta_m \,.$$

We can define a *relative entropy perturbation*[13] between two components

$$S_{ij} \;\equiv\; -3H\left(\frac{\delta\rho_i}{\dot{\bar{\rho}}_i} - \frac{\delta\rho_j}{\dot{\bar{\rho}}_j}\right) \;=\; \frac{\delta_i}{1+w_i} - \frac{\delta_j}{1+w_j} \tag{162}$$

to describe a deviation from the adiabatic case. The relative entropy perturbation is a perturbation in the ratio of the number densities of the two species. For a nonrelativistic species

$$\rho_i = m_i n_i \quad \Rightarrow \quad \delta\rho_i = m_i \delta n_i \qquad \text{and} \qquad \delta_i \equiv \frac{\delta\rho_i}{\bar{\rho}_i} = \frac{\delta n_i}{\bar{n}_i}\,, \tag{163}$$

whereas for an ultrarelativistic species ($\mu \ll T$ and $m \ll T$)

$$\begin{aligned} \rho_i \;&\propto\; T_i^4 \quad \Rightarrow \quad \delta\rho_i = \bar{\rho}_i \cdot 4\frac{\delta T_i}{T_i} \\ n_i \;&\propto\; T_i^3 \quad \Rightarrow \quad \delta n_i = \bar{n}_i \cdot 3\frac{\delta T_i}{T_i} \\ &\Rightarrow \quad \delta_i \equiv \frac{\delta\rho_i}{\bar{\rho}_i} = \frac{4}{3}\frac{\delta n_i}{\bar{n}_i}\,. \end{aligned} \tag{164}$$

For both cases

$$\delta_i = (1+w_i)\frac{\delta n_i}{\bar{n}_i}\,. \tag{165}$$

Thus

$$S_{ij} \;=\; \frac{\delta n_i}{\bar{n}_i} - \frac{\delta n_j}{\bar{n}_j} \;=\; \frac{\delta(n_i/n_j)}{\bar{n}_i/\bar{n}_j}\,. \tag{166}$$

Even if perturbations are initially adiabatic, relative entropy perturbation may develop inside the horizon. We shall encounter such a case in Sec. 8.3.4.

8.2.10 The effect of a homogeneous component

The energy density of the real universe consists of several components. In many cases it is reasonable to ignore the perturbations in some components (since they are relatively small in the scales of interest). We call such components *smooth* and we can add them together into a single smooth component $\rho_s = \bar{\rho}_s$.

[13] There is a connection to entropy/particle of the different components, but we need not concern ourselves with it now. It is not central to this concept, and it is perhaps somewhat unfortunate that it has become customary, for historical reasons, to use the word "entropy" for these perturbations.

Consider the case where we have perturbations in a nonrelativistic ("matter") component ρ_m, and the other components are smooth. Then

$$\rho = \rho_m + \rho_s \tag{167}$$

but

$$\delta\rho = \delta\rho_m \equiv \bar{\rho}_m \delta \,. \tag{168}$$

We write just δ for $\delta\rho_m/\bar{\rho}_m$, since there is no other density perturbation, but note that now $\delta \neq \delta\rho/\bar{\rho}$ (beware of this trap!).

Assuming adiabatic perturbations, we have then from Eq. (141) that

$$\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} + \left[\frac{c_s^2 k^2}{a^2} - 4\pi G\bar{\rho}_m\right]\delta_{\mathbf{k}} = 0 \,. \tag{169}$$

The difference from Eq. (114) is that now the background energy density in the "gravity" term still contains only the matter component $\bar{\rho}_m$, but the expansion law, $a(t)$ and $H(t)$ comes from the full background energy density $\bar{\rho} = \bar{\rho}_m + \bar{\rho}_s$.

Newtonian perturbation theory can be applied even with the presence of relativistic energy components, like radiation and dark energy, as long as they can be considered as smooth components and their perturbations can be ignored. Then they contribute only to the background solution. In this case we have to calculate the background solution using general relativity, i.e., the background solution is a FRW universe, but the perturbation equations are the Newtonian perturbation equations. We can also consider a non-flat (open or closed) FRW universe, as long as we only apply perturbation theory to scales much shorter than the curvature radius (and the Hubble length). Thus the background quantities are to be solved from the Friedmann and energy continuity equations

$$H^2 + \frac{K}{a^2} = \frac{8\pi G}{3}\rho \tag{170}$$

$$\dot{H} + H^2 = -\frac{4\pi G}{3}(\rho + 3p) \tag{171}$$

$$\dot{\rho} = -3H(\rho + p) \,. \tag{172}$$

Example: Matter perturbations in flat vacuum-dominated universe. Consider the case where $\rho_s = \rho_{\rm vac} \gg \rho_m$ and matter is approximated as pressureless (we do not then have to make a separate adiabaticity assumption, since the pressure term does not appear). Then the Jeans equation becomes

$$\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} - 4\pi G\bar{\rho}_m\delta_{\mathbf{k}} = 0 \,. \tag{173}$$

To estimate the relative order of magnitude of the three terms it is better to divide the equation by H^2,

$$H^{-2}\ddot{\delta}_{\mathbf{k}} + 2H^{-1}\dot{\delta}_{\mathbf{k}} - \frac{4\pi G\bar{\rho}_m}{H^2}\delta_{\mathbf{k}} = 0 \,, \tag{174}$$

so that the Hubble time H^{-1} provides the time scale for the time derivatives. Now

$$H^2 = \frac{8\pi G}{3}\rho_{\rm cr} \approx \frac{8\pi G}{3}\rho_{\rm vac} = const \tag{175}$$

and in the last term $\delta_{\mathbf{k}}$ is multiplied with $\frac{3}{2}\bar{\rho}_m/\rho_{\rm vac} \ll 1$, so that we can drop the last term and approximate the Jeans equation by

$$\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} = 0 \,. \tag{176}$$

We see immediately that $\delta_{\mathbf{k}} = const$ is a solution. For the other solution, solve first $\dot{\delta}_{\mathbf{k}}$:

$$\frac{d\dot{\delta}_{\mathbf{k}}}{dt} = -2H\dot{\delta}_{\mathbf{k}} \quad \Rightarrow \quad \frac{d\dot{\delta}_{\mathbf{k}}}{\dot{\delta}_{\mathbf{k}}} = -2H dt \,, \tag{177}$$

whose solution is $\ln \dot{\delta}_{\mathbf{k}} = -2Ht + const$ or $\dot{\delta}_{\mathbf{k}} = Ce^{-2Ht}$. Integrating this gives

$$\delta_{\mathbf{k}} \;=\; Ae^{-2Ht} + B\,, \tag{178}$$

with a constant term and an exponentially decaying term. Thus in a vacuum-dominated universe matter perturbations stay constant (after the decaying term has died out); or to be more precise and referring to the original equation (174), the relative change in $\delta_{\mathbf{k}}$ in a Hubble time is of order $\bar{\rho}_m/\rho_{\rm vac} \ll 1$

We shall do this calculation more accurately later, including the transition from matter domination to vacuum domination. The main lesson now is that the increased expansion rate due to the presence of a smooth component slows down the growth of perturbations.

Exercise: Find the solution for the Jeans equation for pressureless matter perturbations when a) the energy density is dominated by a smooth radiation component b) when there is no other energy component, but the universe has the open geometry ($K < 0$) and is curvature dominated, considering only scales $\ll$ curvature radius.

8.3 Perturbations at subhorizon scales in the real universe

8.3.1 Horizon entry

Newtonian perturbation theory is valid only at subhorizon scales, $k \gg \mathcal{H}$, or $k^{-1} \ll \mathcal{H}^{-1}$. During "normal", decelerating expansion, i.e., after inflation but before the recent onset of dark energy domination, scales are entering the horizon. Short scales enter first, large scales enter later. We have not yet studied what happens to perturbations outside the horizon (for that we need (general) relativistic perturbation theory, to be discussed somewhat later). So, for the present discussion, whatever values the perturbation amplitudes $\delta_{\mathbf{k}}$ have soon after horizon entry, are to be taken as an initial condition, the *primordial perturbation*[14]. Observations actually suggest that different scales enter the horizon with approximately equal perturbation amplitude, whose magnitude is characterized by the number[15] few $\times\, 10^{-5}$.

The history of the different scales after horizon entry, and thus their present perturbation amplitude, depends on at what epoch they enter. The scales which enter during transitions between epochs are thus special scales which should characterize the present structure of the universe. Such important scales are the scale (**exercise**)

$$k_{\rm eq}^{-1} \;=\; (\mathcal{H}_{\rm eq})^{-1} \;\sim\; 13.7\,\Omega_m^{-1} h^{-2}\,{\rm Mpc} \;\equiv\; 13.7\,\omega_m^{-1}\,{\rm Mpc}\,, \tag{179}$$

which enters at the time $t_{\rm eq}$ of matter-radiation equality, and the scale

$$\begin{aligned} k_{\rm dec}^{-1} \;&=\; (\mathcal{H}_{\rm dec})^{-1} \;\sim\; 91\,\Omega_m^{-1/2}\left[1+\frac{\Omega_r}{\Omega_m}(1+z_{\rm dec})\right]^{-1/2} h^{-1}{\rm Mpc} \\ &\equiv\; 91\,\omega_m^{-1/2}\left[1+\frac{\omega_r}{\omega_m}(1+z_{\rm dec})\right]^{-1/2} {\rm Mpc}\,, \end{aligned} \tag{180}$$

which enters at the time $t_{\rm dec}$ ($z_{\rm dec} = 1090$) of photon decoupling. Here $\omega_r = 4.18 \times 10^{-5}$ includes relativistic neutrinos, since the result above only requires them to be relativistic at $t_{\rm dec}$. For $\Omega_\Lambda = 0.7$, $\Omega_m = 0.3$, $h = 0.7$, these scales are

$$\begin{aligned} k_{\rm eq}^{-1} \;&=\; 65\,h^{-1}{\rm Mpc} \;=\; 93\,{\rm Mpc} \\ k_{\rm dec}^{-1} \;&=\; 145\,h^{-1}{\rm Mpc} \;=\; 207\,{\rm Mpc}\,. \end{aligned} \tag{181}$$

The smallest "cosmological" scale is that corresponding to a typical distance between galaxies, about 1 Mpc.[16] This scale entered during the radiation-dominated epoch (well after Big Bang nucleosynthesis).

The scale corresponding to the present "horizon" (i.e. Hubble length) is

$$k_0^{-1} \;=\; (\mathcal{H}_0)^{-1} \;=\; 2998\,h^{-1}{\rm Mpc} \;\sim\; 4300\,{\rm Mpc}\,. \tag{182}$$

Because of the acceleration due to dark energy, this scale is actually *exiting* now, and there are scales, somewhat larger than this, that have briefly entered, and then exited again in the recent past. The horizon entry is not to be taken as an instantaneous process, so these scales were

[14] We shall later redefine *primordial perturbation* to refer to the perturbations at the epoch when all cosmologically interesting scales were well outside the horizon, which is the standard meaning of this concept in cosmology.

[15] Although in coordinate space the relative density perturbation $\delta(\mathbf{x})$ is a dimensionless number, the Fourier quantity $\delta_{\mathbf{k}}$ is not. The size of $\delta_{\mathbf{k}}$ is characterized by the dimensionless value $\mathcal{P}(k)^{1/2}$.

[16] In the present universe, structure at smaller scales has been completely messed up by galaxy formation, so that it bears little relation to the primordial perturbations at these scales. However, observations of the high-redshift universe, especially so-called Lyman-α observations (absorption spectra of high-z quasars, which reveal distant gas clouds along the line of sight), can reveal these structures when they are closer to their primordial state. With such observations, the "cosmological" range of scales can be extended down to ~ 0.1 Mpc.

never really subhorizon enough for the Newtonian theory to apply to them. Thus we shall just consider scales $k^{-1} < k_0^{-1}$. The largest observable scales, of the order of k_0^{-1}, are essentially at their "primordial" amplitude now.

We shall now discuss the evolution of the perturbations at these scales ($k^{-1} < k_0^{-1}$) after horizon entry, using the Newtonian perturbation theory presented in the previous section.

8.3.2 Composition of the real universe

The present understanding is that there are five components to the energy density of the universe,

1. cold dark matter (c)
2. baryonic matter (b)
3. photons (γ)
4. neutrinos (ν)
5. dark energy (d)

(during the time of interest for this section, i.e., from some time after BBN until the present). Thus

$$\rho = \underbrace{\rho_c + \rho_b}_{\rho_m} + \underbrace{\rho_\gamma + \rho_\nu}_{\rho_r} + \rho_d \,. \tag{183}$$

(Note that ρ_c here is the CDM density, not the critical density, for which we write $\rho_{\rm cr}$.) Baryons and photons interact with each other until $t = t_{\rm dec}$, so for $t < t_{\rm dec}$ they have to be discussed as a single component,

$$\rho_{b\gamma} = \rho_b + \rho_\gamma \,. \tag{184}$$

The other components do not interact with each other, except gravitationally, during the time of interest. The fluid description of Sec. 8.2 can only be applied to components whose particle mean free paths are shorter than the scales of interest. After decoupling, photons "free stream" and cannot be discussed as a fluid. On the other hand, the photon component becomes then rather homogeneous quite soon, so we can approximate it as a "smooth" component[17]. The same applies to neutrinos for the whole time since the BBN epoch, until the neutrinos become nonrelativistic. After neutrinos become nonrelativistic, they should be treated as matter (hot dark matter), not radiation. According to observations, the neutrino masses are small enough, not to have a major impact on structure formation. Thus we shall here approximate neutrinos as a smooth radiation component. Dark energy is believed to be relatively smooth. If it is a cosmological constant (vacuum energy) then it is perfectly homogeneous.

The discussion in Sec. 8.2 applies to the case, where ρ can be divided into two components,

$$\rho = \rho_m + \rho_s \,, \tag{185}$$

where the perturbation is only in the matter component ρ_m and $\rho_s = \bar{\rho}_s$ is homogeneous. For perturbations in radiation components and dark energy the Newtonian treatment is not enough. Unfortunately, we do not have quite this two-component case here. Based on the above discussion, a reasonable approximation is given by a separation into three components:

$$t < t_{\rm dec}: \quad \rho = \rho_c + \rho_{b\gamma} + \rho_s \qquad (\rho_s = \rho_\nu + \rho_d) \tag{186}$$

$$t > t_{\rm dec}: \quad \rho = \rho_c + \rho_b + \rho_s \qquad (\rho_s = \rho_\gamma + \rho_\nu + \rho_d) \,. \tag{187}$$

[17] As long as we are interested in density perturbations only. When we are interested in the CMB anisotropy, the momentum distribution of these photons becomes the focus of our attention.

After decoupling, both ρ_c and ρ_b are matter-like ($p \ll \rho$) and we'll discuss in Sec. 8.3.4 how this case is handled. Before decoupling, the situation is more difficult, since $\rho_{b\gamma}$ is not matter-like, the pressure provided by the photons is large. Here we shall be satisfied with a crude approximation for this period.

The most difficult period is that close to decoupling, where the photon mean free path λ_γ is growing rapidly. The fluid description, which we are here using for the perturbations, applies only to scales $\gg \lambda_\gamma$, whereas the photons are smooth only for scales $\ll \lambda_\gamma$. Thus this period can be treated properly only with large numerical "Boltzmann" codes, such as CMBFAST or CAMB.

8.3.3 CDM density perturbations

Cold dark matter is the dominant structure-forming component in the universe (dark energy dominates the energy density at late times, but does not form structure, or, if it does, these structures are very weak, not far from homogeneous). Observations indicate that $\rho_b \sim 0.2\rho_c$. Thus we get a reasonable approximation for the behavior of the CDM perturbations by ignoring the baryon component and equating

$$\rho_m \approx \rho_c \,.$$

The CDM is pressureless, and thus the CDM sound speed is zero, and so is the CDM Jeans length. Thus, for CDM, all scales are larger than the Jeans scale, and we don't get an oscillatory behavior. Instead, perturbations grow at all scales. On the other hand, as we shall discuss in Sec. 8.3.4, perturbations in $\rho_{b\gamma}$ oscillate before decoupling. Therefore the perturbations in $\rho_{b\gamma}$ will be smaller than those in ρ_c, and we can make a (crude) approximation where we treat $\rho_{b\gamma}$ as a homogeneous component before decoupling. This is important, since although $\rho_b \ll \rho_c$, this is not true for $\rho_{b\gamma}$ at earlier, radiation-dominated, times. At decoupling $\rho_b < \rho_\gamma < \rho_c$. Before matter-radiation equality, there is an epoch when $\rho_c < \rho_\gamma$, but $\delta\rho_c > \delta\rho_{b\gamma}$. For simplicity, we now approximate

$$\rho = \rho_m + \rho_r + \rho_d \tag{188}$$

where $\rho_m = \rho_c$ and $\rho_r = \rho_\gamma + \rho_\nu$ is a smooth component (ρ_ν truly smooth, ρ_γ truly smooth after decoupling, and (crudely) approximated as smooth before decoupling). We have ignored baryons, since they are a subdominant part of $\rho_{b\gamma}$ before decoupling, and a subdominant matter component after decoupling. Likewise, ρ_d is also smooth, and becomes important only close to present times.

We can now study the growth of CDM perturbations even during the radiation-dominated period, as the radiation-component is taken as smooth and affects only the expansion rate. We can study it all the way from horizon entry to the present time, or until the perturbations become nonlinear ($\delta_c = \delta\rho_c/\bar{\rho}_c \sim 1$).

We get the equation for the CDM perturbation from Eq. (169) by setting $c_s = 0$ (or rather, $\delta p = 0$; we need not invoke the assumption of adiabaticity, since CDM is pressureless),

$$\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} - 4\pi G\bar{\rho}_m\delta_{\mathbf{k}} = 0\,. \tag{189}$$

Note that the equation is the same for all $\mathbf{k}$ and therefore it applies also in the coordinate space, i.e., for $\delta(\mathbf{x})$. To simplify notation, we drop the subscript $_{\mathbf{k}}$.

We now assume a flat universe, and ignore the ρ_d component (a good approximation at early times[18]), so that the Friedmann equation is

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\bar{\rho}\,,$$

[18] For a flat universe (or an open/closed universe without dark energy) we can also do the late times (see Sec. 8.3.5). Then dark energy (or curvature) must be included in the background evolution, but radiation can be ignored.

where $\bar{\rho} = \bar{\rho}_m + \bar{\rho}_r$ and $\bar{\rho}_m \propto a^{-3}$ and $\bar{\rho}_r \propto a^{-4}$.

A useful trick is to study this as a function of a instead of t or η. We define a new time coordinate,

$$y \equiv \frac{a}{a_{\rm eq}} = \frac{\bar{\rho}_m}{\bar{\rho}_r}\,. \tag{190}$$

($y = 1$ at $t = t_{\rm eq}$.) Now

$$4\pi G\bar{\rho}_m \;=\; 4\pi G\frac{y}{y+1}\bar{\rho} \;=\; \frac{3}{2}\frac{y}{y+1}H^2 \tag{191}$$

and Eq. (189) becomes

$$\ddot{\delta} + 2H\dot{\delta} - \frac{3}{2}\frac{y}{y+1}H^2 = 0\,. \tag{192}$$

Performing the change of variables from t to y (**Exercise**; you may need the 2nd Friedmann equation), we arrive at the equation (where $' \equiv d/dy$)

$$\boxed{\delta'' + \frac{2+3y}{2y(1+y)}\delta' - \frac{3}{2y(1+y)}\delta \;=\; 0\,,} \tag{193}$$

known as the *Meszaros equation*.

It has two solutions, one growing, the other one decaying. The growing solution is

$$\boxed{\delta \;=\; \delta_{\rm prim}\left(1+\frac{3y}{2}\right) \;=\; \delta_{\rm prim}\left(1+\frac{3}{2}\frac{a}{a_{\rm eq}}\right)\,.} \tag{194}$$

We see that the perturbation remains frozen to its primordial value, $\delta \approx \delta_{\rm prim}$, during the radiation-dominated period. By $t = t_{\rm eq}$, it has grown to $\delta = \frac{5}{2}\delta_{\rm prim}$.

During the matter-dominated period, $y \gg 1$, the CDM perturbation grows proportional to the scale factor,

$$\delta \propto y \propto a \propto t^{2/3}\,. \tag{195}$$

In reality, for the case of adiabatic primordial perturbations, there is an additional logarithmic growth factor $\sim \ln(k/k_{\rm eq})$ the CDM perturbations get from the gravitational effect (ignored in the above) of the oscillating radiation perturbation during the radiation-dominated epoch. To get this boost the CDM perturbations must initially be in the same direction (positive or negative) as the radiation perturbations, which is the case for adiabatic primordial perturbations:

For adiabatic primordial perturbations, the baryon, CDM, and radiation perturbations at are related at horizon entry as $\delta_c = \delta_b = \frac{3}{4}\delta_\gamma$. Consider scales that enter during the radiation-dominated epoch ($t < t_{\rm eq} < t_{\rm dec}$). The gravitational effect is dominated initially by the radiation perturbations, which begin to oscillate after horizon entry; the baryon perturbations will oscillate with them until $t_{\rm dec}$. CDM on the other hand, does not see the radiation pressure responsible for the oscillation, it sees only the gravitational effect of the baryon-photon fluid. In the first phase of the oscillation period δ_c is of the same sign as $\delta_{b\gamma}$ so $\delta_{b\gamma}$ adds to the gravitational pull to increase δ_c and since at first $\delta\rho_{b\gamma} > \delta\rho_c$, this additional pull is larger than that of CDM itself, leading to a much faster growth of δ_c (which otherwise would grow very little during the radiation domination). The flow of CDM is accelerated towards CDM overdensities. In the next phase of the oscillation, the sign of $\delta_{b\gamma}$ reverses, and now the pull of $\delta\rho_{b\gamma}$ on CDM is in the opposite direction, and will slow down the flow of CDM towards overdensities. But this is not enough to reverse the CDM flow before the sign of $\delta_{b\gamma}$ changes again and begins to accelerate CDM again towards CDM overdensities. Thus the effect of the radiation oscillations is to increase δ_c stepwise, one step for each oscillation period. As the $\bar{\rho}_\gamma/\bar{\rho}_c$ ratio decreases the relative increases per step decrease; but this effect keeps adding steps until $t_{\rm dec}$. The smaller the scale (the higher

the k) the more steps there are between horizon entry (t_k) and $t_{\rm dec}$, and the larger the first steps. The calculation of this effect is too complicated for this course (I do it in Cosmological Perturbation Theory) but for $k \gg k_{\rm eq}$ the effect is a boost by a factor $\sim 8\ln(k_{\rm eq}/6k)$, so that (194) is modified to

$$\delta_c \;\approx\; \delta_{\rm prim}\left(1+\frac{3}{2}\frac{a}{a_{\rm eq}}\right)8\ln\left(\frac{k}{6k_{\rm eq}}\right) \qquad \text{for } k \gg 6k_{\rm eq} \text{ and } t > t_{\rm dec} \tag{196}$$

(for $k < 6k_{\rm eq}$) the logarithm is negative; this approximate result does not apply for such large scales).

8.3.4 Baryon density perturbations

Although CDM is the dominant matter component in the universe, we cannot directly see it. The main method to observe the density perturbations today is to study the distribution of galaxies. But the part of galaxies that we can see is baryonic. Thus to compare the theory of structure formation to observations, we need to study how perturbations in the baryonic component evolves.

We define the baryon Jeans length as $\lambda_J = 2\pi k_J^{-1}$, where

$$k_J^{-1} = \frac{c_s}{a\sqrt{4\pi G\bar{\rho}_b}}\,, \tag{197}$$

and c_s is the speed of sound for baryons (i.e., in the baryon-photon fluid before decoupling, and in the baryon fluid after decoupling). This definition compares baryon pressure to baryon gravity, so it addresses the question whether baryonic density perturbations can grow under their own gravity. This is not the question we face in reality, since at early times baryons were coupled to photons, and after decoupling the gravity of CDM perturbations dominates. The baryon Jeans length can still be used for order-of-magnitude estimates on at what scales the baryon perturbations can grow (and for the argument that we cannot match observations without CDM).

In general,

$$c_s^2 = \left(\frac{\partial p}{\partial \rho}\right)_\sigma\,, \tag{198}$$

where σ refers to constant entropy per baryon. Since in our case the entropy is completely dominated by photons,

$$s_{b\gamma} \sim s_\gamma = \frac{4\pi^2}{45}T^3 = \frac{2\pi^4}{45\zeta(3)}n_\gamma\,, \tag{199}$$

we have

$$\sigma \equiv \frac{s_{b\gamma}}{n_b} \sim \frac{s_\gamma}{n_b} = \frac{2\pi^4}{45\zeta(3)}\frac{n_\gamma}{n_b} \approx 3.6016\frac{1}{\eta}\,, \tag{200}$$

where η is the baryon-to-photon ratio.

We find the speed of sound by varying $\rho_{b\gamma}$ and $p_{b\gamma}$ *adiabatically*, (i.e., keeping σ, the entropy/baryon constant), which in this case means keeping η constant. Now

$$\begin{aligned}
\rho_b = mn_b = m\eta n_\gamma = m\eta\frac{2\zeta(3)}{\pi^2}T^3 \quad &\Rightarrow \quad \delta\rho_b = \bar{\rho}_b\cdot 3\frac{\delta T}{T}\\
\rho_\gamma = \frac{\pi^2}{15}T^4 \quad &\Rightarrow \quad \delta\rho_\gamma = \bar{\rho}_\gamma\cdot 4\frac{\delta T}{T}\\
p_\gamma = \frac{\pi^2}{45}T^4 \quad &\Rightarrow \quad \delta p_\gamma = \bar{p}_\gamma\cdot 4\frac{\delta T}{T} = \bar{\rho}_\gamma\cdot\frac{4}{3}\frac{\delta T}{T}\,.
\end{aligned}$$

Since $p_b \ll p_\gamma \quad \Rightarrow \quad \delta p_b \ll \delta p_\gamma$, we get

$$c_s^2 = \frac{\delta p}{\delta \rho} = \frac{\delta p_\gamma}{\delta \rho_\gamma + \delta \rho_b} = \frac{\frac{4}{3}\bar{\rho}_\gamma}{4\bar{\rho}_\gamma + 3\bar{\rho}_b} = \frac{1}{3}\frac{1}{1+\frac{3}{4}\frac{\bar{\rho}_b}{\bar{\rho}_\gamma}} \,. \tag{201}$$

This was a *calculation* of the speed of sound, which one gets by varying the pressure and density adiabatically. It is independent of whether the actual perturbations we study are adiabatic or not.

This result, Eq. (201), applies before decoupling. As we go back in time, $\bar{\rho}_b/\bar{\rho}_\gamma \to 0$ and $c_s^2 \to 1/3$. As we approach decoupling, $\bar{\rho}_b$ becomes comparable to (but still smaller than) $\bar{\rho}_\gamma$ and the speed of sound falls, but not by a large factor.

Newtonian perturbation theory applies only to subhorizon scales. The ratio of the (comoving) baryon Jeans length

$$\lambda_J = \frac{2\pi c_s}{a\sqrt{4\pi G \bar{\rho}_b}}$$

to the comoving Hubble length

$$\mathcal{H}^{-1} = \frac{1}{a\sqrt{\frac{8\pi G}{3}\bar{\rho}}}$$

is

$$\frac{\lambda_J}{\mathcal{H}^{-1}} = \mathcal{H}\lambda_J = 2\pi\sqrt{\frac{2\bar{\rho}}{3\bar{\rho}_b}}c_s \,.$$

Thus we see that before decoupling the baryon Jeans length is comparable to the Hubble length, and thus all scales for which our present discussion applies are sub-Jeans. Therefore, if baryon perturbations are adiabatic[19], they oscillate before decoupling[20].

After decoupling, the baryon component sees just its own pressure. This component is now a gas of hydrogen and helium. This gas is monatomic for the epoch we are now interested in. Hydrogen forms molecules only later. For a non-relativistic monatomic gas,

$$c_s^2 = \frac{5T_b}{3m} \,, \tag{202}$$

where we can take $m \approx 1\,\mathrm{GeV}$, since hydrogen dominates. Down until $z \sim 100$, residual free electrons maintain enough interaction between the baryon and photon components to keep $T_b \approx T_\gamma$. After that the baryon temperature falls faster,

$$T_b \propto (1+z)^2 \qquad \text{whereas} \qquad T_\gamma \propto 1+z \tag{203}$$

(as shown in an exercise in Chapter 4). For example, at $1+z = 1000$, soon after decoupling, $T_b = 2725\,\mathrm{K} = 0.2348\,\mathrm{eV}$ and the speed of sound is $c_s = 5930\,\mathrm{m/s}$. The baryon density is $\bar{\rho}_b = \Omega_b(1+z)^3\rho_{\mathrm{cr}} = \omega_b(1+z)^3\, 1.88\times 10^{-26}\,\mathrm{kg/m^3}$, and we get for the Jeans length

$$\lambda_J = (1+z)\frac{\sqrt{\pi}c_s}{\sqrt{G\bar{\rho}_b}} \tag{204}$$

[19] If there is an initial baryon entropy perturbation, i.e., a perturbation in baryon density without an accompanying radiation perturbation, it will initially begin to grow in the same manner as a CDM perturbation, since the pressure perturbation provided by the photons is missing. (Such a baryon entropy perturbation corresponds to a perturbation in the baryon-photon ratio η.) But as the movement of baryons drags the photons with them, a radiation perturbation is generated, and the baryon perturbation begins to oscillate around its initial value (instead of oscillating around zero).

[20] We have not calculated this exactly, since all our calculations have been idealized, i.e., we have used perturbation theory which applies only to matter-dominated perturbations, and here we have ignored the CDM component. But this qualitative feature will hold also in the exact calculation, and this will be enough for us now.

that soon after decoupling

$$\lambda_J(1+z=1000) \;=\; \omega_b^{-1/2}\,0.96\times10^3\,\mathrm{pc} \;=\; \eta_{10}\,0.016\,\mathrm{Mpc} \;\sim\; 0.095\,\mathrm{Mpc}\,, \tag{205}$$

where $\eta_{10}\equiv 10^{10}\eta = 274\,\omega_b$ or $\omega_b = 0.00365\,\eta_{10}$, and the last number is for $\eta_{10}\sim 6$.

We define the baryon Jeans mass

$$M_J \equiv \bar\rho_{b0}\frac{\pi}{6}\lambda_J^3 \tag{206}$$

as the mass of baryonic matter within a sphere whose diameter is λ_J. Note that since λ_J is defined as a comoving distance, we must use here the present (mean) baryon density $\bar\rho_{b0}$. At $1+z=1000$, the baryon Jeans mass is $\omega_b^{-1/2}\,1.3\times10^5\,\mathrm{M}_\odot = \eta_{10}^{-1/2}\,2.1\times10^6\,\mathrm{M}_\odot \sim 9\times10^5\,\mathrm{M}_\odot$ for $\eta_{10}\sim 6$. This corresponds to the mass of a globular cluster and is much less than the mass of a galaxy. Thus, for our purposes, the baryonic component is pressureless after decoupling, i.e., baryon pressure can be ignored in the evolution of perturbations at cosmological scales (greater than $\sim 1\,$Mpc). (The pressure cannot be ignored for smaller scale physics like the formation of individual galaxies.)

After decoupling, the evolution of the baryon density perturbation is governed by the gravitational effect of the dominant matter component, the CDM.

We now have the situation of Sec. 8.2.10, except that we have two matter components,

$$\rho = \rho_c + \rho_b + \rho_s\,, \tag{207}$$

where we approximate $\rho_s = \rho_\gamma + \rho_\nu + \rho_d$ as homogeneous. With the help of Sec. 8.2.7, the discussion is easy to generalize for the present case.

We can ignore the pressure of both ρ_b and ρ_c. Therefore their perturbation equations are

$$\ddot\delta_c + 2H\dot\delta_c = 4\pi G\bar\rho_m\delta \tag{208}$$

$$\ddot\delta_b + 2H\dot\delta_b = 4\pi G\bar\rho_m\delta \tag{209}$$

where $\bar\rho_m = \bar\rho_c + \bar\rho_b$ is the total background matter density and

$$\delta = \frac{\delta\rho_c + \delta\rho_b}{\bar\rho_c + \bar\rho_b} \tag{210}$$

is the total matter density perturbation.

We can now define the *baryon-CDM entropy perturbation*,

$$S_{cb} \equiv \delta_c - \delta_b\,, \tag{211}$$

which expresses how the perturbations in the two components deviate from each other. Subtracting Eq. (209) from (208) we get an equation for this entropy perturbation,

$$\ddot S_{cb} + 2H\dot S_{cb} = 0\,. \tag{212}$$

We assume that the primordial perturbations were adiabatic, so that we had $\delta_b = \delta_c$, i.e, $S_{cb} = 0$ at horizon entry. For large scales, which enter the horizon after decoupling, an S_{cb} never develops, so the evolution of the baryon perturbations is the same as CDM perturbations.

But for scales which enter before decoupling, an S_{cb} develops because the baryon perturbations are then coupled to the photon perturbations, whereas the CDM perturbations are not. After decoupling, $\delta_b \ll \delta_c$, since δ_c has been growing, while δ_b has been oscillating. The initial condition for Eqs. (208,209,212) is then $S_{cb}\sim\delta_c$ ("initial" time here being the time of decoupling

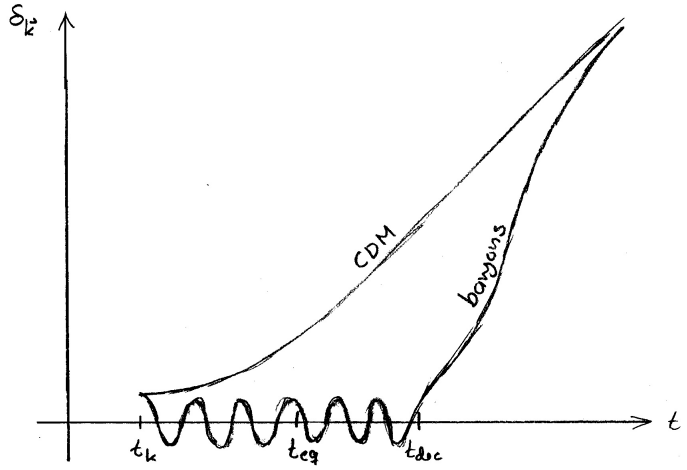


Figure 2: Evolution of the CDM and baryon density perturbations after horizon entry (at $t = t_k$). The figure is just schematic; the upper part is to be understood as having a $\sim$ logarithmic scale; the difference $\delta_c - \delta_b$ stays roughly constant, but the fractional difference becomes negligible as both δ_c and δ_b grow by a large factor.

$t_{\rm dec}$). During the matter-dominated epoch, when $a \propto t^{2/3}$, so that $H = 2/3t$, the solution for S_{cb} is

$$S_{cb} = A + Bt^{-1/3}\,, \tag{213}$$

whereas for δ_c it is, neglecting the effect of baryons on it, from Eq. (123),

$$\delta_c = Ct^{2/3} + Dt^{-1} \sim Ct^{2/3}\,. \tag{214}$$

We call the first term the "growing" and the second term the "decaying" mode (although for S_{cb} the "growing" mode is actually just constant). For δ_c the growing and decaying modes have been growing and decaying since horizon entry, so we can now drop the decaying part of δ_c.

To work out the precise initial conditions, we would need to work out the behavior of S_{cb} during decoupling. However, we really only need to assume that initially there is no strong cancellation between the growing and decaying modes, so that $S_{cb} = \delta_c - \delta_b$ either shrinks or stays roughly constant near the initial value of δ_c. While δ_c grows by a large factor, δ_b must follow it to keep the difference close to the initial small value of δ_c, so that $\delta_b/\delta_c \to 1$.

Thus the baryon density contrast δ_b grows to match the CDM density contrast δ_c (see Fig. 2), and we have eventually $\delta_b = \delta_c = \delta$ to high accuracy.

The baryon density perturbation begins to grow only after $t_{\rm dec}$. Before decoupling the radiation pressure prevents it. Without CDM it would grow only as $\delta_b \propto a \propto t^{2/3}$ after decoupling (during the matter-dominated period; the growth stops when the universe becomes dark energy dominated). Thus it would have grown at most by the factor $a_0/a_{\rm dec} = 1 + z_{\rm dec} \sim 1100$ after decoupling. In the anisotropy of the CMB we observe the baryon density perturbations at $t = t_{\rm dec}$. They are too small (about 10^{-4}) for a growth factor of 1100 to give the present observed large scale structure[21].

With CDM this problem was solved. The CDM perturbations begin to grow earlier, at $t \sim t_{\rm eq}$, and by $t = t_{\rm dec}$ they are much larger than the baryon perturbations. After decoupling

[21] This assumes adiabatic primordial perturbations, since we are seeing δ_γ, not δ_b. For a time, primordial baryon entropy perturbations $S_{b\gamma} = \delta_b - \frac{3}{4}\delta_\gamma$ were considered a possible explanation, but more accurate observations have ruled this model out.

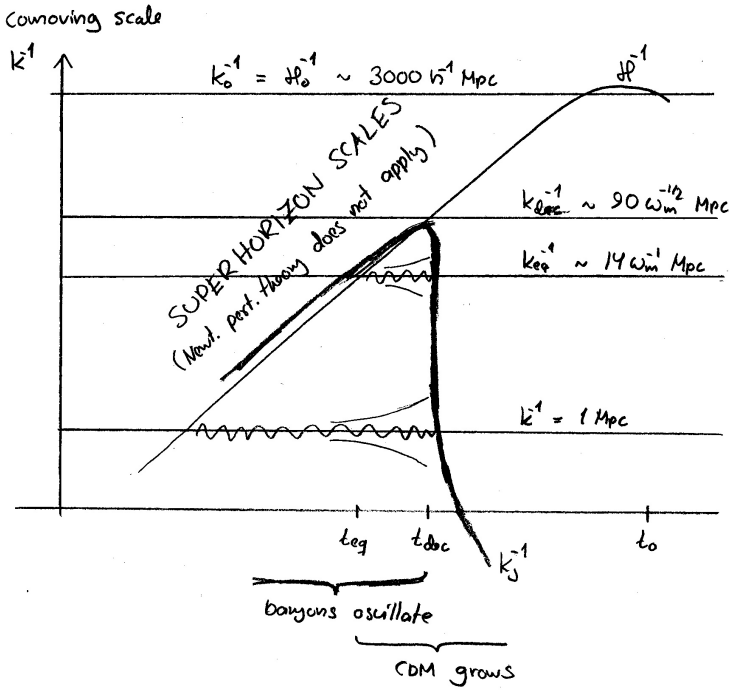


Figure 3: A figure summarizing the evolution of perturbations at different subhorizon scales. The baryon Jeans length k_J^{-1} drops precipitously at decoupling so that all cosmological scales became super-Jeans after decoupling, whereas all subhorizon scales were sub-Jeans before decoupling. The wavy lines symbolize the oscillation of baryon perturbations before decoupling, and the opening pair of lines around them symbolize the $\propto a$ growth of CDM perturbations after $t_{\rm eq}$. There is also an additional weaker (logarithmic) growth of CDM perturbations between horizon entry and $t_{\rm eq}$.

the baryons have lost the support from photon pressure and fall into the CDM gravitational potential wells, catching up with the CDM perturbations.

This allows the baryon perturbations to be small at $t = t_{\rm dec}$ and to grow after that by much more than the factor 10^3, matching observations. This is one of the reasons we are convinced that CDM exists.[22]

The whole subhorizon evolution history of all the different cosmological scales of perturbations is summarized by Fig. 3.

8.3.5 Late-time growth in the ΛCDM model

At late times, dark energy begins to accelerate the expansion, which will slow down the growth of density perturbations. In the ΛCDM model dark energy is just a constant vacuum energy, so it has no perturbations and thus affects just the background. The perturbations are in CDM and baryons, and we can ignore the pressure term in the Jeans equation, since at such small scales where baryon pressure gradients would be important, first-order perturbation theory is not valid anyway at late times. Thus we are facing a similar calculation as we did in Sec. 8.3.3, the solution of Eq. (189),

$$\ddot{\delta}_{\mathbf{k}} + 2H\dot{\delta}_{\mathbf{k}} - 4\pi G\bar{\rho}_m\delta_{\mathbf{k}} = 0\,, \tag{215}$$

[22]Historically, the above situation became clear in the 1980's when the upper limits to CMB anisotropy (which was finally discovered by COBE in 1992) became tighter and tighter. By today we have accurate detailed measurements of the structure of the CMB anisotropy which are compared to detailed calculations including the CDM so the argument is raised to a different level—instead of comparing just two numbers we are now comparing entire power spectra (to be discussed later).

with $\delta_b = \delta_c = \delta$, but instead of radiation we have now vacuum energy contributing to the background solution, which is the Concordance Model discussed in Cosmology I (Chapter 3):

$$a(t) \;=\; \left(\frac{\Omega_m}{\Omega_\Lambda}\right)^{1/3} \sinh^{2/3}\left(\tfrac{3}{2}\sqrt{\Omega_\Lambda}H_0 t\right) . \tag{216}$$

The Hubble parameter is given by

$$H = H_0\sqrt{\Omega_m a^{-3} + \Omega_\Lambda}\,. \tag{217}$$

Again, it is better to use the scale factor as time coordinate. The difference in the power of a in the behavior of the two density components is now 3 instead of 1, which makes the calculation more difficult. We follow here Dodelson[5]. After the change of variable from t to a, (215) becomes (**exercise**)

$$\delta'' + \left(\frac{H'}{H} + \frac{3}{a}\right)\delta' - \frac{3\Omega_m}{2a^5}\left(\frac{H_0}{H}\right)^2\delta = 0\,, \tag{218}$$

where $' \equiv d/da$. The decaying solution is

$$\delta \;\propto\; H \;\propto\; \sqrt{\Omega_m a^{-3} + \Omega_\Lambda} \tag{219}$$

and the growing solution is

$$\delta \;\propto\; H\int^a \frac{dx}{H^3x^3} \;\propto\; \sqrt{\Omega_m a^{-3} + \Omega_\Lambda}\int^a \frac{x^{3/2}dx}{\left(1 + \frac{\Omega_\Lambda}{\Omega_m}x^3\right)^{3/2}} \tag{220}$$

The effect of changing the lower limit of integration can be incorporated in the decaying solution; so we can set the lower limit to 0. (Equation (218) is valid in general for matter perturbations with an additional smooth background component. The first forms of the solutions (219) and (220) are valid when the smooth component is vacuum energy or negative curvature.)

In the limit $a \ll 1$, or rather, $\Omega_\Lambda \ll \Omega_m a^{-3}$, the decaying solution becomes

$$\delta \propto a^{-3/2} \propto t^{-1} \tag{221}$$

and the growing solution becomes

$$\delta \propto a \propto t^{2/3} \tag{222}$$

the familiar results for the matter-dominated universe from Sec. 8.2.6. We can ignore the decaying mode, since it has become completely negligible when the vacuum energy begins to have an effect.

To fix the proportionality coefficient in the growing mode, we write it as

$$\delta = A\left(\Omega_m a^{-3} + \Omega_\Lambda\right)^{1/2}\int_0^a \frac{x^{3/2}dx}{\left(1 + \frac{\Omega_\Lambda}{\Omega_m}x^3\right)^{3/2}} \tag{223}$$

and note that in the limit $\Omega_\Lambda \ll \Omega_m a^{-3}$ it becomes

$$\delta \;\approx\; A\Omega_m^{1/2}a^{-3/2}\int_0^a x^{3/2}dx \;=\; \frac{2}{5}\Omega_m^{1/2}Aa\,. \tag{224}$$

At $a = a_0 = 1$ this would give

$$\frac{2}{5}\Omega_m^{1/2}A \equiv \tilde{\delta} \quad\Rightarrow\quad A = \frac{5}{2}\Omega_m^{-1/2}\tilde{\delta}\,, \tag{225}$$

where we have defined $\tilde{\delta}$ as the value δ would have "now"[23] if there were no vacuum energy, i.e., the universe had stayed matter dominated.

Thus we write (223) as

$$\boxed{\delta \;=\; \tilde{\delta}\,\frac{5}{2}\left(a^{-3}+\frac{\Omega_\Lambda}{\Omega_m}\right)^{1/2}\int_0^a \frac{x^{3/2}dx}{\left(1+\frac{\Omega_\Lambda}{\Omega_m}x^3\right)^{3/2}}\,.} \tag{226}$$

Unfortunately, the integral in (226) does not give an elementary function. (I think it is a so-called hypergeometric function, which does not give much useful information compared to just integrating (226) numerically.) We can see that at late (future) times, when $a \gg 1$, there is very little growth, since the factor outside the integral approaches a constant and for any $a_1 \gg 1$ and $a_2 \gg 1$, the contribution to the integral,

$$\int_{a_1}^{a_2} \frac{x^{3/2}dx}{\left(1+\frac{\Omega_\Lambda}{\Omega_m}x^3\right)^{3/2}} \;\approx\; \left(\frac{\Omega_m}{\Omega_\Lambda}\right)^{3/2}\int_{a_1}^{a_2} x^{-3}\,dx \;=\; \frac{1}{2}\left(\frac{\Omega_m}{\Omega_\Lambda}\right)^{3/2}\left(a_1^{-2}-a_2^{-2}\right) \tag{227}$$

is very small.

It turns out that the integral can be done if we extend it to the infinite future (**exercise**) : As $a \to \infty$,

$$\delta \;\to\; \delta(\infty) \;\equiv\; \tfrac{5}{2}\tilde{\delta}\left(a^{-3}+\frac{\Omega_\Lambda}{\Omega_m}\right)^{1/2}\int_0^\infty \frac{x^{3/2}dx}{\left(1+\frac{\Omega_\Lambda}{\Omega_m}x^3\right)^{3/2}} \;=\; \tfrac{5}{6}\tilde{\delta}\left(\frac{\Omega_m}{\Omega_\Lambda}\right)^{1/3} B(\tfrac{5}{6},\tfrac{2}{3})\,, \tag{228}$$

where

$$B(p,q) \;\equiv\; \int_0^1 t^{p-1}(1-t)^{q-1}dt \;=\; \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \tag{229}$$

is the beta function and

$$B\left(\tfrac{5}{6},\tfrac{2}{3}\right) \approx 1.725\,. \tag{230}$$

Thus the perturbations "freeze", i.e., approach a final value

$$\delta(\infty) = 1.437\left(\frac{\Omega_m}{\Omega_\Lambda}\right)^{1/3}\tilde{\delta}\,. \tag{231}$$

which for $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$ gives

$$\delta(\infty) = 1.084\,\tilde{\delta}\,, \tag{232}$$

i.e., the perturbations will never become much stronger than what they in the matter-dominated model would be already "now". To get the present density perturbation $\delta(a=1)$ one has to do (226) numerically. This is done in Fig. 4, from which one can read that $\delta(a=1) \approx 0.78\,\tilde{\delta}$.

For perturbations that entered horizon well before matter-radiation equality $t_{\rm eq}$, we have from (196) that

$$\tilde{\delta} \approx \delta_{\rm prim}\left(1+\frac{3}{2a_{\rm eq}}\right)8\ln\left(\frac{k}{6k_{\rm eq}}\right)\,, \tag{233}$$

[23]Note that we defined "now" as $a = a_0 = 1$, not as $t = t_0$; or in more physical terms as $T = T_0 = 2.725\,\mathrm{K}$. The comparison situation ($\tilde{}$) we have in mind is that the early universe (where vacuum energy has no effect) is the same as in the ΛCDM model, but there is no vacuum energy to accelerate the expansion at late times, so that by "now" the expansion rate, i.e., H_0 is smaller than we observe in reality. The present matter density $\rho_{m0} = (3/8\pi G)\Omega_m H_0^2$ is the same as in the ΛCDM model, but $\tilde{\Omega}_m = 1$, so $\tilde{H}_0 = \Omega_m^{1/2}H_0$. The age of the universe is $\tilde{t}_0 = \frac{2}{3}\tilde{H}_0^{-1} = \frac{2}{3}\Omega_m^{-1/2}H_0^{-1}$, which for $h = 0.7$ and $\Omega_m = 0.3$ gives $\tilde{t}_0 = 17.0\times10^9$ years, instead of the $t_0 = 13.5\times10^9$ years of the ΛCDM model.

assuming that this is still $\ll 1$ so that first-order perturbation theory remains valid, and remembering that this was an approximate result that ignored the effect of the baryon-photon oscillations on CDM during the radiation-dominated epoch. For larger scales, and for what $\delta_{\rm prim}$ is, we need the remaining sections of this chapter.

For $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $h = 0.7$, we have $k_{\rm eq}^{-1} = 65\,h^{-1}$Mpc and $a_{\rm eq} = 1/3603$. Equation (233) gives then for the scale $k^{-1} = 8\,h^{-1}$Mpc,

$$\tilde{\delta} \approx 13\,000\delta_{\rm prim} \qquad \text{and} \qquad \delta(a=1) \approx 10\,000\delta_{\rm prim}. \tag{234}$$

Observationally, the variance of the top-hat-filtered density field of the galaxy distribution today is ≈ 1 at this scale. Because of the galaxy bias b_g, the corresponding variance for the matter distribution is less by factor b_g^{-2}, but still not far from 1, meaning that the linear perturbation theory approximation is beginning to break.

8.3.6 Growth function

Inside the horizon, during the matter- and dark-energy-dominated epochs the linear growth of perturbations is independent of scale (once the decaying mode has died out and ignoring the subcosmological scales where pressure gradients have a role). Thus it can be described by a function that depends on time (or scale factor, or redshift) only, called the *growth function*,

$$D(a) \equiv \frac{\delta(a)}{\delta_{\rm ref}} \tag{235}$$

where $\delta(a)$ is the density perturbation ($\delta_{\mathbf{k}}$ or $\delta(\mathbf{x})$; $D(a)$ is the same function for any $\mathbf{k}$ or $\mathbf{x}$) when scale factor is a and $\delta_{\rm ref}$ is it at some reference time. The choice of reference time fixes the normalization of D. We define the *growth rate*

$$f \equiv \frac{d\ln D}{d\ln a} = \frac{d\ln\delta}{d\ln a} = \frac{a}{\delta}\frac{d\delta}{da}\,, \tag{236}$$

which is independent of this normalization.

For the ΛCDM model of Sec. 8.3.5, we get from (226) (**exercise**)

$$f(a) \;=\; \frac{1}{1+\frac{\Omega_\Lambda}{\Omega_m}a^3}\left(\tfrac{5}{2}a\frac{\tilde{\delta}}{\delta} - \tfrac{3}{2}\right) \tag{237}$$

It turns out that a good approximation for the growth rate is

$$f(a) \approx \Omega_m(a)^\gamma\,, \qquad \text{where } \gamma = 0.55\,, \tag{238}$$

where γ is called the *growth index*. (This result assumes General Relativity, and the measurement of the growth index from galaxy surveys is a way of testing gravity theory.) We plot D, f, and the approximation (238) for ΛCDM in Fig. 4.

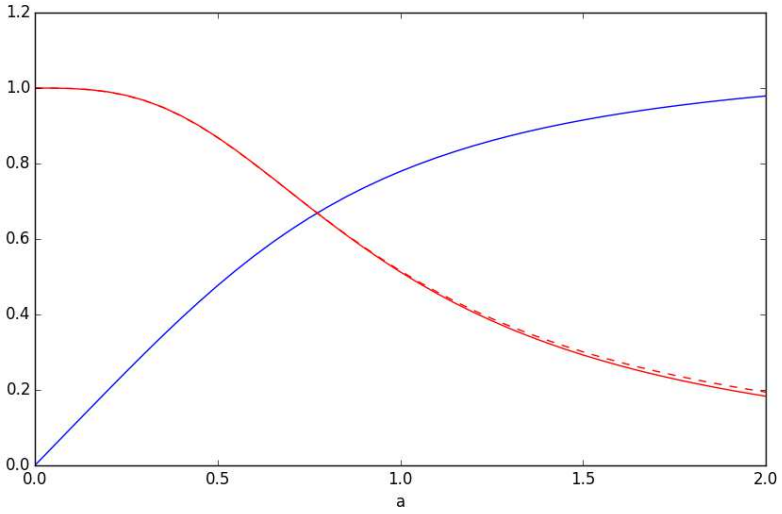


Figure 4: The growth function $D(a)$ (blue, with normalization $\delta_{\rm ref} = \tilde{\delta}$), growth rate $f(a)$ (red) and the approximation (238) (red, dashed) for ΛCDM with $\Omega_m = 0.3$.

8.4 Relativistic perturbation theory

For scales comparable to, or larger than the Hubble scale, Newtonian perturbation theory does not apply, because we can no more ignore the curvature of spacetime. Therefore we need to use (general) relativistic perturbation theory. Instead of the Newtonian equations of gravity and fluid mechanics, the fundamental equation is now the *Einstein equation* of general relativity (GR). We assume a background solution, which is homogeneous and isotropic, i.e., a solution of the Friedmann equations, and study small perturbations around it. This particular choice of the background solution means that we are doing a particular version of relativistic perturbation theory, called *cosmological perturbation theory*.

The evolution of the perturbations while they are well outside the horizon is simple, but the mathematical machinery needed for its description is complicated. This is due to the coordinate freedom of general relativity. For the background solution we had a special coordinate system (time slicing) of choice, the one where the $t = const$ slices are homogeneous. The perturbed universe is no more homogeneous, it is just "close to homogeneous", and therefore we no more have a unique choice for the coordinate system. We should now choose a coordinate system where the universe is close to homogeneous on the time slices, but there are many different possibilities for such slicing. This freedom of choosing the coordinate system in the perturbed universe is called *gauge freedom*, and a particular choice is called a *gauge*.[24] The most important part of the choice of gauge is the choice of the time coordinate, because it determines the slicing of the spacetime into $t = const$ slices, "universe at time t". Sometimes the term 'gauge' is used to refer only to this slicing.

Because the perturbations are defined in terms of the chosen coordinate system, they look different in different gauges. We can, for example, choose the gauge so that the perturbation in one scalar quantity, e.g., proper energy density, disappears, by choosing the $\rho = const$ 3-surfaces as the time slices (this is called "the uniform energy density gauge").

[24] If you are familiar with *gauge field theories*, like electrodynamics, the concept of 'gauge' may look different here. The mathematical similarity appears when the perturbation equations are developed. In relativistic perturbation theory gauge has this geometric origin (this is where the use of the word "gauge" comes from), unlike electrodynamics.

The true nature of gravitation is spacetime curvature, so perturbations should be described in terms of curvature.

We leave the actual development of cosmological perturbation theory to a more advanced course, and just summarize here some basic concepts and results.

In the Newtonian theory gravity was represented by a single function, the gravitational potential Φ. In GR, gravity is manifested in the geometry of spacetime, described in terms of the metric. Thus in addition to the density, pressure, and velocity perturbations, we have a perturbation in the metric. The perturbed metric tensor is

$$g_{\mu\nu} = \bar{g}_{\mu\nu} + \delta g_{\mu\nu} \,. \tag{239}$$

For the background metric, $\bar{g}_{\mu\nu}$, we choose that of the flat Friedmann-Robertson-Walker universe,

$$ds^2 = \bar{g}_{\mu\nu} dx^\mu dx^\nu = -dt^2 + a(t)^2(dx^2 + dy^2 + dz^2) \tag{240}$$

The restriction to the flat case is an important simplification, because it allows us to Fourier expand our perturbations in terms of plane waves.[25] Fortunately the real universe appears to be flat, or at least close to it. And earlier it was even flatter. Inflation predicts a flat universe.

For the metric perturbation, we have now 10 functions $\delta g_{\mu\nu}(t, \mathbf{x})$. So there appears to be ten degrees of freedom. Four of them are not physical degrees of freedom, since they just correspond to our freedom in choosing the four coordinates. So there are 6 real degrees of freedom.

Two of these metric degrees of freedom couple to density and pressure perturbations and the irrotational velocity perturbation. These are the *scalar perturbations*. Two couple to the rotational velocity perturbation to make up the *vector perturbations*. The remaining two are not coupled to the cosmic fluid at all[26], and are called *tensor perturbations*. They are gravitational waves, which do not exist in Newtonian theory.

The vector perturbations decay in time, and are not produced by inflation, so they are the least interesting. Although the tensor perturbations also are not related to growth of structure, they are produced in inflation and affect the cosmic microwave background anisotropy and polarization. Different inflation models produce tensor perturbations with different amplitudes and spectral indices (to be explained later), so they are an important diagnostic of inflation. No tensor perturbations have been detected in the CMB so far, but they could be detected in the future with more sensitive instruments if their amplitude is large enough.[27]

Since the three kinds of perturbations evolve independently of each other, they can be studied separately. We shall first concentrate on the scalar perturbations, returning to the tensor perturbations later.

8.4.1 Gauges for scalar perturbations

Consider now scalar perturbations. The gauges discussed in the following assume scalar perturbations.

The perturbations appear different in the different gauges. When needed, we use superscripts to indicate in which gauge the quantity is defined: C for the comoving gauge and N for the Newtonian gauge. Some other gauges are the synchronous gauge (S), spatially flat gauge (Q), and the uniform energy density gauge (U).

[25] In the Newtonian case this restriction was not necessary, and we could apply it to any Friedmann model, as there is no curvature of spacetime in the Newtonian view, and only the expansion law $a(t)$ of the Friedmann model is used. The Newtonian theory of course is only valid for small scales were the curvature can indeed be ignored.

[26] This is true in first-order perturbation theory in the perfect fluid approximation, but not in general.

[27] Typically, large-field inflation models produce tensor perturbations with much larger amplitude than small-field inflation models. In the latter case they are likely to be too small to be detectable.

There are two common ways to specify a gauge, i.e., the choice of coordinate system in the perturbed universe:

- A statement about the relation of the coordinate system to the fluid perturbation. This will lead to some condition on the metric perturbations.
- A statement about the metric perturbations. This will then lead to some condition on the coordinate system.

The two gauges (C and N) we shall refer to in the following, give an example of each.

The *comoving gauge* is defined so that the space coordinate lines $\mathbf{x} = const$ follow fluid flow lines, and the time slice, the $t = const$ hypersurface is orthogonal to them. Thus the velocity perturbation is zero in this gauge,

$$v^C = 0\,. \tag{241}$$

The *conformal-Newtonian* gauge, also called the longitudinal gauge, or the zero-shear gauge, and sometimes, for short, just the Newtonian gauge, is defined by requiring the metric to be of the form

$$ds^2 = -(1+2\Phi)dt^2 + a^2(1-2\Psi)(dx^2+dy^2+dz^2)\,. \tag{242}$$

This means that we require

$$\delta g_{0i} = 0, \quad \delta g_{11} = \delta g_{22} = \delta g_{33}, \quad \text{and} \quad g_{ij} = 0 \quad \text{for} \quad i \neq j\,. \tag{243}$$

(This is possible for scalar perturbations). The two metric perturbations, $\Phi(t,\mathbf{x})$ and $\Psi(t,\mathbf{x})$ are called *Bardeen potentials*.[28] Φ is also called the Newtonian potential, since in the Newtonian limit ($k \gg \mathcal{H}$ and $p \ll \rho$), it becomes equal to the Newtonian gravitational potential perturbation. Thus we can use the same symbol for it. Ψ is also called the Newtonian curvature perturbation, because it determines the curvature of the 3-dimensional $t = const$ subspaces, which are flat in the unperturbed universe (since it is the flat FRW universe).

It turns out that the difference $\Phi - \Psi$ is caused only by anisotropic stress (or anisotropic pressure). We shall here consider only the case of a perfect fluid. For a perfect fluid the pressure (or stress) is necessary isotropic. Thus we have only a single metric perturbation[29]

$$\Psi = \Phi \tag{244}$$

The density perturbations in these two gauges become equal in the limit $k \gg \mathcal{H}$, and we can then identify them with the "usual" density perturbation δ of Newtonian theory.

8.4.2 Evolution at superhorizon scales

When the perturbations are outside the horizon (meaning that the wavelength of the Fourier mode we are considering is much longer than the Hubble length), very little happens to them, and we can find quantities which remain constant for superhorizon scales. Such a quantity is the (comoving) curvature perturbation $\mathcal{R}(\mathbf{x})$, which describes how curved is the $t = const$ slice in

[28] Warning: The sign conventions for Ψ differ, and many authors call them Ψ and Φ instead.

[29] In reality, neutrinos develop anisotropic pressure after neutrino decoupling. Therefore the two Bardeen potentials actually differ from each other by about 10 % between the times of neutrino decoupling and matter-radiation equality. After the universe becomes matter-dominated, the neutrinos become unimportant, and Ψ and Φ rapidly approach each other. The same happens to photons after photon decoupling, but the universe is then already matter-dominated, so they do not cause a significant $\Psi - \Phi$ difference.

the comoving gauge.[30] *For adiabatic perturbations, the curvature perturbation $\mathcal{R}$ stays constant in time outside the horizon.*

Using gauge transformation equations $\mathcal{R}$ can be related to the metric in the Newtonian gauge. The result is

$$\mathcal{R} \;=\; -\frac{5+3w}{3+3w}\Phi - \frac{2}{3+3w}H^{-1}\dot{\Phi}\,, \tag{247}$$

where $w \equiv \bar{p}/\bar{\rho}$.

Because $\mathcal{R}_{\mathbf{k}}$ stays constant while $k \ll \mathcal{H}$, it is a very useful quantity for "carrying" the perturbations from their generation at horizon exit during inflation to horizon entry at later times. We now define *the primordial perturbation* to refer to the perturbation at the epoch when it is well outside the horizon. For adiabatic perturbations, the primordial perturbation is completely characterized by the set of these constant values $\mathcal{R}_{\mathbf{k}}$. We shall later discuss how the primordial perturbation is generated by inflation, and how these superhorizon values $\mathcal{R}_{\mathbf{k}}$ are determined by it.

However, we would like to describe the perturbation in more "familiar" terms, the gravitational potential perturbation Φ and the density perturbation δ. When $\mathcal{R}_k$ remains constant this turns out to be easy. Eq. (247) can be written as a differential equation for $\Phi_{\mathbf{k}}$,

$$\frac{2}{3}H^{-1}\dot{\Phi}_{\mathbf{k}} + \frac{5+3w}{3}\Phi_{\mathbf{k}} \;=\; -(1+w)\mathcal{R}_{\mathbf{k}}\,. \tag{248}$$

During any period, when also $w = const$, the solution of this equation is

$$\Phi_{\mathbf{k}} \;=\; -\frac{3+3w}{5+3w}\mathcal{R}_{\mathbf{k}} + \text{a decaying part}\,. \tag{249}$$

Thus, after w has stayed constant for some time, the Bardeen potential has settled to the constant value

$$\boxed{\Phi_{\mathbf{k}} = -\frac{3+3w}{5+3w}\mathcal{R}_{\mathbf{k}}} \qquad (w = const)\,. \tag{250}$$

In particular, we have the relations

$$\Phi_{\mathbf{k}} \;=\; -\frac{2}{3}\mathcal{R}_{\mathbf{k}} \qquad (\text{rad.dom},\ w = \tfrac{1}{3}) \tag{251}$$

$$\Phi_{\mathbf{k}} \;=\; -\frac{3}{5}\mathcal{R}_{\mathbf{k}} \qquad (\text{mat.dom},\ w = 0)\,. \tag{252}$$

After the potential has entered the horizon, we can use the Newtonian perturbation theory result, Eq. (106), which gives the density perturbation as

$$\delta_{\mathbf{k}} \;=\; -\left(\frac{k}{a}\right)^2 \frac{\Phi_{\mathbf{k}}}{4\pi G\bar{\rho}} \;=\; -\frac{2}{3}\left(\frac{k}{aH}\right)^2 \Phi_{\mathbf{k}} \;=\; -\frac{2}{3}\left(\frac{k}{\mathcal{H}}\right)^2 \Phi_{\mathbf{k}}\,, \tag{253}$$

[30] Technically, $\mathcal{R}$ is defined in terms of the trace of the space part of the comoving gauge metric perturbation ($-\Psi$ is the corresponding quantity in the Newtonian gauge), and it is related to the scalar curvature ${}^{(3)}R^C$ of the *comoving gauge* time slice (the ${}^{(3)}$ reminds us that we are considering a 3-dimensional subspace, and the C refers to the comoving gauge) so that

$${}^{(3)}R^C = -4a^{-2}\nabla^2\mathcal{R}\,. \tag{245}$$

For Fourier components we have then that

$$\mathcal{R}_{\mathbf{k}} \equiv \frac{1}{4}\left(\frac{a}{k}\right)^2 {}^{(3)}R^C_{\mathbf{k}}\,. \tag{246}$$

Another similar quantity is the (uniform-density-gauge) curvature perturbation ζ that is defined the same way, but for the uniform-density-gauge time slice. For superhorizon scales they are equal, $\mathcal{R} = \zeta$ (in the limit $k \ll \mathcal{H}$).

where we used the background relation

$$H^2 = \frac{8\pi G}{3}\bar{\rho} \quad \Rightarrow \quad \boxed{4\pi G\bar{\rho} = \frac{3}{2}H^2}. \tag{254}$$

The problem is to get $\Phi_{\mathbf{k}}$ from its superhorizon epoch where it is constant (as long as $w = const$) through the horizon entry to its subhorizon epoch where it evolves according to Newtonian theory. For scales k which enter while the universe is matter dominated, this is easy, since in this case Φ_k stays constant the whole time (until dark energy becomes important).

Thus we can relate the constant values of $\Phi_{\mathbf{k}}$, and the corresponding subhorizon density perturbations $\delta_{\mathbf{k}}$ during the matter-dominated epoch to the primordial perturbations $\mathcal{R}_{\mathbf{k}}$ by

$$\begin{aligned} \Phi_{\mathbf{k}} &= -\frac{3}{5}\mathcal{R}_{\mathbf{k}} \qquad \text{(mat.dom)} \\ \delta_{\mathbf{k}} &= -\frac{2}{3}\left(\frac{k}{\mathcal{H}}\right)^2 \Phi_{\mathbf{k}} \;=\; \frac{2}{5}\left(\frac{k}{\mathcal{H}}\right)^2 \mathcal{R}_{\mathbf{k}} \;\propto\; \frac{1}{(aH)^2} \propto t^{2/3} \propto a \end{aligned} \tag{255}$$

Note that by $\mathcal{R}_{\mathbf{k}}$ we refer always to the constant primordial value, when we use it in equations, like (255), that give other quantities at later times.

For perturbations which enter during the radiation-dominated epoch, the potential $\Phi_{\mathbf{k}}$ does not stay constant. We learned earlier, that in this case the density perturbations oscillate with roughly constant amplitude, which means that the amplitude for the potential Φ must decay $\propto a^2\bar{\rho} \propto a^{-2}$. This oscillation applies to the baryon-photon fluid, whereas the CDM density perturbations grow slowly. After the universe becomes matter dominated, it is these CDM perturbations that matter.

We shall now make a crude estimate how the amplitudes of these smaller-scale perturbations during the matter-dominated epoch are related to the primordial perturbations. These perturbations enter during the radiation-dominated epoch. Assume that the relation $\Phi_{\mathbf{k}} = -\frac{2}{3}\mathcal{R}_{\mathbf{k}}$ holds all the way to horizon entry ($k = \mathcal{H}$). Assume then that the Newtonian relation (253) holds already. Then

$$\delta_{\mathbf{k}} \;\approx\; -\frac{2}{3}\left(\frac{k}{\mathcal{H}}\right)^2 \Phi_{\mathbf{k}} \;=\; -\frac{2}{3}\Phi_k \approx \frac{4}{9}\mathcal{R}_{\mathbf{k}} \tag{256}$$

at horizon entry. The universe is now radiation-dominated, and therefore $\delta_{r\mathbf{k}} = \delta_{\mathbf{k}}$. We are assuming primordial adiabatic perturbations and therefore the adiabatic relations $\delta_c = \frac{3}{4}\delta_r$, $\delta_\gamma = \delta_r$ hold at superhorizon scales. Assume that these relations hold until horizon entry. After that $\delta_{\gamma\mathbf{k}}$ begins to oscillate, whereas $\delta_{c\mathbf{k}}$ grows slowly. Thus we have that at horizon entry

$$\delta_{c\mathbf{k}} \approx \frac{3}{4}\delta_{\mathbf{k}} \approx \frac{1}{3}\mathcal{R}_{\mathbf{k}}\,. \tag{257}$$

Ignoring the slow growth of δ_c we get that $\delta_{c\mathbf{k}}$ stays at this value until the universe becomes matter-dominated at $t = t_{\rm eq}$, after which we can approximate $\delta_{\mathbf{k}} \approx \delta_{c\mathbf{k}}$ and $\delta_{\mathbf{k}}$ begins to grow according to the matter-dominated law, $\propto 1/\mathcal{H}^2$.

Thus

$$\delta_{\mathbf{k}}(t_{\rm eq}) \approx \frac{1}{3}\mathcal{R}_{\mathbf{k}} \tag{258}$$

and

$$\delta_{\mathbf{k}}(t) \approx \frac{1}{3}\mathcal{R}_{\mathbf{k}}\left(\frac{\mathcal{H}_{\rm eq}}{\mathcal{H}}\right)^2 = \frac{1}{3}\mathcal{R}_{\mathbf{k}}\left(\frac{k_{\rm eq}}{\mathcal{H}}\right)^2 \qquad \text{for } t > t_{\rm eq}\,, \tag{259}$$

as long as the universe stays matter dominated.

8.4.3 Transfer function

For large scales ($k \ll k_{\rm eq}$) which enter the horizon during the matter-dominated epoch, we got

$$\delta_{\mathbf{k}}(t) = \frac{2}{5}\left(\frac{k}{\mathcal{H}}\right)^2 \mathcal{R}_{\mathbf{k}} \qquad (k \ll k_{\rm eq})\,, \tag{260}$$

for as long as the universe stays matter dominated.

This is a simple result, and we use this as a reference for the more complicated result at smaller scales. That is, we define a *transfer function* $T(k,t)$ so that

$$\delta_{\mathbf{k}}(t) = \frac{2}{5}\left(\frac{k}{\mathcal{H}}\right)^2 T(k,t)\mathcal{R}_{\mathbf{k}} \tag{261}$$

where $\mathcal{R}_{\mathbf{k}}$ refers to the primordial perturbation. Thus by definition $T(k,t) = 1$ for $k \ll k_{\rm eq}$.[31]

Using the rough estimate from the previous subsection we get that

$$T(k,t) \approx \frac{5}{6}\left(\frac{k_{\rm eq}}{k}\right)^2 \tag{262}$$

during the matter-dominated epoch, where we can drop the factor $\frac{5}{6}$, since this is anyway just a rough estimate.

Once we are well into the matter-dominated era, perturbations at all scales grow $\propto a \propto 1/(aH)^2$ and the transfer function becomes independent of time,[32]

$$\begin{aligned} T(k) &= 1 && k \ll k_{\rm eq} \\ T(k) &\sim \left(\frac{k_{\rm eq}}{k}\right)^2 && k \gg k_{\rm eq} \end{aligned} \tag{263}$$

A more accurate calculation, including the gravitational effect of baryon-photon oscillations on the CDM perturbations, and assuming adiabatic primordial perturbations, adds a logarithmic growth factor and gives

$$T(k) \approx 12\left(\frac{k_{\rm eq}}{k}\right)^2 \ln\left(\frac{k}{6k_{\rm eq}}\right) \qquad k \gg k_{\rm eq} \tag{264}$$

Note that that logarithm is negative for $k < 6k_{\rm eq}$; the equation is not supposed to apply yet for this low k.

According to present understanding, the universe becomes dark energy dominated as we approach the present time. The equation-of-state parameter w begins to decrease (becomes negative) and therefore Φ begins to change again. The growth of the density perturbations is slowed down as we saw in Sec. 8.3.5. The effect is not very big since the universe has expanded by less than a factor of 2 after the onset of dark energy domination, but it is important in detailed matching of observations and theory.

Our calculation of the growth of structure has been just a rough approximation. A detailed calculation including all relevant effects has to be done numerically. There are publicly available computer programs (such as CMBFAST and CAMB) that do this (you give your favorite values for the cosmological parameters as input). The exact result can be given in form of the transfer function $T(k)$ we defined above. The main corrections to our simple results, (263) and (264), include:

[31] With the given definition for $T(k,t)$, this holds for $t \ll t_0$, i.e, before we entered the present dark-energy-dominated epoch.

[32] We shall later define other transfer functions, but this is *the* transfer function $T(k)$ of structure formation theory. It relates the perturbations inside the horizon during the matter-dominated epoch to the primordial perturbations, and it is independent of time.

1) The transition from $k \ll k_{\rm eq}$ behavior to $k \gg k_{\rm eq}$ behavior is, of course, smooth.

2) The effect of *baryon acoustic oscillations* (i.e., the oscillations of $\delta_{b\gamma}$ before decoupling, which leave a trace in δ_b) shows up as a small-amplitude wavy pattern in the $k > k_{\rm eq}$ part of the transfer function, since different modes k were at a different phase of the oscillation when that ended around $t_{\rm dec}$.

We have calculated everything using linear perturbation theory. This breaks down when the perturbations become large, $\delta(\mathbf{x}) \sim 1$. We say that the perturbation becomes nonlinear. This has happened for the smaller scales, $k^{-1} < 10$ Mpc by now. When the perturbation becomes nonlinear, i.e., an overdense region becomes significantly denser (say, twice as dense) as the average density of the universe, it collapses rapidly, and forms a gravitationally bound structure, e.g. a galaxy or a cluster of galaxies. Further collapse is prevented by the angular momentum of the structure. Galaxies in a cluster and stars (and CDM particles) in a galaxy orbit around the center of mass of the bound structure.

8.4.4 Tensor perturbations

In addition to scalar and vector perturbations, in general relativistic perturbation theory we have tensor perturbations. They have the nice property that we do not have to worry about different gauges, since they are gauge invariant in the sense that, if we first do a gauge transformation and then separate out the scalar, vector, and tensor parts, the tensor part has remained unchanged.

These are perturbations of the metric that for one Fourier mode take the form

$$\begin{aligned} ds^2 &= -dt^2 + a(t)^2\left[(1+h)dx^2 + (1-h)dy^2 + dz^2\right] \\ &= a(\eta)^2\left[-d\eta^2 + (1+h)dx^2 + (1-h)dy^2 + dz^2\right] \end{aligned} \tag{265}$$

where

$$h = h_{\mathbf{k}}(t)e^{ikz} \tag{266}$$

is the perturbation and η is conformal time. In (265) we have chosen the z axis in the direction of the wave vector, so that $\mathbf{k} = k\hat{k}$ and $\mathbf{k}\cdot\mathbf{x} = kz$. Since the metric is a real quantity, in (265) and (271) h should be interpreted as the real part of h; like one should always do when one makes physical interpretations for a single Fourier mode. Remember that when one sums over Fourier components the imaginary parts of $h_{\mathbf{k}}(t)e^{ikz} + h_{-\mathbf{k}}(t)e^{-ikz}$ cancel since $h_{-\mathbf{k}} = h^*_{\mathbf{k}}$, and thus the imaginary parts have no physical significance, they are just a mathematical convenience.

The effect of the tensor perturbation is to stretch space in one direction (here x if h is positive) and compress it in the other direction (here y) orthogonal to the wave vector of the Fourier mode. In (265) we also chose the orientation of the x and y axes so that they correspond to these stretch/compress directions. But of course the perturbation could be oriented differently. We get the other possibilities by rotating the pattern around the wave vector $\mathbf{k}$ by some angle φ, which is mathematically equivalent to rotating the coordinate system by angle $-\varphi$.

In matrix form the metric is

$$[g_{\mu\nu}] = a^2\begin{bmatrix} -1 & & & \\ & 1+h & & \\ & & 1-h & \\ & & & 1 \end{bmatrix} \tag{267}$$

After rotation by φ around the z axis it becomes

$$[g_{\mu\nu}] = a^2\begin{bmatrix} 1 & & & \\ & \cos\varphi & -\sin\varphi & \\ & \sin\varphi & \cos\varphi & \\ & & & 1 \end{bmatrix}\begin{bmatrix} -1 & & & \\ & 1+h & & \\ & & 1-h & \\ & & & 1 \end{bmatrix}\begin{bmatrix} 1 & & & \\ & \cos\varphi & \sin\varphi & \\ & -\sin\varphi & \cos\varphi & \\ & & & 1 \end{bmatrix} \tag{268}$$

Rotation by 45°, i.e., $\cos\varphi = \sin\varphi = 1/\sqrt{2}$, gives

$$[g_{\mu\nu}] = a^2 \begin{bmatrix} -1 & & & \\ & 1 & h & \\ & h & 1 & \\ & & & 1 \end{bmatrix} \tag{269}$$

We call (267) the + mode and (269) the × mode. An arbitrary orientation of the stretch/compress pattern can be obtained as a linear combination of these two modes, so that the general form of the tensor perturbation is

$$[g_{\mu\nu}] = a^2 \begin{bmatrix} -1 & & & \\ & 1+h_+ & h_\times & \\ & h_\times & 1-h_+ & \\ & & & 1 \end{bmatrix} \tag{270}$$

or

$$\begin{aligned} ds^2 &= -dt^2 + a(t)^2 \left[(1+h_+)dx^2 + 2h_\times dxdy + (1-h_+)dy^2 + dz^2\right] \\ &= a(\eta)^2 \left[-d\eta^2 + (1+h_+)dx^2 + 2h_\times dxdy + (1-h_+)dy^2 + dz^2\right] \end{aligned} \tag{271}$$

for a Fourier mode in the z direction. Thus we have two Fourier amplitudes $h_{+\mathbf{k}}(t)$ and $h_{\times\mathbf{k}}(t)$ for each wave vector $\mathbf{k}$. In the following we mostly write just $h(t)$ to represent an arbitrary such mode.

The evolution equation for $h(t)$,

$$\ddot{h} + 3H\dot{h} + \left(\frac{k}{a}\right)^2 h = 0 \qquad \Leftrightarrow \qquad H^{-2}\ddot{h} + 3H^{-1}\dot{h} + (k/\mathcal{H})^2\, h = 0\,, \tag{272}$$

can be obtained from the Einstein equation. This derivation is beyond the level of this course, but the equation has a simple and plausible form: it is the wave equation with a damping term $3H\dot{h}$; the wave velocity is the speed of light = 1.

For superhorizon scales we can ignore the last term, and we get $h = const$ as a solution and another solution where $\dot{h} \equiv dh/dt \propto a^{-3}$ so it also approaches a constant. Thus tensor perturbations remain essentially constant outside the horizon.

For evolution inside the horizon we get oscillatory solutions and then it is better to work with conformal time. The $h(\eta)$ evolution equation is

$$h'' + 2\mathcal{H}h' + k^2 h = 0 \qquad \Leftrightarrow \qquad \mathcal{H}^{-2}h'' + 2\mathcal{H}^{-1}h' + (k/\mathcal{H})^2\, h = 0\,, \tag{273}$$

where $' \equiv d/d\eta$. If we first ignore the middle term, we get solutions of the form $h \propto e^{\pm ik\eta}$, where $-$ represents a wave moving in the $\mathbf{k}$ direction and + in the $-\mathbf{k}$ direction. These are *gravitational waves*. They propagate at the speed of light and they are transverse waves. During one half-period of the wave oscillation, space is stretched in one direction orthogonal to the direction of propagation, and compressed in the other orthogonal direction. During the next half-period the opposite happens. The amplitude of the stretching is given by h, meaning that the maximum stretching is by factor $1 + |h|$ and the maximum compression is by factor $1 - |h|$.

The middle term in (273) represents the damping of gravitational terms due to the expansion of the universe. Write

$$h(\eta) = A(\eta)e^{-ik\eta} \tag{274}$$

and insert this into (273) to get

$$A'' + 2\mathcal{H}A' - 2ik(A' + \mathcal{H}A) = 0\,. \tag{275}$$

For $k \gg \mathcal{H}$, the part $2ik(A' + \mathcal{H}A)$ dominates the left-hand side, and we get

$$A' + \mathcal{H}A \;=\; A' + \frac{a'}{a}A \;=\; \frac{1}{a}(aA)' \;=\; 0 \quad \Rightarrow \quad aA = const \quad \Rightarrow \quad A \propto a^{-1}\,. \tag{276}$$

Thus gravitational waves are damped inside the horizon as a^{-1} independent of the expansion law.

For simple expansion laws one can also solve Eq. (273) exactly, covering also horizon entry/exit. These solutions are Bessel functions.

8.5 Nonlinear growth

When δ grows the evolution becomes nonlinear, requiring a more complicated discussion. One can get further with higher-order perturbation theory, or what is called the Zeldovich approximation, but eventually one has to resort to numerical simulations. We shall not discuss these in this course. The spherically symmetric special case can be done analytically by basing it on solutions for FRW universes with different densities. We do it below for an overdensity in a flat matter-dominated background universe.

8.5.1 Closed Friedmann model

In Cosmology I we derived the expansion law for the closed ($\Omega > 1$) matter-dominated FRW universe. It cannot be given in closed form as $a(t)$, but can be given in terms of an auxiliary variable, the *development angle* ψ, as

$$\begin{aligned} a(\psi) &= a_i \frac{\Omega_i}{2(\Omega_i - 1)}(1 - \cos\psi) = a(\psi)\frac{\Omega(\psi)}{2[\Omega(\psi) - 1]}(1 - \cos\psi) \\ t(\psi) &= H_i^{-1}\frac{\Omega_i}{2(\Omega_i - 1)^{3/2}}(\psi - \sin\psi) = H(\psi)^{-1}\frac{\Omega(\psi)}{2[\Omega(\psi) - 1]^{3/2}}(\psi - \sin\psi)\,, \end{aligned} \qquad (277)$$

where a_i, Ω_i, and H_i are the scale factor, density parameter, and Hubble parameter at some reference time t_i (usually chosen as the present time t_0, but below we will instead choose t_i to be some early time, when Ω is still very close to 1). In the second forms we took advantage of the fact that we can choose t_i to be any time during the development and replaced it with the "current" time. See Fig. 5 for the shape of $a(t)$. This curve is called a *cycloid*. (It is the path made by a point at the rim of a wheel.) From (277) we solve

$$\Omega(\psi) = \frac{2}{1 + \cos\psi}\,. \qquad (278)$$

Calculating $da/dt = da/d\psi \times d\psi/dt$ we find (**exercise**)

$$H(\psi) = 2H_i\frac{(\Omega_i - 1)^{3/2}}{\Omega_i}\frac{\sin\psi}{(1 - \cos\psi)^2}\,. \qquad (279)$$

The matter density is given by

$$\rho(\psi) = \rho_i\left(\frac{a_i}{a(\psi)}\right)^3 = 8\rho_i\frac{(\Omega_i - 1)^3}{\Omega_i^3(1 - \cos\psi)^3}\,. \qquad (280)$$

The scale factor reaches a maximum $a_{\rm ta}$ (and the density a minimum) at the "turnaround" time $t_{\rm ta}$, when $\psi = \pi$, so that

$$a_{\rm ta} = a_i\frac{\Omega_i}{\Omega_i - 1}\,, \qquad t_{\rm ta} = \frac{\pi}{2}H_i^{-1}\frac{\Omega_i}{(\Omega_i - 1)^{3/2}}\,, \qquad \text{and} \qquad \rho(t_{\rm ta}) = \rho_i\frac{(\Omega_i - 1)^3}{\Omega_i^3}\,. \qquad (281)$$

At this point $H = 0$ and then the universe begins to shrink. Since

$$\rho_i = \frac{3\Omega_i H_i^2}{8\pi G} \qquad \text{we have} \qquad \rho(t_{\rm ta}) = \frac{3\pi}{32Gt_{\rm ta}^2}\,. \qquad (282)$$

The universe collapses at $t_{\rm coll} = 2t_{\rm ta}$, when $\psi = 2\pi$ and $a = 0$ again.

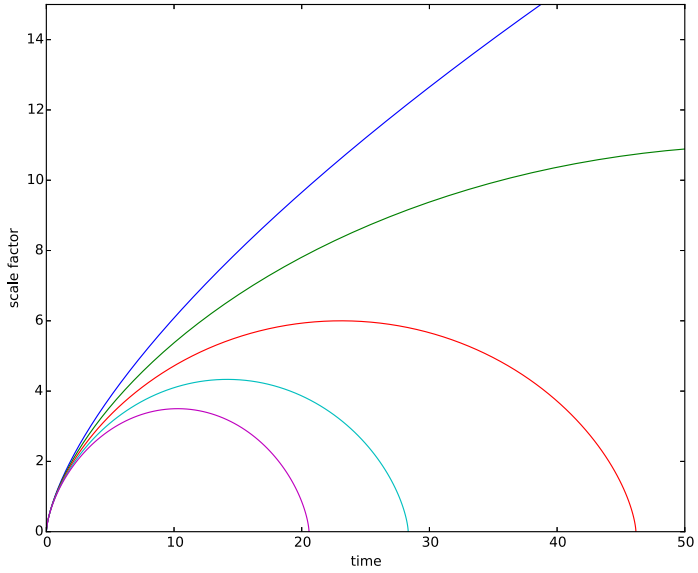


Figure 5: The expansion law for the flat matter-dominated universe (blue) and for closed matter-dominated universes with different initial values $\Omega_i > 1$ for the density parameter. Both axes are linear, the units are arbitrary.

8.5.2 Spherical collapse

The expansion law (277) will hold also for a spherically symmetric overdense region within a flat ($\Omega = 1$) matter-dominated FRW universe. Denote the quantities for this flat background universe by $\bar{a}, \bar{H}, \bar{\rho}$. (Time t is the same for both solutions and $\bar{\Omega} = 1$, so we don't need notations for them.) The background universe has

$$\bar{H}^2 = \frac{8\pi G}{3}\bar{\rho} = \left(\frac{2}{3t}\right)^2 \quad \Rightarrow \quad \bar{\rho} = \frac{1}{6\pi G t^2} \tag{283}$$

Thus we see that at $t_{\rm ta}$, the density of the overdense region is

$$\rho(t_{\rm ta}) = \frac{9\pi^2}{16}\bar{\rho}(t_{\rm ta}) \approx 5.5517\bar{\rho}(t_{\rm ta})\,, \tag{284}$$

i.e., at the turnaround time the density contrast has the value

$$\delta_{\rm ta} = \frac{9\pi^2}{16} - 1 \approx 4.4417\,. \tag{285}$$

Until then the overdense region has been expanding, although slower than the surrounding background universe. At turnaround the overdense region begins to shrink (in terms of proper distance).

The preceding applies both for an overdense region with homogeneous density and for one with a spherically symmetric density profile. In the latter case, we have to apply it separately for each spherical shell, and the density ρ refers, not to the density of the shell, but to the mean density within the shell, as it is the total mass within the shell that is responsible for the gravity affecting the expansion or contraction of the shell. To avoid shell crossing the density profile has to decrease outward, so that outer shells do not collapse before inner shells.[33]

[33] We should also include in our model an underdense region around our overdense region so that their combined mean density equals that of the background universe, so as not to affect the evolution of the surroundings.

In linear perturbation theory, which applies when $\delta \ll 1$, density perturbations in the flat matter-dominated universe grow as

$$\delta^{\rm lin} \propto a \propto t^{2/3} \,. \tag{286}$$

When the density contrast δ becomes large it begins to grow faster. Compare now the linear growth law to the above result for δ at turnaround.

The initial density contrast δ_i is given by $\rho_i = (1+\delta_i)\bar\rho_i$. On the other hand

$$\bar H_i^2 = \frac{8\pi G}{3}\bar\rho_i \qquad \text{and} \qquad H_i^2 = \frac{8\pi G}{3}\Omega_i\rho_i \tag{287}$$

so that

$$1+\delta_i = \Omega_i \frac{H_i^2}{\bar H_i^2} \qquad \text{or at any time} \qquad 1+\delta = \Omega\frac{H^2}{\bar H^2} \,. \tag{288}$$

Thus the density contrast is not simply given by $\Omega - \bar\Omega = \Omega - 1$, since also the Hubble parameters are different for the two solutions. We can sort out the separate contributions from $\Omega_i - 1$ and $(H_i/\bar H_i)^2$ at an early time when $\Omega - 1 \ll 1$ and $\psi \ll 1$, by expanding Ω, H and $\bar H$ from (278), (279) and (283&277) in terms of ψ (**exercise**) to get

$$\Omega_i \approx 1 + \tfrac{1}{4}\psi_i^2 \qquad \text{and} \qquad \frac{H_i^2}{\bar H_i^2} \approx 1 - \tfrac{1}{10}\psi^2 \quad \Rightarrow \quad 1+\delta_i \approx 1 + \tfrac{3}{20}\psi^2 \quad \Rightarrow \quad \delta_i \approx \tfrac{3}{5}(\Omega_i - 1) \,. \tag{289}$$

We can now give the linear prediction for the density contrast at turnaround time[34]:

$$\delta_{\rm ta}^{\rm lin} \;=\; \frac{\bar a_{\rm ta}}{\bar a_i}\delta_i \;=\; \left(\frac{t_{\rm ta}}{t_i}\right)^{2/3}\delta_i \;\approx \left(\frac{3\pi}{4}\right)^{2/3}\frac{\delta_i}{\Omega_i - 1} \;\approx\; \frac{3}{5}\left(\frac{3\pi}{4}\right)^{2/3} \;\approx\; 1.0624 \,, \tag{290}$$

where we approximated

$$t_{\rm ta} \;\approx\; \frac{\pi}{2}\bar H_i^{-1}\frac{1}{(\Omega_i - 1)^{3/2}} \qquad \text{and} \qquad t_i = \tfrac{2}{3}\bar H_1^{-1} \,. \tag{291}$$

Thus we conclude that density perturbations begin to collapse when the linear prediction is $\delta \sim 1$, at which time the true density perturbation is already over 4 times stronger.

The collapse is completed at $t_{\rm coll} = 2t_{\rm ta}$, when the linear prediction gives

$$\delta_{\rm coll}^{\rm lin} \;=\; 2^{2/3}\delta_{\rm ta}^{\rm lin} \;\approx\; 1.6865 \,. \tag{292}$$

The above special case can be extended to the situation where the background universe is a closed or open Friedmann model (i.e., a matter-dominated FRW universe), and to the ΛCDM model, with more complicated math.

8.5.3 Without spherical symmetry

I suppose these idealized cases would lead to a supermassive black hole at the center of symmetry (for perturbations at cosmological scales, for a smaller scale perturbation we might end up with a star). In reality overdensities are never exactly spherically symmetric. The deviation from spherical symmetry increases as the collapse progresses. For an ellipsoidal overdensity the flattest direction collapses first leading first to a "Zeldovich pancake", and the second flattest next leading then to an elongated structure. In the situation where the density refers to a number density of galaxies instead of a smooth continuous density, the galaxies will pass the center point at various distances (instead of colliding at the center as in the perfectly spherically symmetric

[34] Note that Kolb&Turner[6], p. 328, misses the factor 3/5 .

case), after which they will move away from the center and will be decelerated, eventually falling back in and ending up orbiting the center, forming a cluster of galaxies.

For the real universe the different distance scales are in a different stage of the collapse. The largest distance scales are still “falling in”, leading to flattened structures at the largest scales and elongated structures, “filaments”, at somewhat smaller scales. These structures surround rounder underdense regions, “voids”. Smaller scales have already collapsed into galaxy clusters.

8.6 Perturbations during inflation

So far we have developed perturbation theory describing the substance filling the universe in fluid terms, i.e., giving the perturbations in terms of $\delta\rho$ and δp. During inflation the universe is dominated by a scalar field, the inflaton φ, so it is better to give the perturbation directly as a perturbation in the inflaton field,

$$\varphi(t,\mathbf{x}) = \bar{\varphi}(t) + \delta\varphi(t,\mathbf{x})\,. \tag{293}$$

8.6.1 Evolution of inflaton perturbations

In Minkowski space the field equation for a scalar field is

$$\ddot{\varphi} - \nabla^2\varphi + V'(\varphi) = 0\,. \tag{294}$$

In the flat Friedmann-Robertson-Walker universe (the background universe) the field equation is

$$\ddot{\varphi} + 3H\dot{\varphi} - a^{-2}\nabla^2\varphi + V'(\varphi) = 0\,. \tag{295}$$

(Here $\nabla = \nabla_{\mathbf{x}}$, i.e., with respect to the comoving coordinates $\mathbf{x}$, and therefore the factor $1/a$ appears in front of it.)

We ignore for the moment the perturbation in the spacetime metric and just insert (293) into Eq. (295),

$$(\bar{\varphi}+\delta\varphi)^{\cdot\cdot} + 3H(\bar{\varphi}+\delta\varphi)^{\cdot} - a^{-2}\nabla^2(\bar{\varphi}+\delta\varphi) + V'(\bar{\varphi}+\delta\varphi) = 0\,. \tag{296}$$

Here $V'(\bar{\varphi}+\delta\varphi) = V'(\bar{\varphi}) + V''(\bar{\varphi})\delta\varphi$ and $\bar{\varphi}(t)$ is the homogeneous background solution from our earlier discussion of inflation. Thus $\nabla^2\bar{\varphi} = 0$, and $\bar{\varphi}$ satisfies the background equation

$$\ddot{\bar{\varphi}} + 3H\dot{\bar{\varphi}} + V'(\bar{\varphi}) = 0\,. \tag{297}$$

Subtracting the background equation from the full equation (296) we get the perturbation equation

$$\delta\ddot{\varphi} + 3H\delta\dot{\varphi} - a^{-2}\nabla^2\delta\varphi + V''(\bar{\varphi})\delta\varphi = 0 \tag{298}$$

In Fourier space we have

$$\delta\ddot{\varphi}_{\mathbf{k}} + 3H\delta\dot{\varphi}_{\mathbf{k}} + \left[\left(\frac{k}{a}\right)^2 + m^2(\bar{\varphi})\right]\delta\varphi_{\mathbf{k}} = 0\,, \tag{299}$$

or

$$H^{-2}\delta\ddot{\varphi}_{\mathbf{k}} + 3H^{-1}\delta\dot{\varphi}_{\mathbf{k}} + \left[\left(\frac{k}{aH}\right)^2 + \frac{m^2}{H^2}\right]\delta\varphi_{\mathbf{k}} = 0\,, \tag{300}$$

where

$$m^2(\bar{\varphi}) \equiv V''(\bar{\varphi})\,. \tag{301}$$

During inflation, H and m^2 change slowly. Thus we make now an approximation where we treat them as constants. If the slow-roll approximation is valid, $m^2 \ll H^2$, since

$$\frac{m^2}{H^2} = 3M_{\rm Pl}^2\frac{V''}{V} = 3\eta \ll 1\,. \tag{302}$$

Thus we can ignore the m^2/H^2 in Eq. (300)[35]. The general solution becomes then

$$\delta\varphi_{\mathbf{k}}(t) = A_{\mathbf{k}}w_k(t) + B_{\mathbf{k}}w_k^*(t)\,, \tag{305}$$

[35]The general solution to (299), when H and m^2 are constants, is

$$\delta\varphi_{\mathbf{k}}(t) = a^{-3/2}\left[A_{\mathbf{k}}J_{-\nu}\left(\frac{k}{aH}\right) + B_{\mathbf{k}}J_{\nu}\left(\frac{k}{aH}\right)\right]\,, \tag{303}$$

where

$$w_k(t) = \left(i + \frac{k}{aH}\right) \exp\left(\frac{ik}{aH}\right) . \tag{306}$$

(**Exercise:** Show that this is a solution of (299) when $H = const$ and $m^2 = 0$.) The time dependence of (305) is in

$$a = a(t) \propto e^{Ht} . \tag{307}$$

Well before horizon exit, $k \gg aH$, the argument of the exponent is large. As $a(t)$ increases the solution oscillates rapidly and its amplitude is damped. *After horizon exit*, $k \ll aH$, the solution stops oscillating and *approaches the constant value* $i(A_{\mathbf{k}} - B_{\mathbf{k}})$.

We have cheated by ignoring the metric perturbation. We should use GR and write the curved-spacetime field equation using the perturbed metric. Perturbations in a scalar field couple only to scalar perturbations, so we need to consider scalar perturbations only. For example, in the conformal-Newtonian gauge the correct perturbation equation is

$$\delta\ddot{\varphi}^N_{\mathbf{k}} + 3H\delta\dot{\varphi}^N_{\mathbf{k}} + \left[\left(\frac{k}{a}\right)^2 + V''(\bar{\varphi})\right] \delta\varphi^N_{\mathbf{k}} \;=\; -2\Phi_{\mathbf{k}} V(\bar{\varphi}) + \left(\dot{\Phi}_{\mathbf{k}} + 3\dot{\Psi}_{\mathbf{k}}\right) \dot{\bar{\varphi}} . \tag{308}$$

That is, there are additional terms which are first order in the metric and zeroth order (background) in the scalar field φ.

Fortunately, it is possible to choose the gauge so that the terms with the metric perturbations are negligible *during inflation*[36], and the previous calculation applies in such a gauge. The comoving gauge is not such a gauge, so a gauge transformation is required to obtain the comoving gauge curvature perturbation $\mathcal{R}$. Gauge transformations are beyond the scope of these lectures, but the result is

$$\mathcal{R} = -H \frac{\delta\varphi}{\dot{\bar{\varphi}}} . \tag{309}$$

Thus it is clear what we want from inflation. We want to find the inflaton perturbations $\delta\varphi_{\mathbf{k}}$ some time after horizon exit. We can use the constant value the solution (305) approaches after horizon exit. Then Eq. (309) gives us $\mathcal{R}_k$, which remains constant while the scale k is outside the horizon, and is indeed the primordial $\mathcal{R}_{\mathbf{k}}$ discussed in the previous section. And then we can use the results of Sec. 8.4 to get $\delta_{\mathbf{k}}$.

We are still missing the initial conditions for the solution (305). These are determined by quantum fluctuations, which we shall discuss in Sec. 8.6.3. Quantum fluctuations produce the initial conditions in a random manner, so that we can predict only their statistical properties. It turns out that the quantum fluctuations are a *Gaussian process*, a term which specifies certain statistical properties, which we shall discuss next before returning to the application to inflaton fluctuations.

8.6.2 Statistical properties of Gaussian perturbations

The statistical (Gaussian) nature of the inflaton perturbations $\delta\varphi(\mathbf{x})$ are inherited later by other perturbations, which depend linearly on them. Let us therefore discuss a generic Gaussian

where J_ν is the Bessel function of order ν and

$$\nu = \sqrt{\frac{9}{4} - \frac{m^2}{H^2}} . \tag{304}$$

With $m^2 = 0$, $\nu = \frac{3}{2}$. Bessel functions of half-integer order are spherical Bessel functions which can be expressed in terms of trigonometric functions, or $e^{\pm ikx}$.

[36] One such gauge is the *spatially flat gauge* Q. For scalar perturbations it is possible to choose the time coordinate so that the time slices have Euclidean geometry. This leads to the spatially flat gauge. (There are still perturbations in the spacetime curvature; they show up when one considers the time direction).

perturbation

$$g(\mathbf{x}) = \sum_{\mathbf{k}} g_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}}\,, \tag{310}$$

where the set of Fourier coefficients $\{g_{\mathbf{k}}\}$ is a result of a *statistically homogeneous and isotropic Gaussian random process*. We assume $g(\mathbf{x})$ is real, so that $g_{-\mathbf{k}} = g^*_{\mathbf{k}}$. We write $g_{\mathbf{k}}$ in terms of its real and imaginary part,

$$g_{\mathbf{k}} = \alpha_{\mathbf{k}} + i\beta_{\mathbf{k}}\,. \tag{311}$$

For Fourier analysis of statistically homogeneous and isotropic random perturbations, see sections (8.1.1, 8.1.3, 8.1.4), where the probability distribution was treated as unknown. The new ingredient (in addition to the assumption that the perturbations are small, allowing the use of first-order perturbation theory, which we introduced in Sec. 8.2), is that the probability distribution is known to be Gaussian. This means that

$$\begin{aligned}\mathrm{Prob}(g_{\mathbf{k}}) &= \frac{1}{2\pi s^2_{\mathbf{k}}} \exp\left(-\frac{1}{2}\frac{|g_{\mathbf{k}}|^2}{s^2_{\mathbf{k}}}\right) \\ &= \frac{1}{\sqrt{2\pi}s_{\mathbf{k}}} \exp\left(-\frac{1}{2}\frac{\alpha^2_{\mathbf{k}}}{s^2_{\mathbf{k}}}\right) \times \frac{1}{\sqrt{2\pi}s_{\mathbf{k}}} \exp\left(-\frac{1}{2}\frac{\beta^2_{\mathbf{k}}}{s^2_{\mathbf{k}}}\right)\,,\end{aligned} \tag{312}$$

i.e., the real and imaginary parts are independent Gaussian random variables[37] with equal variance $s^2_{\mathbf{k}}$.

The *expectation value* of a quantity which depends on $g_{\mathbf{k}}$ as $f(g_{\mathbf{k}})$ is given by

$$\langle f(g_{\mathbf{k}})\rangle \equiv \int f(g_{\mathbf{k}})\mathrm{Prob}(g_{\mathbf{k}}) d\alpha_{\mathbf{k}} d\beta_{\mathbf{k}}\,, \tag{313}$$

where the integral is over the complex plane, i.e.,

$$\int_{-\infty}^{\infty} d\alpha_{\mathbf{k}} \int_{-\infty}^{\infty} d\beta_{\mathbf{k}}\,.$$

We immediately get (**exercise**) the *mean*

$$\langle g_{\mathbf{k}}\rangle = 0 \tag{314}$$

and *variance*

$$\langle |g_{\mathbf{k}}|^2\rangle = 2s^2_{\mathbf{k}} \tag{315}$$

of $g_{\mathbf{k}}$.

The distribution has one free parameter, the real positive number $s_{\mathbf{k}}$ which gives the width (determines the variance) of the distribution. From statistical isotropy and homogeneity follows that $s_{\mathbf{k}} = s(k)$ and

$$\langle g^*_{\mathbf{k}} g_{\mathbf{k}'}\rangle = 0 \qquad \text{for} \quad \mathbf{k} \neq \mathbf{k}'\,. \tag{316}$$

We can combine Eqs. (315) and (316) into a single equation,

$$\langle g^*_{\mathbf{k}} g_{\mathbf{k}'}\rangle \;=\; 2\delta_{\mathbf{k}\mathbf{k}'} s^2_{\mathbf{k}} \;=\; \delta_{\mathbf{k}\mathbf{k}'}\langle |g_{\mathbf{k}}|^2\rangle \;=\; \frac{\delta_{\mathbf{k}\mathbf{k}'}}{V} P_g(k) \;=\; \frac{2\pi^2\delta_{\mathbf{k}\mathbf{k}'}}{Vk^3}\mathcal{P}_g(k)\,, \tag{317}$$

where

$$\mathcal{P}_g(k) \;\equiv\; \left(\frac{L}{2\pi}\right)^3 4\pi k^3 \langle |g_{\mathbf{k}}|^2\rangle \;=\; \frac{V}{2\pi^2} k^3 \langle |g_{\mathbf{k}}|^2\rangle\,, \tag{318}$$

[37] $g_{\mathbf{k}}$ is a complex Gaussian random variable and $\alpha_{\mathbf{k}}$ and $\beta_{\mathbf{k}}$ are real Gaussian random variables.

which gives the dependence of the variance of $g_{\mathbf{k}}$ on the wave number k, is the *power spectrum* of g.

Going back to coordinate space, we find

$$\langle g(\mathbf{x})\rangle = \left\langle \sum_{\mathbf{k}} g_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} \right\rangle = \sum_{\mathbf{k}} \langle g_{\mathbf{k}}\rangle e^{i\mathbf{k}\cdot\mathbf{x}} = 0 \tag{319}$$

The square of the perturbation can be written as

$$g(\mathbf{x})^2 = \sum_{\mathbf{k}} g^*_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{x}} \sum_{\mathbf{k}'} g_{\mathbf{k}'} e^{i\mathbf{k}'\cdot\mathbf{x}} \tag{320}$$

since $g(\mathbf{x})$ is real. The typical amplitude of the perturbation is described by the variance, the expectation value of this square,

$$\begin{aligned} \langle g(\mathbf{x})^2\rangle &= \sum_{\mathbf{k}\mathbf{k}'} \langle g^*_{\mathbf{k}} g_{\mathbf{k}'}\rangle e^{i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{x}} = \sum_{\mathbf{k}} \langle |g_{\mathbf{k}}|^2\rangle = 2\sum_{\mathbf{k}} s^2_{\mathbf{k}}\,. = \left(\frac{2\pi}{L}\right)^3 \sum_{\mathbf{k}} \frac{1}{4\pi k^3}\mathcal{P}_g(k) \\ &\to \frac{1}{4\pi}\int \frac{d^3k}{k^3}\mathcal{P}_g(k) = \int_0^\infty \frac{dk}{k}\mathcal{P}_g(k) = \int_{-\infty}^\infty \mathcal{P}_g(k) d\ln k\,. \end{aligned} \tag{321}$$

Note that there is no $\mathbf{x}$-dependence in the result, since this is an expectation value. $g(\mathbf{x})^2$ of course varies from place to place, but its expectation value from the random process is the same everywhere—the perturbed universe is statistically homogeneous.

Thus the power spectrum of g gives the contribution of a logarithmic scale interval to the variance of $g(\mathbf{x})$. *For Gaussian perturbations, the power spectrum gives a complete statistical description.* All statistical quantities can be calculated from it.

In practice the integration is not extended all the way from $k = 0$ to $k = \infty$. Rather, there is usually some largest and smallest relevant scale, which introduce natural cutoffs at both ends of the integral. The largest relevant scale could be the size of the observable universe: The perturbation $g(\mathbf{x})$ represents a deviation from the background quantity, but the best estimate we have for the background may be the average taken over the observable universe. Then perturbations at larger scales contribute to our estimate of the background value instead of contributing to the perturbation away from it. The smallest relevant scale could be the resolution of the observational survey considered. For example, density perturbations are observed as perturbations in the number density of galaxies; such a number density can only be meaningfully defined at scales larger than the typical separation between galaxies.

It can be shown (under weak assumptions about the power spectrum), that statistically homogeneous and isotropic Gaussian perturbations are ergodic, so that we do not need to make a separate assumption of ergodicity.[38]

An alternative definition for the power spectrum is

$$P_g(k) \equiv V\langle |g_{\mathbf{k}}|^2\rangle \tag{322}$$

While this definition is simpler, the result for the variance of $g(\mathbf{x})$ in terms of it and thus the interpretation is more complicated. Because of the common use of this latter definition, we shall make reference to both power spectra, and distinguish them by the different typeface. They are related by

$$P_g(k) = \frac{2\pi^2}{k^3}\mathcal{P}_g(k)\,. \tag{323}$$

[38] Liddle & Lyth [2], in Sec. 4.3.3, make this claim but do not provide a proof.

8.6.3 Generation of primordial perturbations from inflation

Subhorizon scales during inflation are microscopic[39] and therefore quantum effects are important. Thus we should study the inflaton field using quantum field theory.

This goes beyond the level of this course, so we have relegated the discussion into an appendix. The basic idea is that for scales that are inside horizon there are quantum fluctuations, called *vacuum fluctuations*, in the inflaton field. For a homogeneous inflaton field, the Fourier amplitudes $\delta\varphi_{\mathbf{k}}$ of its perturbations would be identically zero, but analogous to a quantum harmonic oscillator, it is not possible for them to stay there, but instead they fluctuate around this value.

We saw in Sec. 8.6.1 that the classical solutions to the evolution of $\delta\varphi_{\mathbf{k}}$ reach a constant value after horizon exit (in the approximation $H = const$ during horizon exit). The quantum treatment gives that at this stage we can neglect further quantum fluctuations and treat $\delta\varphi_{\mathbf{k}}$ classically—the fluctuations "freeze".

The final result is that well after horizon exit, $k \ll \mathcal{H}$, the Fourier amplitudes $\delta\varphi_{\mathbf{k}}$ have acquired a power spectrum

$$\mathcal{P}_\varphi(k) \equiv V\frac{k^3}{2\pi^2}\langle|\delta\varphi_{\mathbf{k}}|^2\rangle = \left(\frac{H}{2\pi}\right)^2 . \tag{324}$$

After this we can ignore further quantum effects and treat the later evolution of the inflaton field, both the background and the perturbation, classically. The effect of the vacuum fluctuations was to produce "out of nothing" the perturbations $\delta\varphi_{\mathbf{k}}$. We can't predict their individual values; their production from quantum fluctuations is a random process. We can only calculate their statistical properties. Closer investigation reveals that this is a Gaussian random process. All $\delta\varphi_{\mathbf{k}}$ acquire their values as independent random variables (except for the reality condition $\delta\varphi_{-\mathbf{k}} = \delta\varphi^*_{\mathbf{k}}$) with a Gaussian probability distribution. Thus all statistical information is contained in the power spectrum $\mathcal{P}_\varphi(k)$.

The result (324) was obtained treating H as a constant. However, over long time scales, H does change. The main purpose of the preceding discussion was to follow the inflaton perturbations through the horizon exit. After the perturbation is well outside the horizon, we switch to other variables, namely the curvature perturbation $\mathcal{R}_{\mathbf{k}}$, which, unlike $\delta\varphi_{\mathbf{k}}$, remains constant outside the horizon, even though H changes. Therefore we have to use for each scale k a value of H which is representative for the evolution of that particular scale through the horizon. That is, we choose the value of H at horizon exit,[40] so that $aH = k$. Thus we write our power spectrum result as

$$\boxed{\mathcal{P}_\varphi(k) = V\frac{k^3}{2\pi^2}\langle|\delta\varphi_{\mathbf{k}}|^2\rangle = \left(\frac{H}{2\pi}\right)^2_{aH=k},} \tag{325}$$

to signify that the value of H for each k is to be taken at horizon exit of that particular scale. Equation (325) is our main result from inflaton fluctuations.

8.6.4 Transfer functions

Since the inflaton fluctuations are assumed to be the origin of structure, all later perturbations are related to the inflaton perturbations $\delta\varphi_{\mathbf{k}}$. As long as all inhomogeneities are small ("per-

[39] We later give an upper limit to the inflation energy scale, i.e., $V(\varphi)$ at the time cosmological scales exited the horizon, $V^{1/4} < 1.9 \times 10^{16}$ GeV. From $H^2 = V(\varphi)/3M_{\rm Pl}^2$ we have $H < 9 \times 10^{13}$ GeV or for the Hubble length $H^{-1} > 2.3 \times 10^{-30}$ m. This is a lower limit to the horizon size, but it is not expected to be very many orders of magnitude larger.

[40] One can do a more precise calculation, where one takes into account the evolution of $H(t)$. The result is that one gets a correction to the amplitude of $\mathcal{P}_\mathcal{R}(k)$, which is first order in slow-roll parameters, and a correction to its spectral index n which is second order in the slow-roll parameters.

turbations"), the relationship is linear. We can express these linear relationships as *transfer functions* $T(t,k)$, e.g.,

$$g_{\mathbf{k}}(t) = T_{g\varphi}(t,k)\delta\varphi_{\mathbf{k}}(t_k)\,. \tag{326}$$

The linearity implies several things:

1. The Fourier coefficient $g_{\mathbf{k}}$ depends only on the Fourier coefficient of $\delta\varphi$ corresponding to the same wave vector $\mathbf{k}$, not on any other $\mathbf{k}'$.

2. The relationship is linear, so that if $\delta\varphi_{\mathbf{k}}$ were, e.g., twice as big, then so would $g_{\mathbf{k}}$ be.

3. The perturbations of g inherit the Gaussian statistics of $\delta\varphi$.

We could also define transfer functions relating perturbations at any two different times, t and t', and call them $T(t,t',k)$, but here we are referring to the inflaton perturbations at the time of horizon exit, t_k, which is different for different k. Actually, by $\delta\varphi_{\mathbf{k}}(t_k)$ we mean the constant value the perturbation approaches after horizon exit in the $H = const = H_k$ approximation.

That the transfer function depends only on the magnitude k results from the fact that physical laws are isotropic. The transfer function of Eq. (326) will then relate the power spectra of $\{g_{\mathbf{k}}(t)\}$ and $\{\delta\varphi_{\mathbf{k}}(t_k)\}$ as

$$\mathcal{P}_g(t,k) = T_{g\varphi}(t,k)^2\mathcal{P}_\varphi(k)\,. \tag{327}$$

The transfer functions thus incorporate all the physics that determines how structure evolves.

For the largest scales, $k^{-1} \gg 10h^{-1}\,\mathrm{Mpc}$, the perturbations are still small today, and one needs not go beyond the transfer function. For smaller scales, corresponding to galaxies and galaxy clusters, the inhomogeneities have become large at late times, and the physics of structure growth has become nonlinear. This nonlinear evolution is typically studied using large numerical simulations. Fortunately, the relevant scales are small enough that Newtonian physics is usually sufficient.

We are now in position to put together all the results we obtained. From Eq. (309)

$$\mathcal{R}_{\mathbf{k}} = -H\frac{\delta\varphi_{\mathbf{k}}}{\dot{\bar{\varphi}}}\,, \tag{328}$$

so that

$$T_{\mathcal{R}\varphi}(k) = -\frac{H_k}{\dot{\bar{\varphi}}(t_k)} \tag{329}$$

and

$$\mathcal{P}_{\mathcal{R}}(k) = \left(\frac{H}{\dot{\bar{\varphi}}}\right)^2\mathcal{P}_\varphi(k) = \left[\left(\frac{H}{\dot{\bar{\varphi}}}\right)\left(\frac{H}{2\pi}\right)\right]^2_{\mathcal{H}=k}\,, \tag{330}$$

where we used the result (325).

This primordial spectrum is the starting point for calculating structure formation (discussed already) and the CMB anisotropy (Chapter 9). Thus CMB and large-scale structure observations can be used to constrain $\mathcal{P}_{\mathcal{R}}$ together with other cosmological parameters.

8.6.5 Generation of primordial gravitational waves

The quantum fluctuations at subhorizon scales during inflation apply also to the spacetime itself. We do not yet have a complete theory of quantum gravity, so we do not know how spacetime behaves in the Planck era. At lower energy scales the spacetime fluctuations are smaller and for small perturbations around a FRW universe we can use the linearized equations for metric perturbations, for which quantization is straightforward. In fact, the proper treatment of the generation of inflaton perturbations, where we include the scalar metric perturbations in the

inflaton perturbation equation (see Eq. 308), contains also the quantum treatment of scalar metric perturbations.

Likewise, we have quantum fluctuations of tensor metric perturbations during inflation. These do not couple to density perturbations, but they become classical gravitational waves after horizon exit. These *primordial gravitational waves* have an effect on CMB anisotropy and polarization.

In the quantum treatment, $(M_{\rm Pl}/\sqrt{2})h$ fluctuates like a scalar field, so that in inflation the gravitational wave amplitudes h acquire a spectrum

$$\mathcal{P}_h(k) \;\equiv\; 4\frac{V}{2\pi^2}k^3\langle|h_{\mathbf{k}}|^2\rangle \;=\; 4\frac{2}{M_{\rm Pl}^2}\left(\frac{H}{2\pi}\right)^2_{\mathcal{H}=k} \;=\; \frac{8}{M_{\rm Pl}^2}\left(\frac{H}{2\pi}\right)^2_{\mathcal{H}=k} \tag{331}$$

(the factor 4 in this customary definition is related to the way h appears in several places in the metric and to there being two modes for each $\mathbf{k}$).

The *tensor-to-scalar ratio* is the ratio of the two primordial spectra (331) and (330),

$$r \;\equiv\; \frac{\mathcal{P}_h(k)}{\mathcal{P}_{\mathcal{R}}(k)} \;=\; \frac{8}{M_{\rm Pl}^2}\left(\frac{\dot{\varphi}}{H}\right)^2_{\mathcal{H}=k}\,. \tag{332}$$

8.7 The primordial spectrum

8.7.1 Primordial spectrum from slow-roll inflation

The final result of the previous section is thus that inflation generates primordial scalar perturbations $\mathcal{R}_{\mathbf{k}}$ with the power spectrum

$$\mathcal{P}_{\mathcal{R}}(k) \;=\; \left[\left(\frac{H}{\dot\varphi}\right)\left(\frac{H}{2\pi}\right)\right]^2_{\mathcal{H}=ak} \;=\; \frac{1}{4\pi^2}\left(\frac{H^2}{\dot\varphi}\right)^2_{t=t_k}. \tag{333}$$

and primordial tensor perturbations with the power spectrum

$$\mathcal{P}_h(k) \;=\; \frac{8}{M_{\rm Pl}^2}\left(\frac{H}{2\pi}\right)^2_{t=t_k}. \tag{334}$$

In this section φ and $\dot\varphi$ refer to the background values.

Applying the slow-roll equations

$$H^2 = \frac{V}{3M_{\rm Pl}^2} \quad\text{and}\quad 3H\dot\varphi = -V' \quad\Rightarrow\quad \frac{\dot\varphi}{H} = -M_{\rm Pl}^2\frac{V'}{V}$$

these become

$$\begin{aligned}\mathcal{P}_{\mathcal{R}}(k) &= \frac{1}{12\pi^2}\frac{1}{M_{\rm Pl}^6}\frac{V^3}{V'^2} \;=\; \frac{1}{24\pi^2}\frac{1}{M_{\rm Pl}^4}\frac{V}{\varepsilon}\\ \mathcal{P}_h(k) &= \frac{2}{3\pi^2}\frac{V}{M_{\rm Pl}^4},\end{aligned} \tag{335}$$

where ε is the slow-roll parameter. The tensor-to-scalar ratio is thus

$$r \;\equiv\; \frac{\mathcal{P}_h(k)}{\mathcal{P}_{\mathcal{R}}(k)} \;=\; 16\varepsilon\,. \tag{336}$$

According to present observational CMB and large-scale structure data, the amplitude of the primordial power spectrum is about

$$\mathcal{P}_{\mathcal{R}}(k)^{1/2} \approx 5\times 10^{-5} \tag{337}$$

at cosmological scales. This gives a constraint on inflation

$$\left(\frac{V}{\varepsilon}\right)^{1/4} \;\approx\; 24^{1/4}\sqrt{\pi}\sqrt{5\times10^{-5}}M_{\rm Pl} \;\approx\; 0.028M_{\rm Pl} \;=\; 6.8\times10^{16}\,{\rm GeV}\,. \tag{338}$$

The best chance of detecting primordial gravitational waves is based on their effect on CMB. They have not been observed so far and the present upper limit is about

$$r<0.1 \quad\Rightarrow\quad \mathcal{P}_h(k)^{1/2} < 1.5\times10^{-5} \quad\text{and}\quad \varepsilon<0.006\,. \tag{339}$$

This implies an upper limit to the inflation energy scale

$$V^{1/4} \;\approx\; \varepsilon^{1/4}0.028M_{\rm Pl} \;<\; 0.008M_{\rm Pl} \;=\; 1.9\times10^{16}\,{\rm GeV}\,. \tag{340}$$

Since during inflation, V and V' change slowly while a wide range of scales k exit the horizon, $\mathcal{P}_{\mathcal{R}}(k)$ and $\mathcal{P}_h(k)$ should be slowly varying functions of k. We define the *spectral indices* n_s and n_t of the primordial spectra as

$$\begin{aligned} n_s(k)-1 &\equiv \frac{d\ln\mathcal{P}_{\mathcal{R}}}{d\ln k}\\ n_t(k) &\equiv \frac{d\ln\mathcal{P}_h}{d\ln k}\,.\end{aligned} \tag{341}$$

(The -1 is in the definition of n_s for historical reasons, to match with the definition in terms of density perturbations, see Sec. 8.7.2.) If the spectral index is independent of k, we say that the spectrum is *scale free*. In this case the primordial spectra have the *power-law* form

$$\mathcal{P}_\mathcal{R}(k) = A_s^2 \left(\frac{k}{k_p}\right)^{n_s-1} \qquad \text{and} \qquad \mathcal{P}_h(k) = A_t^2 \left(\frac{k}{k_p}\right)^{n_t} , \tag{342}$$

where k_p is some chosen reference scale, "pivot scale", and A_s and A_t are the amplitudes at this pivot scale.

If the power spectrum is constant,

$$\mathcal{P} = \text{const.} , \tag{343}$$

corresponding to $n_s = 1$ and $n_t = 0$, we say that the spectrum is *scale invariant*. A scale-invariant scalar spectrum is also called the *Harrison–Zeldovich* spectrum.

If $n_s \neq 1$ or $n_t \neq 0$, the spectrum is called *tilted*. A tilted spectrum is called *red*, if $n_s < 1$ (more structure at large scales), and *blue* if $n_s > 1$ (more structure at small scales).

Using Eqs. (335) and (341) we can calculate the spectral index for slow-roll inflation.

Since $\mathcal{P}(k)$ is evaluated from Eqs. (333) and (334) or (335) when $k = aH$,

$$\frac{d\ln k}{dt} = \frac{d\ln(aH)}{dt} = \frac{\dot a}{a} + \frac{\dot H}{H} = (1-\varepsilon)H ,$$

where we used $\dot H = -\varepsilon H^2$ (in the slow-roll approximation) in the last step. Thus

$$\frac{d}{d\ln k} = \frac{1}{1-\varepsilon}\frac{1}{H}\frac{d}{dt} = \frac{1}{1-\varepsilon}\frac{\dot\varphi}{H}\frac{d}{d\varphi} = -\frac{M_\mathrm{Pl}^2}{1-\varepsilon}\frac{V'}{V}\frac{d}{d\varphi} \approx -M_\mathrm{Pl}^2\frac{V'}{V}\frac{d}{d\varphi} . \tag{344}$$

Let us first calculate the scale dependence of the slow-roll parameters:

$$\frac{d\varepsilon}{d\ln k} = -M_\mathrm{Pl}^2\frac{V'}{V}\frac{d}{d\varphi}\left[\frac{M_\mathrm{Pl}^2}{2}\left(\frac{V'}{V}\right)^2\right] = M_\mathrm{Pl}^4\left[\left(\frac{V'}{V}\right)^4 - \left(\frac{V'}{V}\right)^2\frac{V''}{V}\right] = 4\varepsilon^2 - 2\varepsilon\eta \tag{345}$$

and, in a similar manner (**exercise**),

$$\frac{d\eta}{d\ln k} = \ldots = 2\varepsilon\eta - \xi , \tag{346}$$

where we have defined a third slow-roll parameter

$$\xi \equiv M_\mathrm{Pl}^4\frac{V'}{V^2}V''' . \tag{347}$$

The parameter ξ is typically second-order small in the sense that $\sqrt{|\xi|}$ is of the same order of magnitude as ε and η. (Therefore it is sometimes written as ξ^2, although nothing forces it to be positive.)

We are now ready to calculate the spectral indices:

$$\begin{aligned} n_s - 1 &= \frac{1}{\mathcal{P}_\mathcal{R}}\frac{d\mathcal{P}_\mathcal{R}}{d\ln k} = \frac{\varepsilon}{V}\frac{d}{d\ln k}\left(\frac{V}{\varepsilon}\right) = \frac{1}{V}\frac{dV}{d\ln k} - \frac{1}{\varepsilon}\frac{d\varepsilon}{d\ln k} \\ &= -M_\mathrm{Pl}^2\frac{V'}{V}\cdot\frac{1}{V}\frac{dV}{d\varphi} - 4\varepsilon + 2\eta = -6\varepsilon + 2\eta \\ n_t &= \frac{1}{\mathcal{P}_h}\frac{d\mathcal{P}_h}{d\ln k} = -M_\mathrm{Pl}^2\frac{V'}{V}\frac{1}{V}\frac{dV}{d\varphi} = -M_\mathrm{Pl}^2\left(\frac{V'}{V}\right)^2 = -2\varepsilon . \end{aligned} \tag{348}$$

Since $\varepsilon > 0$, the tensor spectrum is necessarily red. (This follows already from (334), since H is decreasing, or from (335) since V is decreasing.) Slow-roll requires $\varepsilon \ll 1$ and $|\eta| \ll 1$, so *both spectra are close to scale invariant*. For scalar perturbations this is verified by observation. Based on CMB anisotropy data from the Planck satellite, the Planck Collaboration [4] finds

$$n_s = 0.965 \pm 0.004\,. \tag{349}$$

If one were able to measure all three values n_s, r, and n_t from observations, one could solve from them the slow-roll parameters ε and η and moreover, check the *consistency condition*

$$n_t = -\frac{r}{8} \tag{350}$$

for single-field slow-roll inflation. This consistency condition is the only truly quantitative prediction of the inflation scenario (as opposed to some specific inflation model) – all the other predictions (Ω_k very small, n_s close to 1 and n_t close to 0, primordial perturbations Gaussian) are of qualitative nature, not a specific number not equal to 0 or 1.

Unfortunately, the existing upper limit to r already means that it will be difficult to ever determine the spectral index n_t with sufficient accuracy to distinguish between $n_t = -r/8$ and $n_t = 0$. The most sensitive probe to primordial gravitational waves is provided by polarization of CMB on which they will imprint a characteristic pattern (discussed briefly in the next chapter). The theoretical limit to detection is $r \sim 10^{-4}$ and there are proposals[41] for future CMB satellite missions that could reach $r \sim 10^{-3}$. If r is significantly larger than these detection limits, after detection one could still measure n_t accurately enough to distinguish, say, $n_t \approx -1$, $n_t \approx 0$ (which includes the case $n_t = -r/8$), and $n_t \approx 1$ from each other. There have been other proposals (other than inflation) for very-early-universe physics, which predict primordial tensor perturbations that deviate from scale invariance this much or more.

Detection of primordial gravitational waves, i.e., measurement of r, would be enough to determine ε and η and thus the inflation energy scale from Eq. (338).

One can also calculate the scale-dependence of the spectral index (**exercise**):

$$\frac{dn_s}{d\ln k} = 16\varepsilon\eta - 24\varepsilon^2 - 2\xi\,. \tag{351}$$

It is second order in slow-roll parameters, so it's expected to be even smaller than the deviation from scale invariance, $n_s - 1$. Planck Collaboration finds it consistent with zero to accuracy $\mathcal{O}(10^{-2})$, as expected.

Cosmologically observable scales have a range of about $\Delta\ln k \sim 10$. Planck measured the CMB anisotropy over a range $\Delta\ln k \sim 6$ (missing the shortest scales, where the CMB is expected to have negligible anisotropy). Some inflation models have $|n_s - 1|$, r, and $|dn_s/d\ln k|$ larger than the Planck results, while others do not. These observations already ruled out many inflation models.

Example: Consider the simple inflation model

$$V(\varphi) = \frac{1}{2}m^2\varphi^2\,.$$

In Chapter 7 we already calculated the slow-roll parameters for this model:

$$\varepsilon = \eta = 2\frac{M_{\rm Pl}^2}{\varphi^2}$$

[41] See, e.g., `http://www.core-mission.org/`

and we immediately see that $\xi = 0$. Thus

$$\begin{aligned} n_s &= 1 - 6\varepsilon + 2\eta \;=\; 1 - 8\left(\frac{M_{\rm Pl}}{\varphi}\right)^2 \\ \frac{dn_s}{d\ln k} &= 16\varepsilon\eta - 24\varepsilon^2 - 2\xi \;=\; -32\left(\frac{M_{\rm Pl}}{\varphi}\right)^4 \\ r &= 16\varepsilon \;=\; 32\left(\frac{M_{\rm Pl}}{\varphi}\right)^2 \\ n_t &= -2\varepsilon \;=\; -4\left(\frac{M_{\rm Pl}}{\varphi}\right)^2 \end{aligned}$$

To get the numbers out, we need the values of φ when the relevant cosmological scales exited the horizon. The number of inflation e-foldings after that should be about $N \sim 50$. We have

$$N(\varphi) \;=\; \frac{1}{M_{\rm Pl}^2}\int_{\varphi_{\rm end}}^{\varphi}\frac{V}{V'}d\varphi \;=\; \frac{1}{M_{\rm Pl}^2}\int\frac{\varphi}{2} \;=\; \frac{1}{4M_{\rm Pl}^2}\left(\varphi^2 - {\varphi_{\rm end}}^2\right)\,,$$

and we estimate $\varphi_{\rm end}$ from $\varepsilon(\varphi_{\rm end}) = 2M_{\rm Pl}^2/{\varphi_{\rm end}}^2 = 1 \quad\Rightarrow\quad \varphi_{\rm end} = \sqrt{2}M_{\rm Pl}$ to get

$$\varphi^2 = {\varphi_{\rm end}}^2 + 4M_{\rm Pl}^2 N = 2M_{\rm Pl}^2 + 4M_{\rm Pl}^2 N \approx 4M_{\rm Pl}^2 N\,.$$

Thus

$$\left(\frac{M_{\rm Pl}}{\varphi}\right)^2 = \frac{1}{4N}$$

and

$$\begin{aligned} n_s &= 1 - \frac{2}{N} \;\approx\; 0.96 \\ \frac{dn_s}{d\ln k} &= -\frac{2}{N^2} \;\approx\; -0.0008 \\ r &= \frac{8}{N} \;\approx\; 0.16 \\ n_t &= -\frac{1}{N} \;\approx\; -0.02 \end{aligned}$$

We see that this model is ruled out by the observed upper limit $r < 0.1$.[42]

8.7.2 Scale invariance of the primordial power spectrum

Inflation predicts and observations give evidence for an almost scale invariant primordial power spectrum. Let us forget the "almost" for a moment and discuss what it means for the primordial spectrum to be scale invariant.

The primordial spectrum is something we have at superhorizon scales, where we have discussed it in terms of the comoving curvature perturbation $\mathcal{R}$, and we are calling it scale invariant, when

$$\mathcal{P}_{\mathcal{R}}(k) = A_s^2 = \text{const.} \tag{352}$$

We would like the spectrum in terms of more familiar concepts like the density perturbation, but at superhorizon scales the density perturbation is gauge dependent.

For small scales the perturbation spectrum gets modified when the scales enter the horizon, but for large scales $k \ll k_{\rm eq}$ the spectrum maintains its primordial shape, at least as long as

[42] There was enormous excitement in early 2014, when the BICEP2 collaboration[7] claimed to have detected the effect of primordial gravitational waves with $r = 0.20^{+0.07}_{-0.05}$ in CMB polarization, consistent with this inflation model. However, it turned out that their data was contaminated by polarized emission from dust in our own galaxy.[8]

the universe stays matter dominated. This allows the discussion of the primordial spectrum at subhorizon scales, where we can talk about the density perturbations without specifying a gauge.

From Eq. (255), the gravitational potential and density perturbation are related to the curvature perturbation as

$$\begin{aligned} \Phi_{\mathbf{k}} &= -\frac{3}{5}\mathcal{R}_{\mathbf{k}} \qquad \text{(mat.dom)} \\ \delta_{\mathbf{k}} &= -\frac{2}{3}\left(\frac{k}{\mathcal{H}}\right)^2 \Phi_{\mathbf{k}} \;=\; \frac{2}{5}\left(\frac{k}{\mathcal{H}}\right)^2 \mathcal{R}_{\mathbf{k}}\,, \end{aligned} \tag{353}$$

giving

$$\mathcal{P}_\Phi(k) = \frac{9}{25}\mathcal{P}_\mathcal{R}(k) = \frac{9}{25}A_s^2 = const \tag{354}$$

$$\begin{aligned} \mathcal{P}_\delta(t,k) &= \frac{4}{9}\left(\frac{k}{\mathcal{H}}\right)^4 \mathcal{P}_\Phi(k) = \frac{4}{25}\left(\frac{k}{\mathcal{H}}\right)^4 \mathcal{P}_\mathcal{R}(k) \\ &= \frac{4}{25}\left(\frac{k}{\mathcal{H}}\right)^4 A_s^2 \propto t^{4/3}k^4 \end{aligned} \tag{355}$$

Thus perturbations in the gravitational potential are scale invariant, but perturbations in density are not. Instead the density perturbation spectrum is steeply rising, meaning that there is much more structure at small scales than at large scales. Thus the scale invariance refers to the gravitational aspect of perturbations, which in the Newtonian treatment is described by the gravitational potential, and in the GR treatment by spacetime curvature.

The relation between density and gravitational potential perturbations reflects the nature of gravity: A 1% overdense region 100 Mpc across generates a much deeper potential well than a 1% overdense region 10 Mpc across, since the former has 1000 times more mass. Therefore we need much stronger density perturbations at smaller scales to have an equal contribution to Φ.

However, if we extrapolate Eq. (355) back to horizon entry, $k = \mathcal{H}$, we get

$$\delta_H^2(k) \equiv \text{“}\mathcal{P}_\delta(k, t_k)\text{”} \equiv \frac{4}{25}\mathcal{P}_\mathcal{R}(k) = \left(\frac{2}{5}A_s\right)^2 = const \tag{356}$$

Thus for scale-invariant primordial perturbations, *density perturbations of all scales enter the horizon with the same amplitude*, $\delta_H = (2/5)A_s \sim 2 \times 10^{-5}$. Since the density perturbation at the horizon entry is actually a gauge-dependent quantity, and our extension of the above Newtonian relation up to the horizon scale is not really allowed, this statement should be taken just qualitatively (hence the quotation marks around the $\mathcal{P}_\delta$). As such, it applies also to the smaller scales which enter during the radiation-dominated epoch, since the perturbations only begin to evolve after horizon entry.

What is the deep reason that inflation generates (almost) scale invariant perturbations? During inflation the universe is almost a de Sitter universe, which has the metric

$$ds^2 = -dt^2 + e^{2Ht}(dx^2 + dy^2 + dz^2)$$

with $H = const$. In GR we learn that it is an example of a “maximally symmetric spacetime”. In addition to being homogeneous (in the space directions), it also looks the same at all times. Therefore, as different scales exit at different times they all obtain the same kind of perturbations.

In terms of the other definition of the power spectrum, $P(k) \equiv (2\pi^2/k^3)\mathcal{P}(k)$, the relations (355) for scale-invariant perturbations give

$$\begin{aligned} P_\mathcal{R}(k) \;&\propto\; k^{-3}\mathcal{P}_R \;\propto\; k^{-3} \\ P_\delta(k) \;&\propto\; k^{-3}\mathcal{P}_\delta \;\propto\; k^4 P_\mathcal{R} \;\propto\; k\mathcal{P}_R \;\propto\; k \end{aligned} \tag{357}$$

For $\mathcal{P}_\mathcal{R}(k) \propto k^{n-1}$ we have $P_\delta(k) \propto k^n$. This is the reason for the -1 in the definition of the spectral index in terms of $\mathcal{P}_\mathcal{R}$—it was originally defined in terms of P_δ.

8.8 The power spectrum today

8.8.1 Density perturbations

From Eq. (261), the density perturbation spectrum at late times is

$$\mathcal{P}_\delta(k) = \frac{4}{25}\left(\frac{k}{\mathcal{H}}\right)^4 T(k,t)^2 \mathcal{P}_\mathcal{R}(k) \tag{358}$$

where, from Eq. (263)

$$\begin{aligned} T(k) &= 1 && \text{for} \quad k \ll k_\mathrm{eq} \\ T(k) &\sim \left(\frac{k_\mathrm{eq}}{k}\right)^2 \ln k && \text{for} \quad k \gg k_\mathrm{eq}\,. \end{aligned}$$

Thus the present-day density power spectrum rises steeply $\propto k^4$ at large scales, but turns at $\sim k_\mathrm{eq}$ to become less steep (growing $\sim \ln k$) at small scales. This is because the growth of density perturbations was inhibited while the perturbations were inside the horizon during the radiation-dominated epoch. The $\sim \ln k$ factor comes from the slow growth of CDM perturbations during this time.

Thus the structure in the universe appears stronger at smaller scales, down to $k_\mathrm{eq}^{-1} \sim 100\,\mathrm{Mpc}$. The $\sim 100\,\mathrm{Mpc}$ scale is indeed quite prominent in large scale structure surveys, like the 2dF-GRS and SDSS galaxy distribution surveys. Towards smaller scales the structure keeps getting stronger, but now more slowly. However, perturbations are now so large that first-order perturbation theory begins to fail, and that limit is crossed at around $k^{-1} \sim k_\mathrm{nl}^{-1} \sim 10\,\mathrm{Mpc}$. Nonlinear effects cause the density power spectrum to rise more steeply than calculated by perturbation theory at scales smaller than this.

The present-day density power spectrum $\mathcal{P}_\delta(k)$ can be determined observationally from the distribution of galaxies (Fig. 7). The quantity plotted is usually $P_\delta(k)$. It should go as

$$\begin{aligned} P_\delta(k) &\propto k^n && \text{for} \quad k \ll k_\mathrm{eq} \\ P_\delta(k) &\propto k^{n-4} \ln k && \text{for} \quad k \gg k_\mathrm{eq}\,. \end{aligned} \tag{359}$$

See Fig. 8.

8.8.2 Primordial gravitational waves

We found that outside the horizon tensor perturbations remain constant,

$$h_\mathbf{k}(t) = h_{\mathbf{k},\mathrm{prim}} = const\,, \tag{360}$$

whereas inside the horizon they become gravitational waves whose amplitude decays

$$|h_\mathbf{k}(t)| \propto a^{-1}\,. \tag{361}$$

Define the transfer function for gravitational waves

$$T_h(k) \equiv \frac{|h_\mathbf{k}(t_0)|}{h_{\mathbf{k},\mathrm{prim}}}\,, \tag{362}$$

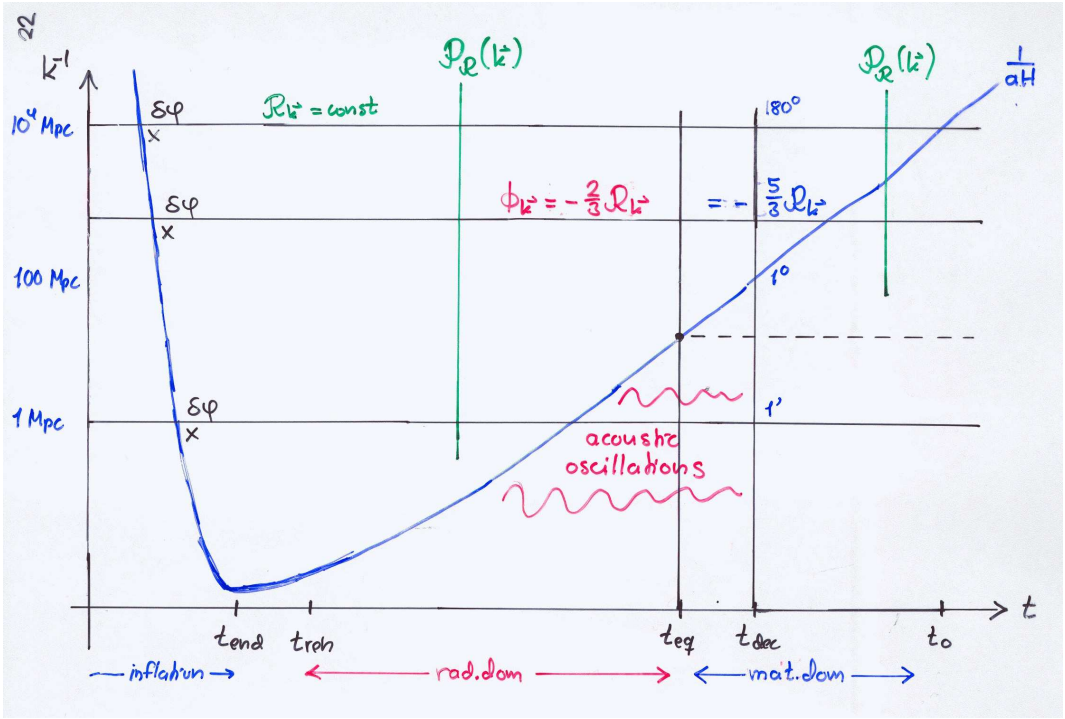


Figure 6: The whole picture of structure formation theory from quantum fluctuations during inflation to the present-day power spectrum at t_0.

so that the present-day power spectrum of primordial gravitational waves is

$$\mathcal{P}_{\rm grav}(k, t_0) = T_h(k)^2 \mathcal{P}_h(k)\,. \tag{363}$$

Make the approximation that the transition from (360) to (361) is instantaneous at horizon entry defined as

$$k = \mathcal{H} = aH\,. \tag{364}$$

Denote these values of a, H, and $\mathcal{H}$ by a_k, H_k, and $\mathcal{H}_k$. Then

$$T_h(k) \;=\; \frac{a_k}{a_0} \;=\; a_k\,. \tag{365}$$

The shape of the transfer function is determined by the rate at which different comoving scales k enter horizon as the universe expands. This is determined by the evolution of the comoving Hubble distance $\mathcal{H}^{-1}$.

In the matter-dominated universe

$$a \propto t^{2/3} \quad \text{and} \quad H = \frac{2}{3t} \propto a^{-3/2} \quad \Rightarrow \quad \mathcal{H} \propto a^{-1/2}\,. \tag{366}$$

Make first the approximation that the universe is still matter dominated. Then

$$T_h(k) \;=\; \frac{a_k}{a_0} \;=\; \left(\frac{\mathcal{H}_k}{\mathcal{H}_0}\right)^{-2} \;=\; \left(\frac{k}{a_0 H_0}\right)^{-2} \qquad (H_0 < k < k_{\rm eq}) \tag{367}$$

for scales that entered during the matter-dominated epoch.

To correct this result for the effect of dark energy at late times, we note that because of dark energy, the comoving Hubble distance $\mathcal{H}^{-1} = (aH)^{-1}$ stopped growing and began to shrink,

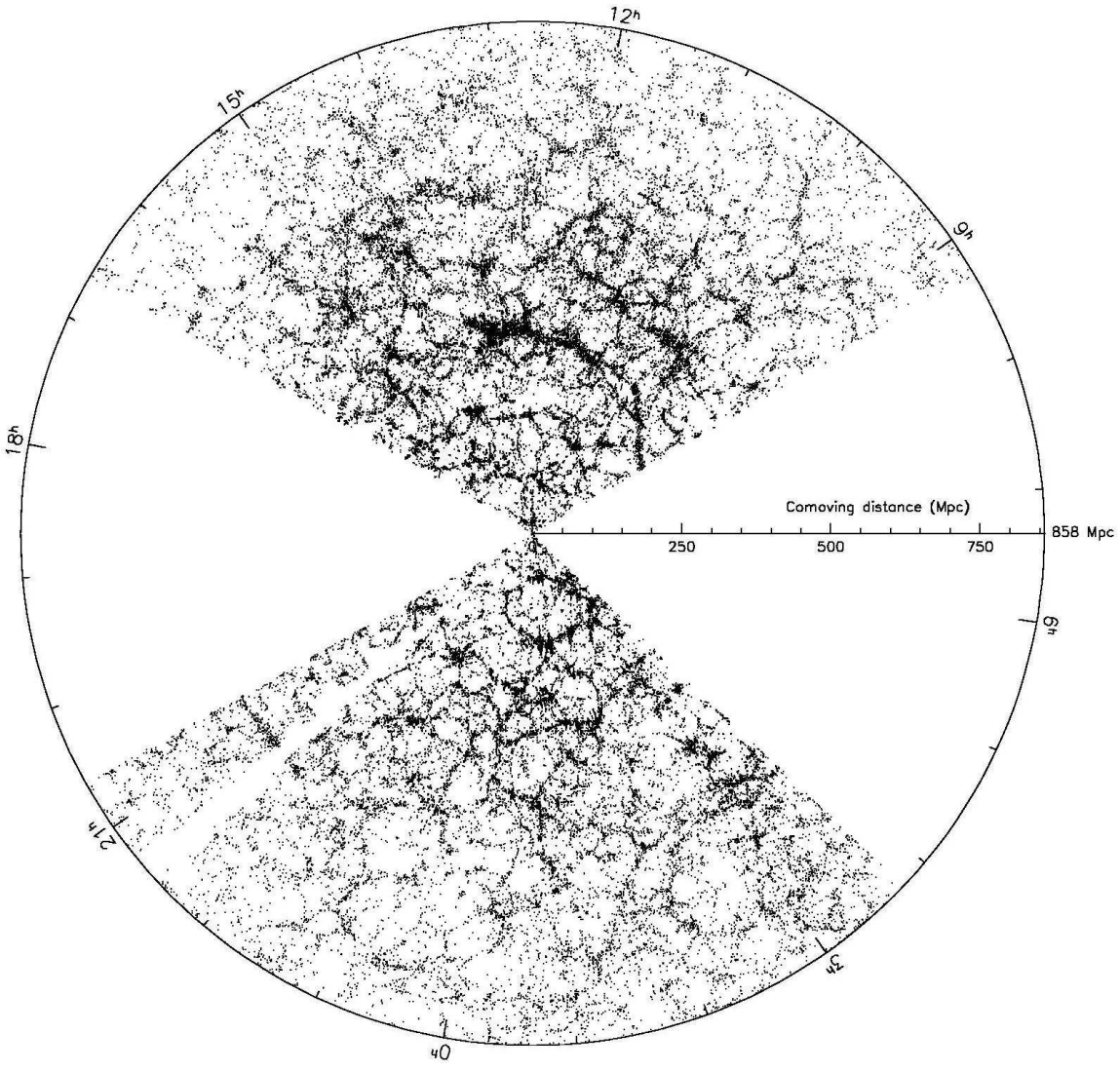


Figure 7: Distribution of galaxies according to the Sloan Digital Sky Survey (SDSS). This figure shows galaxies that are within 2° of the equator and closer than 858 Mpc (assuming $H_0 = 71$ km/s/Mpc). Figure from astro-ph/0310571[9].

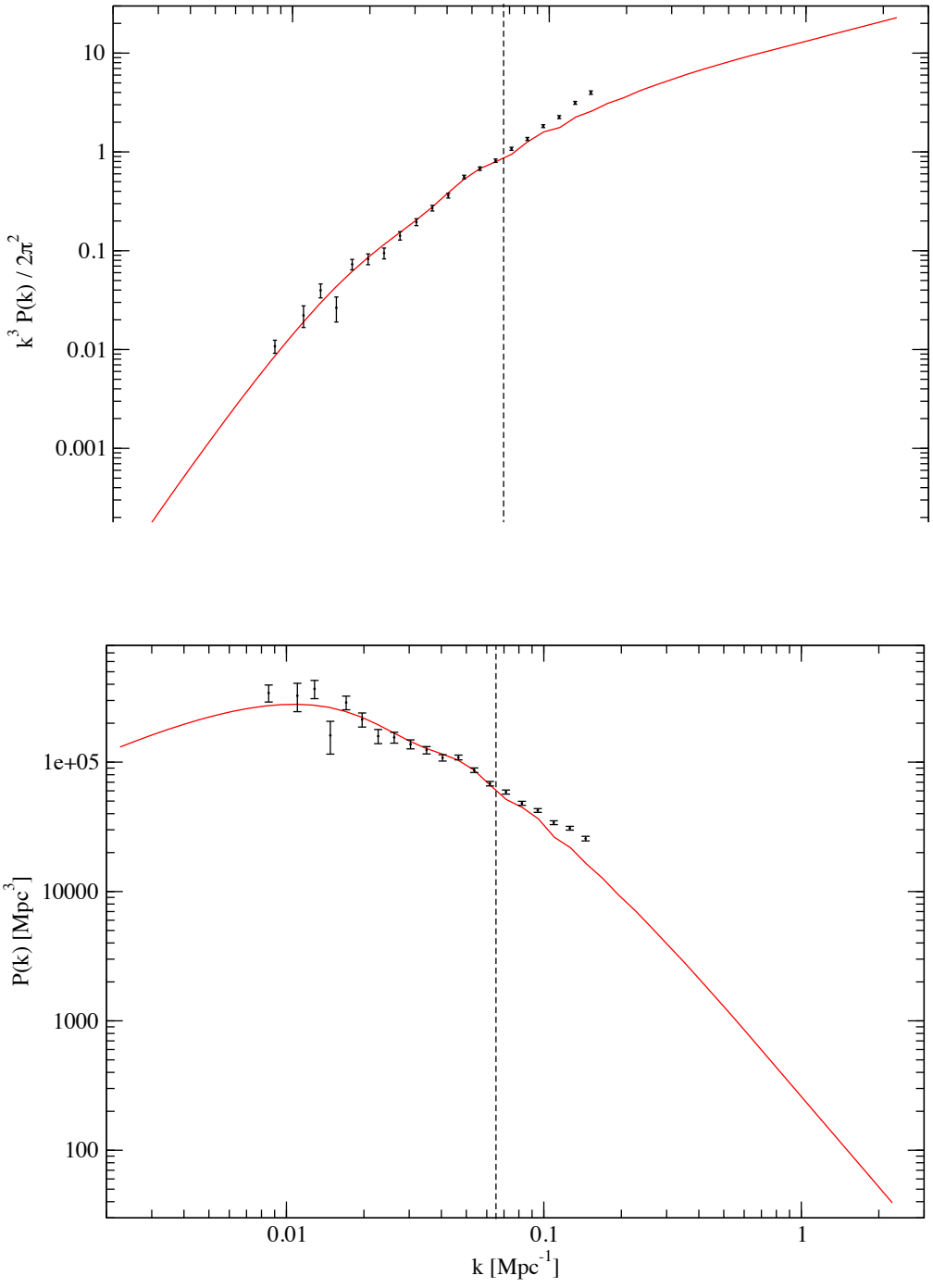


Figure 8: The matter power spectrum from the SDSS obtained using luminous red galaxies [10]. The top figure shows $\mathcal{P}_\delta(k)$ and the bottom figure $P_\delta(k)$. A Hubble constant value $H_0 = 71.4$ km/s/Mpc has been assumed for this figure. (These galaxy surveys only obtain the scales up to the Hubble constant, and therefore the observed $P_\delta(k)$ is usually shown in units of h Mpc^{-1}, so that no value for H_0 need to be assumed.) The black bars are the observations and the red curve is a theoretical fit, from linear perturbation theory, to the data. The bend in $P(k)$ at $k_{\text{eq}} \sim 0.01$ Mpc^{-1} is clearly visible in the bottom figure. Linear perturbation theory fails when $\mathcal{P}(k) \gtrsim 1$, and therefore the data points do not follow the theoretical curve to the right of the dashed line (representing an estimate on how far linear theory can be trusted). Figure by R. Keskitalo.

so that the scale $k = H_0$ is actually exiting now, and it entered at an earlier time t_1 when the expansion was still (barely) matter dominated. Thus the above result for $T_h(k)$ should apply (roughly) at that earlier time:

$$T_h(t_1, k) \;=\; \left(\frac{k}{a_1 H_1}\right)^{-2} \;=\; \left(\frac{k}{a_0 H_0}\right)^{-2} \qquad (H_0 < k < k_{\rm eq}) \tag{368}$$

While the scale $k = H_0$ was inside the horizon, the universe expanded by about a factor of two, so the correct transfer function is about half of (367).

Exercise: Extend the result (367) to scales $k > k_{\rm eq}$. You can make the approximation where the transition from radiation-dominated expansion law to matter-dominated expansion law is instantaneous at $t_{\rm eq}$. (This approximation actually underestimates $T_h(k > k_{\rm eq})$ by a factor that roughly compensates the overestimation in (367) from ignoring dark energy at late times.)

Gravitational waves were detected for the first time on September 14, 2015 at the LIGO observatory. These were not primordial gravitational waves; they were caused by a collision of two black holes about 400 Mpc from here, and they were observed only for about 0.2 seconds. The peak amplitude was $h \approx 10^{-21}$. LIGO is sensitive to frequencies near 100 Hz, and with further refinements it is expected to reach a sensitivity of $h = 10^{-22}$. Assume the primordial tensor perturbations had amplitude $h = 10^{-5}$ (close to the upper limit from CMB observations). What is their amplitude today at the 100 Hz frequency?

ESA is planning to launch a space gravitational wave observatory (LISA) in 2034. It would have similar sensitivity as LIGO, but for frequencies lower by a factor 10^{-4}. What do you conclude about the prospect for observing primordial gravitational waves this way?

References

[1] J.A. Peacock: Cosmological Physics (Cambridge University Press 1999), Chapter 16

[2] A.R. Liddle and D.H. Lyth: Cosmological Inflation and Large-Scale Structure (Cambridge University Press 2000)

[3] Planck Collaboration, Astronomy & Astrophysics **594**, A13 (2016), arXiv:1502.01589

[4] Planck Collaboration, *Planck 2018 results. VI. Cosmological parameters*, arXiv:1807.06209

[5] S. Dodelson: Modern Cosmology (Academic Press 2003), Chapter 7

[6] E.W. Kolb and M.S. Turner: The Early Universe (Addison-Wesley 1990)

[7] P.A.R. Ade et al., Phys. Rev. Lett. **112**, 241101 (2014), arXiv:1403.3985

[8] Planck Collaboration, Astronomy & Astrophysics **586**, A133 (2016), arXiv:1409.5738

[9] J. Richard Gott III et al., *A Map of the Universe*, Astrophys. J. **624**, 463 (2005), astro-ph/0310571

[10] M. Tegmark et al., *Cosmological Constraints from the SDSS Luminous Red Galaxies*, Phys. Rev. **D74**, 123507 (2006), astro-ph/0608632

9 Cosmic Microwave Background Anisotropy

9.1 Introduction

The cosmic microwave background (CMB) is isotropic to a high degree. This tells us that the early universe was rather homogeneous at the time ($t = t_{\mathrm{dec}} \approx 380\,000$ years) the CMB was formed. However, with precise measurements we can detect a low-level anisotropy in the CMB (Fig. 1) which reflects the small perturbations in the early universe.

This anisotropy was first detected by the COBE (Cosmic Background Explorer) satellite in 1992, which mapped the whole sky in three microwave frequencies. The angular resolution of COBE was rather poor, 7°, meaning that only features larger than this were detected. Measurements with better resolution, but covering only small parts of the sky were then performed using instruments carried by balloons to the upper atmosphere, and ground-based detectors located at high altitudes. A significant improvement came with the WMAP (Wilkinson Microwave Anisotropy Probe) satellite, which made observations for nine years, from 2001 to 2010.

The best CMB anisotropy data to date, covering the whole sky, has been provided by the Planck satellite (Fig. 2). Planck was launched by the European Space Agency (ESA), on May 14th, 2009, to an orbit around the L2 point of the Sun-Earth system, 1.5 million kilometers from the Earth in the anti-Sun direction. Planck made observations for over four years, from August 12th, 2009 until October 23rd, 2013. The first major release of Planck results was in 2013 [1] and the second release in 2015 [2]. Final Planck results are expected in 2018.

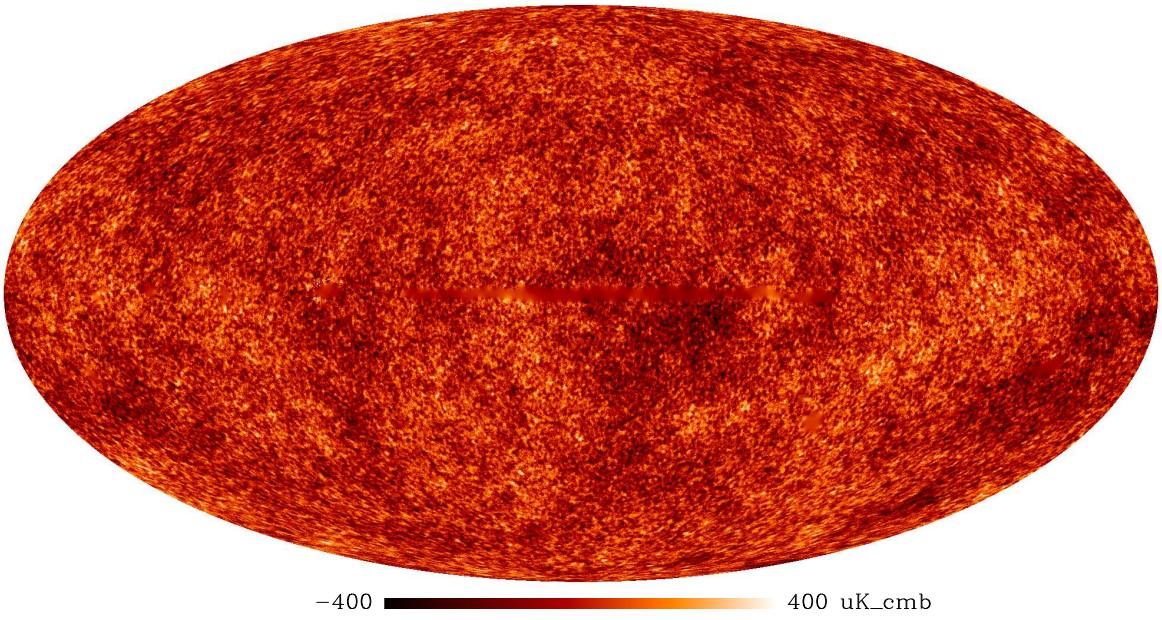


Figure 1: Cosmic microwave background. The figure shows temperature variations from $-400\,\mu$K to $+400\,\mu$K around the mean temperature (2.725 K) over the whole sky, in galactic coordinates. The color is chosen to mimic the true color of CMB at the time it was formed, when it was visible orange-red light, but the brightness variation (the anisotropy) is hugely exaggerated by the choice of color scale. The fuzzy regions, notable especially in the galactic plane, are regions of the sky where microwave radiation from our own galaxy or nearby galaxies makes it difficult to separate out the CMB. (ESA/Planck data).

Planck observed the entire sky twice in a year. The satellite repeated these observations year after year, and the results become gradually more accurate, since the effects of instrument noise averaged out and various instrument-related systematic effects could be determined and corrected better with repeated observations.

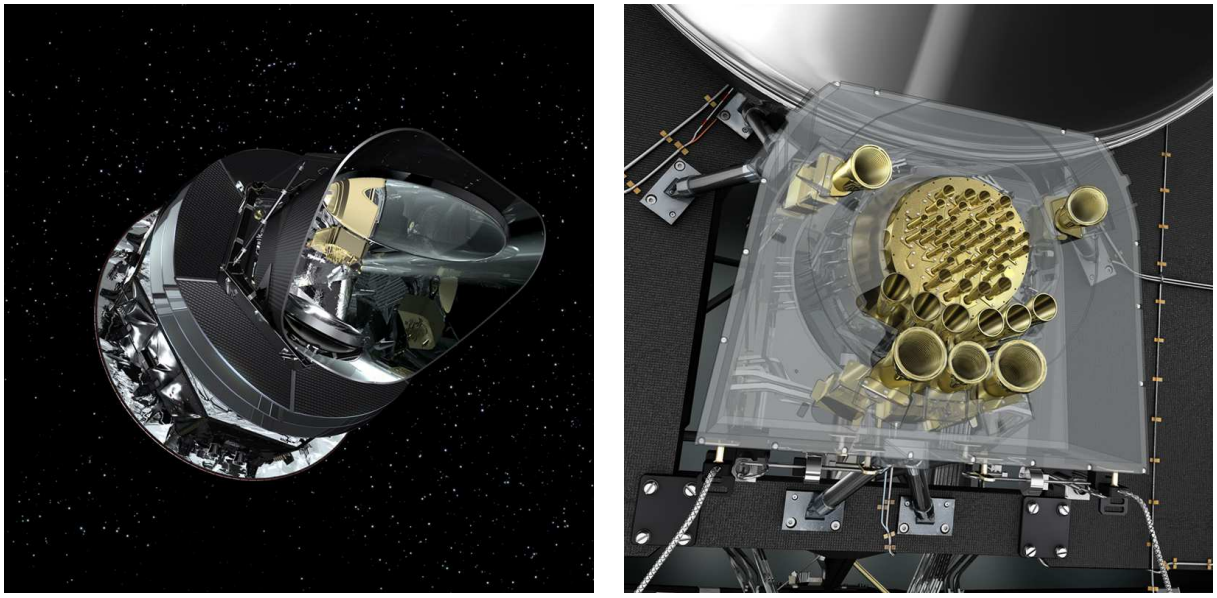


Figure 2: The Planck satellite and its microwave receivers. The larger horns are for receiving lower frequencies and the smaller horns for higher frequencies.

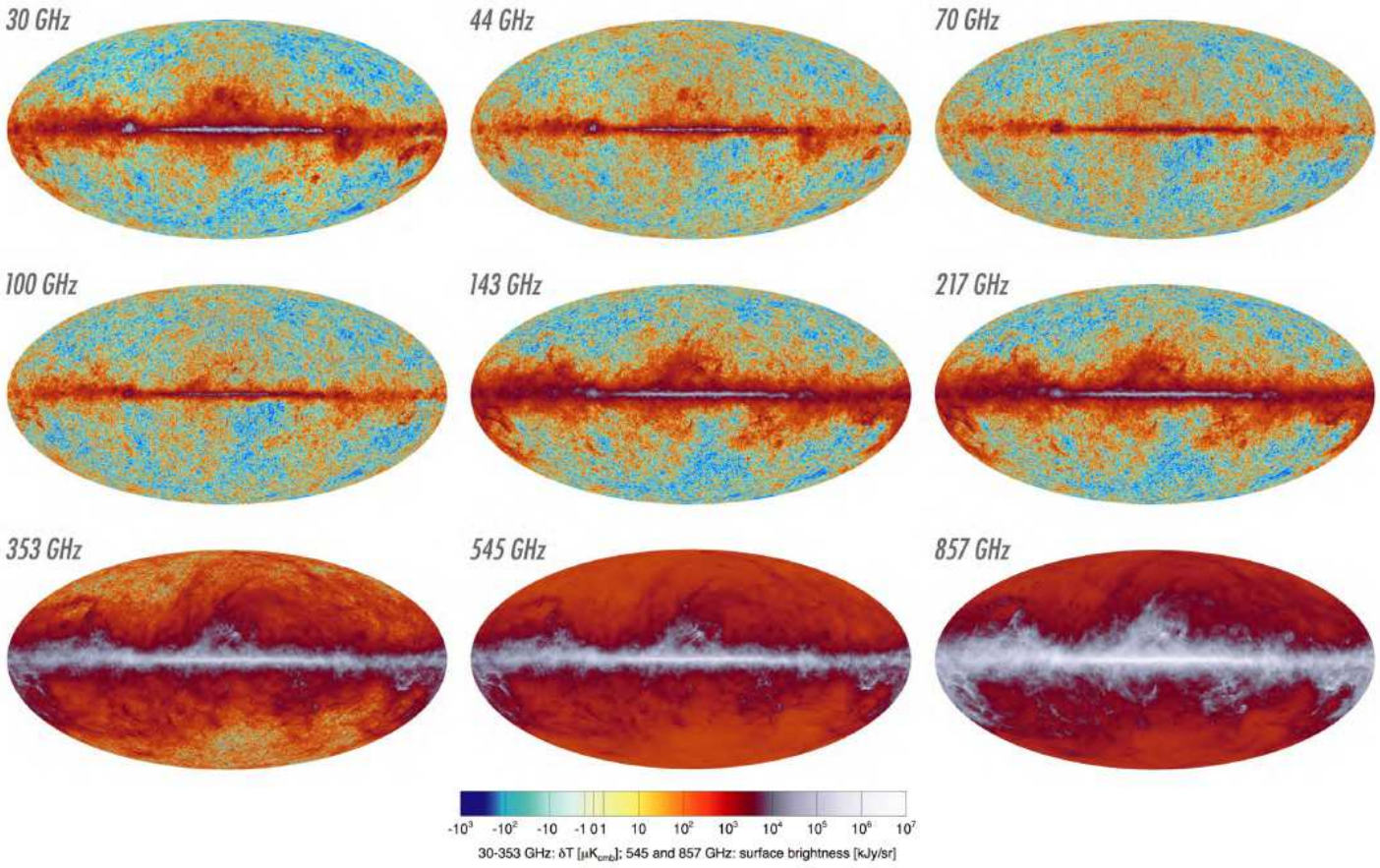


Figure 3: Brightness of the sky in the nine Planck frequency bands. These sky maps are in galactic coordinates so the Milky Way lies horizontally. From [2].

In addition to the CMB, there is microwave radiation from our own galaxy and other galaxies, called *foreground* by those who study CMB. This radiation can be separated from the CMB based on its different electromagnetic spectrum. To enable this *component separation*, Planck observed at 9 different frequency bands; the lowest one centered at 30 GHz and the highest at 857 GHz (Fig. 3). There were two different instruments on Planck, using different technologies to detect the variations in the microwave radiation. The Low Frequency Instrument (LFI) used radiometers for the 30, 44, and 70 GHz bands. The High Frequency Instrument (HFI) used bolometers for the bands from 100 GHz to 857 GHz. HFI is the barrel-shaped instrument at the center in Fig. 2 right panel and LFI was wrapped around it. With the additional help of WMAP and ground-based data 8 different foreground components could be distinguished (Fig. 4).

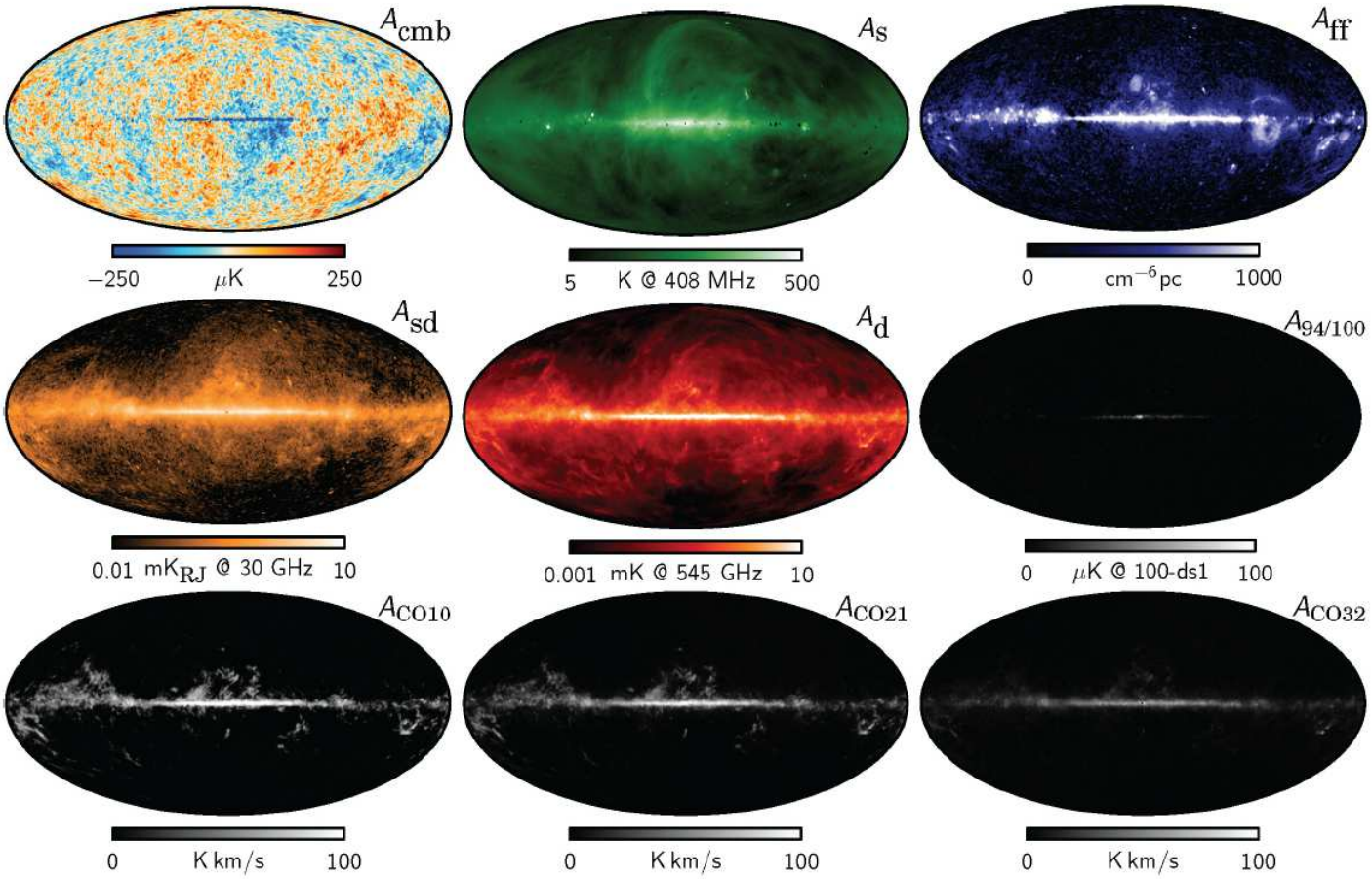


Figure 4: Result from Planck component separation. Also WMAP data and ground-based 408 MHz data was used. The extracted nine different components of the microwave radiation from top left to bottom right are: 1) CMB; 2) synchrotron radiation generated by relativistic cosmic-ray electrons accelerated by the galactic magnetic field; 3) "free-free emission" (bremsstrahlung) from electron-ion collisions; 4) emission from spinning galactic dust grains due to their electric dipole moment; 5) thermal emission from galactic dust (the typical dust temperatures are of order 20 K, so the dust thermal spectrum is peaked at much higher frequencies than CMB); 6) spectral line emission from HCN, CN, HCO, CS, and other molecules; 7) spectral line emission from the CO (carbon monoxide) $J = 1 \rightarrow 0$ transition; 8) CO $J = 2 \rightarrow 1$ line; 9) CO $J = 3 \rightarrow 2$ line (these emission lines from transitions between the four lowest rotation states of the CO molecule map the distribution of carbon monoxide in the Milky Way). From [2].

Figures 5–7 show the observed variation δT in the temperature of the CMB on the sky (red means hotter than average, blue means colder than average).

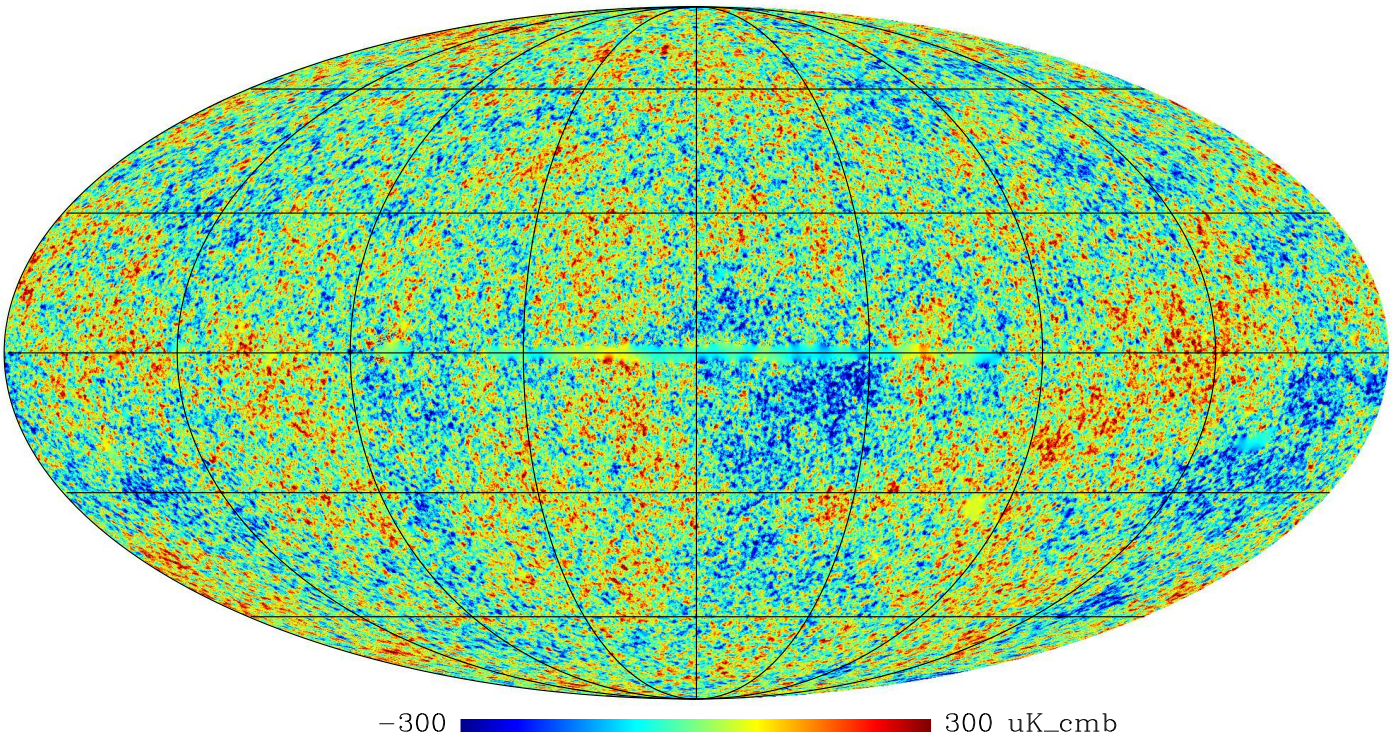


Figure 5: Cosmic microwave background: Fig. 1 reproduced in false color to bring out the patterns more clearly. The color range corresponds to CMB temperature variations from $-300\,\mu$K (blue) to $+300\,\mu$K (red) around the mean temperature. (ESA/Planck data).

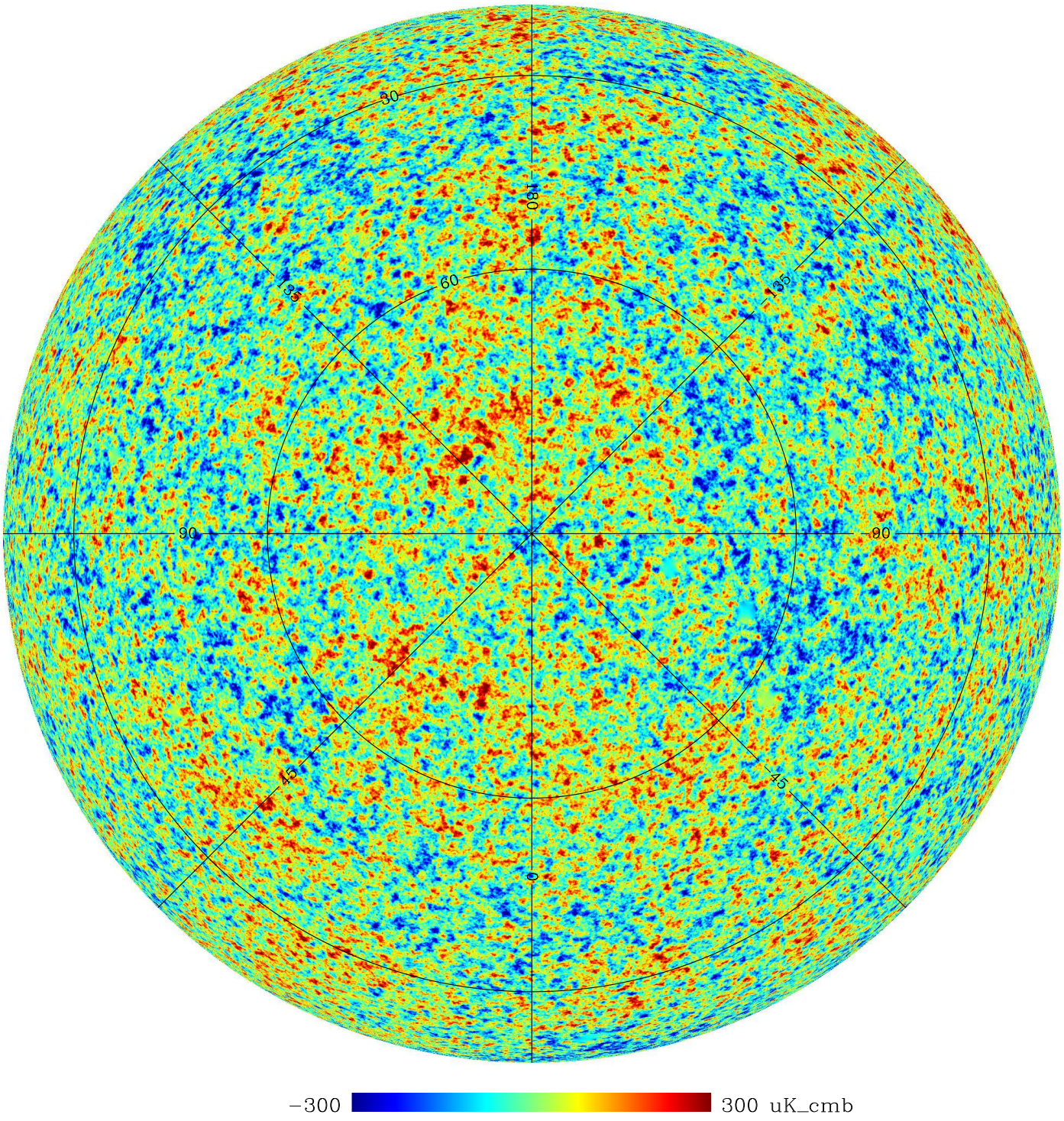


Figure 6: The northern galactic hemisphere of the CMB sky (ESA/Planck data).

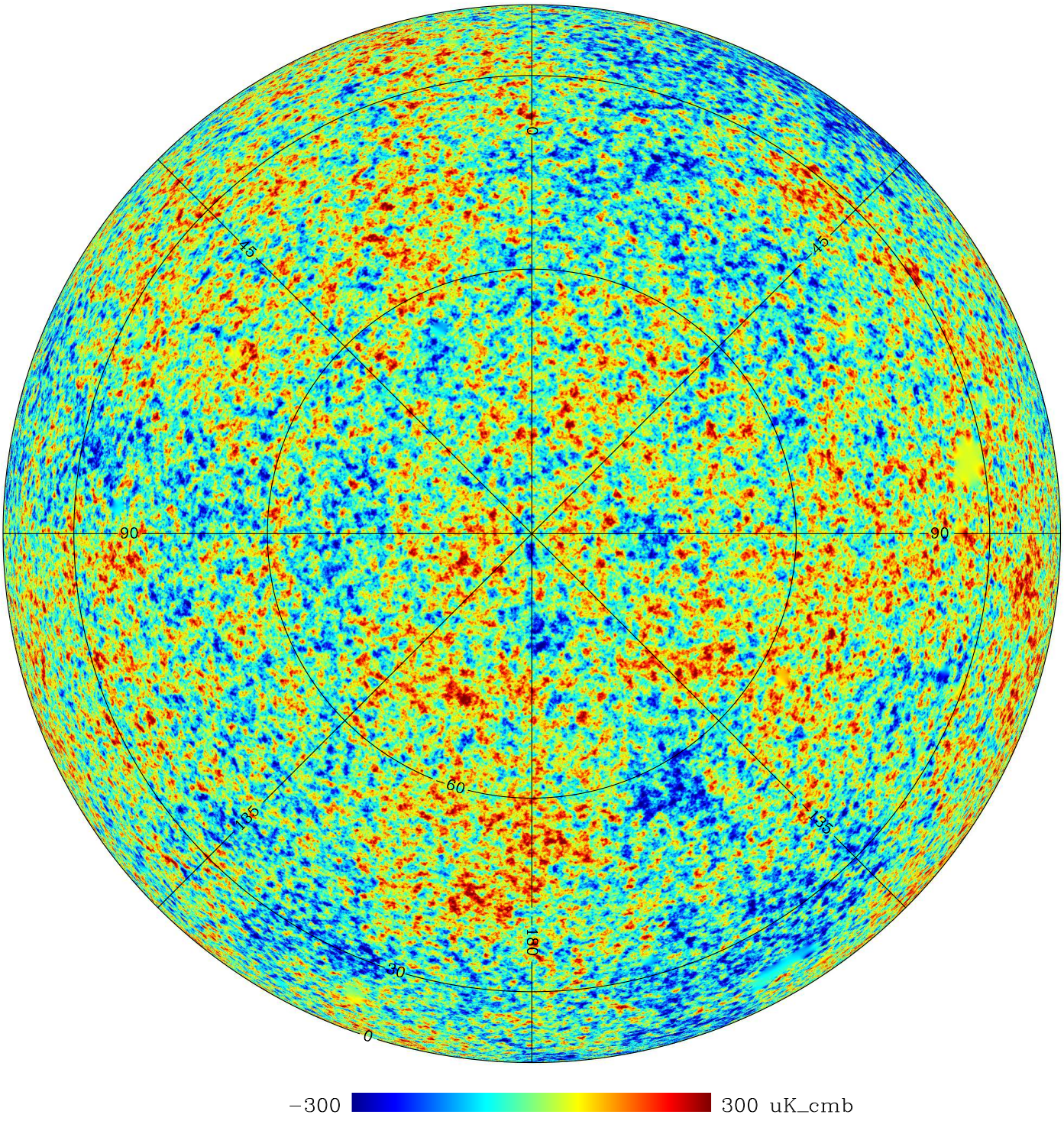


Figure 7: The southern galactic hemisphere of the CMB sky. The conspicuous cold region around $(-150°, -55°)$ is called the Cold Spot. The yellow smooth spot at $(-80°, -35°)$ in galactic coordinates is a region where the CMB is obscured by the Large Magellanic Cloud, and the light blue spot at $(-150°, -20°)$ is due to the Orion Nebula. (ESA/Planck data).

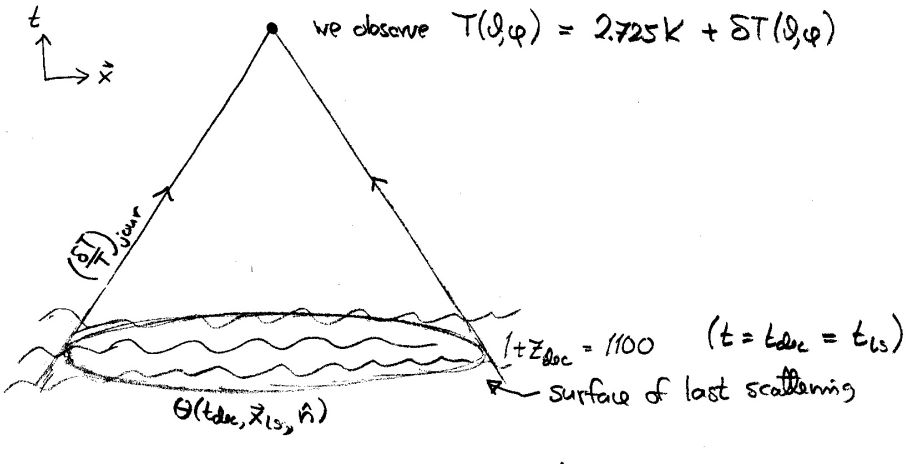


Figure 8: The observed CMB temperature anisotropy gets a contribution from the last scattering surface, $(\delta T/T)_{\rm intr} = \Theta(t_{\rm dec}, \mathbf{x}_{\rm ls}, \hat{\mathbf{n}})$ and from along the photon's journey to us, $(\delta T/T)_{\rm jour}$.

The photons we see as the CMB, have traveled to us from where our past light cone intersects the hypersurface corresponding to the time $t = t_{\rm dec}$ of photon decoupling. This intersection forms a sphere which we shall call the *last scattering surface*.[1] We are at the center of this sphere, except that timewise the sphere is located in the past.

The observed temperature anisotropy is due to two contributions, an *intrinsic* temperature variation at the surface of last scattering and a variation in the redshift the photons have suffered during their "journey" to us,

$$\left(\frac{\delta T}{T}\right)_{\rm obs} = \left(\frac{\delta T}{T}\right)_{\rm intr} + \left(\frac{\delta T}{T}\right)_{\rm jour} . \qquad (1)$$

See Fig. 8.

The first term, $\left(\frac{\delta T}{T}\right)_{\rm intr}$ represents the temperature variation of the photon gas at $t = t_{\rm dec}$. We also include in it the Doppler effect from the motion of this photon gas. At that time the larger scales we see in the CMB sky were still outside the horizon, so we have to pay attention to the gauge choice. In fact, the separation of $\delta T/T$ into the two components in Eq. (1) is gauge-dependent. If the time slice $t = t_{\rm dec}$ dips further into the past in some location, it finds a higher temperature, but the photons from there also have a longer way to go and suffer a larger redshift, so that the two effects balance each other. We can calculate in any gauge we want, getting different results for $(\delta T/T)_{\rm intr}$ and $(\delta T/T)_{\rm jour}$ depending on the gauge, but their sum $(\delta T/T)_{\rm obs}$ is gauge independent. It has to be, being an observed quantity.

One might think that $(\delta T/T)_{\rm intr}$ should be equal to zero, since in our earlier discussion of recombination and decoupling we identified decoupling with a particular temperature $T_{\rm dec} \sim 3000\,{\rm K}$. This kind of thinking corresponds to a particular gauge choice where the $t = t_{\rm dec}$ time slice coincides with the $T = T_{\rm dec}$ hypersurface. In this gauge $(\delta T/T)_{\rm intr} = 0$, except for the Doppler effect (we are not going to use this gauge). Anyway, it is not true that all photons have their last scattering exactly when $T = T_{\rm dec}$. Rather they occur during a rather large temperature interval and time period. The zeroth-order (background) time evolution of the temperature of the photon distribution is the same before and after last scattering, $T \propto a^{-1}$, so it does not matter how we draw the artificial separation line, the time slice $t = t_{\rm dec}$ separating the fluid and free particle treatments of the photons. See Fig. 9.

[1] Or the *last scattering sphere*. "Last scattering surface" often refers to the entire $t = t_{\rm dec}$ time slice.

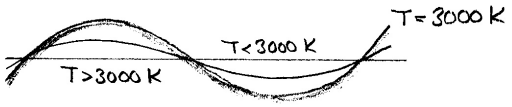


Figure 9: Depending on the gauge, the $T = T_{\rm dec}$ surface may, or (usually) may not coincide with the $t = t_{\rm dec}$ time slice.

9.2 Multipole analysis

The CMB temperature anisotropy is a function over a sphere (the celestial sphere, or the unit sphere of directions $\hat{\mathbf{n}}$). In analogy with the Fourier expansion in 3D space, we separate out the contributions of different angular scales by doing a multipole expansion,

$$\frac{\delta T}{T_0}(\theta,\phi) = \sum a_{\ell m} Y_{\ell m}(\theta,\phi) \tag{2}$$

where the sum runs over $\ell = 1, 2, \ldots \infty$ and $m = -\ell, \ldots, \ell$, giving $2\ell + 1$ values of m for each ℓ. The functions $Y_{\ell m}(\theta,\phi)$ are the *spherical harmonics* (see Fig. 10), which form an orthonormal set of functions over the sphere, so that we can calculate the multipole coefficients $a_{\ell m}$ from

$$a_{\ell m} = \int Y^*_{\ell m}(\theta,\phi)\frac{\delta T}{T_0}(\theta,\phi)d\Omega\,. \tag{3}$$

Definition (2) gives dimensionless $a_{\ell m}$. Often they are defined without the $T_0 = 2.725\,\mathrm{K}$ in Eq. (2), and then they have the dimension of temperature and are usually given in units of μK. Here θ and ϕ are spherical coordinates, $d\Omega \equiv d\cos\theta d\phi$, θ ranges from 0 to π and ϕ ranges from 0 to 2π.[2]

The sum begins at $\ell = 1$, since $Y_{00} = $ const. and therefore we must have $a_{00} = 0$ for a quantity which represents a deviation from average. The dipole part, $\ell = 1$, is dominated by the Doppler effect due to the motion of the solar system with respect to the last scattering surface, and we cannot separate out from it the *cosmological dipole* caused by large scale perturbations. Therefore we are here interested only in the $\ell \geq 2$ part of the expansion.

Another notation for $Y_{\ell m}(\theta,\phi)$ is $Y_{\ell m}(\hat{\mathbf{n}})$, where $\hat{\mathbf{n}}$ is a unit vector whose direction is specified by the angles θ and ϕ.

9.2.1 Spherical harmonics

We list here some useful properties of the spherical harmonics.

They are orthonormal functions on the sphere, so that

$$\int d\Omega\; Y_{\ell m}(\theta,\phi)Y^*_{\ell' m'}(\theta,\phi) = \delta_{\ell\ell'}\delta_{mm'}\,. \tag{4}$$

They are elementary complex functions and are related to the *associated Legendre functions* $P^m_\ell(x)$ by

$$Y_{\ell m}(\theta,\phi) = (-1)^m\sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}}P^m_\ell(\cos\theta)e^{im\phi}\,. \tag{5}$$

[2]They can also be given in degrees, the *colatitude* θ ranging from 0° (North) to 180° (South) and the *longitude* ϕ from 0° to 360°. There are a number of different astronomical coordinate systems (equatorial, ecliptic, galactic) in use, with their own historical conventions for the coordinate names, symbols, and units. Typically they involve the *latitude* $90° - \theta$ instead of the colatitude, so that North is at +90° and South at −90°, and the longitude is usually given between −180° and +180°, e.g., in Fig. 7.

Legendre polynomials
$P_0(x) = 1$
$P_1(x) = x$
$P_2(x) = \frac{1}{2}(3x^2 - 1)$
$P_3(x) = \frac{1}{2}(5x^3 - 3x)$
$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$

Associated Legendre functions $P_\ell^m(x) = P_\ell^m(\cos\theta)$
$P_1^1(x) = \sqrt{1-x^2} = \sin\theta$
$P_2^1(x) = 3x\sqrt{1-x^2} = 3\cos\theta\sin\theta$
$P_2^2(x) = 3(1-x^2) = 3\sin^2\theta$

Spherical harmonics
$Y_0^0(\theta,\phi) = \frac{1}{\sqrt{4\pi}}$
$Y_1^1(\theta,\phi) = -\sqrt{\frac{3}{8\pi}}\sin\theta e^{i\phi}$
$Y_1^0(\theta,\phi) = \sqrt{\frac{3}{4\pi}}\cos\theta$
$Y_2^2(\theta,\phi) = \sqrt{\frac{5}{96\pi}}3\sin^2\theta e^{i2\phi}$
$Y_2^1(\theta,\phi) = -\sqrt{\frac{5}{24\pi}}3\sin\theta\cos\theta e^{i\phi}$
$Y_2^0(\theta,\phi) = \sqrt{\frac{5}{4\pi}}\left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right)$

Spherical Bessel functions
$j_0(x) = \dfrac{\sin x}{x}$
$j_1(x) = \dfrac{\sin x}{x^2} - \dfrac{\cos x}{x}$
$j_2(x) = \left(\dfrac{3}{x^3} - \dfrac{1}{x}\right)\sin x - \dfrac{3}{x^2}\cos x$

Table 1: Legendre functions, spherical harmonics, and spherical Bessel functions.

Thus the θ-dependence is in $P_\ell^m(\cos\theta)$ and the ϕ-dependence is in $e^{im\phi}$. The functions P_ℓ^m are real and

$$Y_{\ell,-m} = (-1)^m Y_{\ell m}^*\,, \tag{6}$$

so that

$$Y_{\ell 0} = \sqrt{\frac{2\ell+1}{4\pi}}P_\ell(\cos\theta) \qquad \text{is real.} \tag{7}$$

The functions $P_\ell \equiv P_\ell^0$ are called *Legendre polynomials*. See Table 9.2.1 for examples of these functions for $\ell \le 2$.

Summing over the m corresponding to the same multipole number ℓ gives the *addition theorem*

$$\sum_m Y_{\ell m}^*(\theta',\phi')Y_{\ell m}(\theta,\phi) = \frac{2\ell+1}{4\pi}P_\ell(\cos\vartheta)\,, \tag{8}$$

where ϑ is the angle between $\hat{\mathbf{n}} = (\theta, \phi)$ and $\hat{\mathbf{n}}' = (\theta', \phi')$, i.e., $\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}' = \cos\vartheta$. For $\hat{\mathbf{n}} = \hat{\mathbf{n}}'$ this becomes

$$\sum_m |Y_{\ell m}(\theta, \phi)|^2 = \frac{2\ell + 1}{4\pi} \tag{9}$$

(since $P_\ell(1) = 1$ always).

We shall also need the expansion of a plane wave in terms of spherical harmonics,

$$e^{i\mathbf{k}\cdot\mathbf{x}} = 4\pi \sum_{\ell m} i^\ell j_\ell(kx) Y_{\ell m}(\hat{\mathbf{x}}) Y^*_{\ell m}(\hat{\mathbf{k}}) \,. \tag{10}$$

Here $\hat{\mathbf{x}}$ and $\hat{\mathbf{k}}$ are the unit vectors in the directions of $\mathbf{x}$ and $\mathbf{k}$, and the j_ℓ are the spherical Bessel functions.

9.2.2 Theoretical angular power spectrum

The CMB anisotropy is due to primordial perturbations, and therefore it reflects their Gaussian nature. Because one gets the values of the $a_{\ell m}$ from the other perturbation quantities through linear equations (in first-order perturbation theory), the $a_{\ell m}$ are also (complex) Gaussian random variables. Since they represent a deviation from the average temperature, their expectation value is zero,

$$\langle a_{\ell m} \rangle = 0 \,. \tag{11}$$

From statistical isotropy follows that the $a_{\ell m}$ are independent random variables so that

$$\langle a_{\ell m} a^*_{\ell' m'} \rangle = 0 \qquad \text{if } \ell \neq \ell' \text{ or } m \neq m'. \tag{12}$$

Since $\delta T/T_0$ is real,

$$a_{\ell,-m} = (-1)^m a^*_{\ell,m} \,. \tag{13}$$

Although thus $a_{\ell,-m}$ and $a_{\ell m}$ are not independent of each other, we still have $\langle a_{\ell m} a^*_{\ell,-m} \rangle = 0$ (**exercise**), so that (12) is satisfied even in this case. For each ℓ, there are $2\ell + 1$ independent real random variables: $a_{\ell 0}$ (which is always real), and $\operatorname{Re} a_{\ell m}$ and $\operatorname{Im} a_{\ell m}$ for $m = 1, \ldots, \ell$.

The quantity we want to calculate from theory is the variance $\langle |a_{\ell m}|^2 \rangle$ to get a prediction for the typical size of the $a_{\ell m}$. From statistical isotropy also follows that these expectation values depend only on ℓ not m. (The ℓ are related to the angular size of the anisotropy pattern, whereas the m are related to "orientation" or "pattern". See Fig. 10.) Since $\langle |a_{\ell m}|^2 \rangle$ is independent of m, we can define

$$C_\ell \equiv \langle |a_{\ell m}|^2 \rangle = \frac{1}{2\ell + 1} \sum_m \langle |a_{\ell m}|^2 \rangle \,, \tag{14}$$

and altogether we have

$$\langle a_{\ell m} a^*_{\ell' m'} \rangle = \delta_{\ell\ell'} \delta_{mm'} C_\ell \,. \tag{15}$$

This function C_ℓ (of integers $\ell \geq 2$) is called the (theoretical) *angular power spectrum*. It is analogous to the power spectrum $\mathcal{P}(k)$ of density perturbations. For Gaussian perturbations, the C_ℓ contains all the statistical information about the CMB temperature anisotropy. And this is all we can predict from theory. Thus the analysis of the CMB anisotropy consists of calculating the angular power spectrum from the observed CMB (a map like Figure 5) and comparing it to the C_ℓ predicted by theory.[3]

[3] In addition to the temperature anisotropy, the CMB also has another property, its polarization. There are two additional power spectra related to the polarization, C_ℓ^{EE} and C_ℓ^{BB}, and one related to the correlation between temperature and polarization, C_ℓ^{TE}.

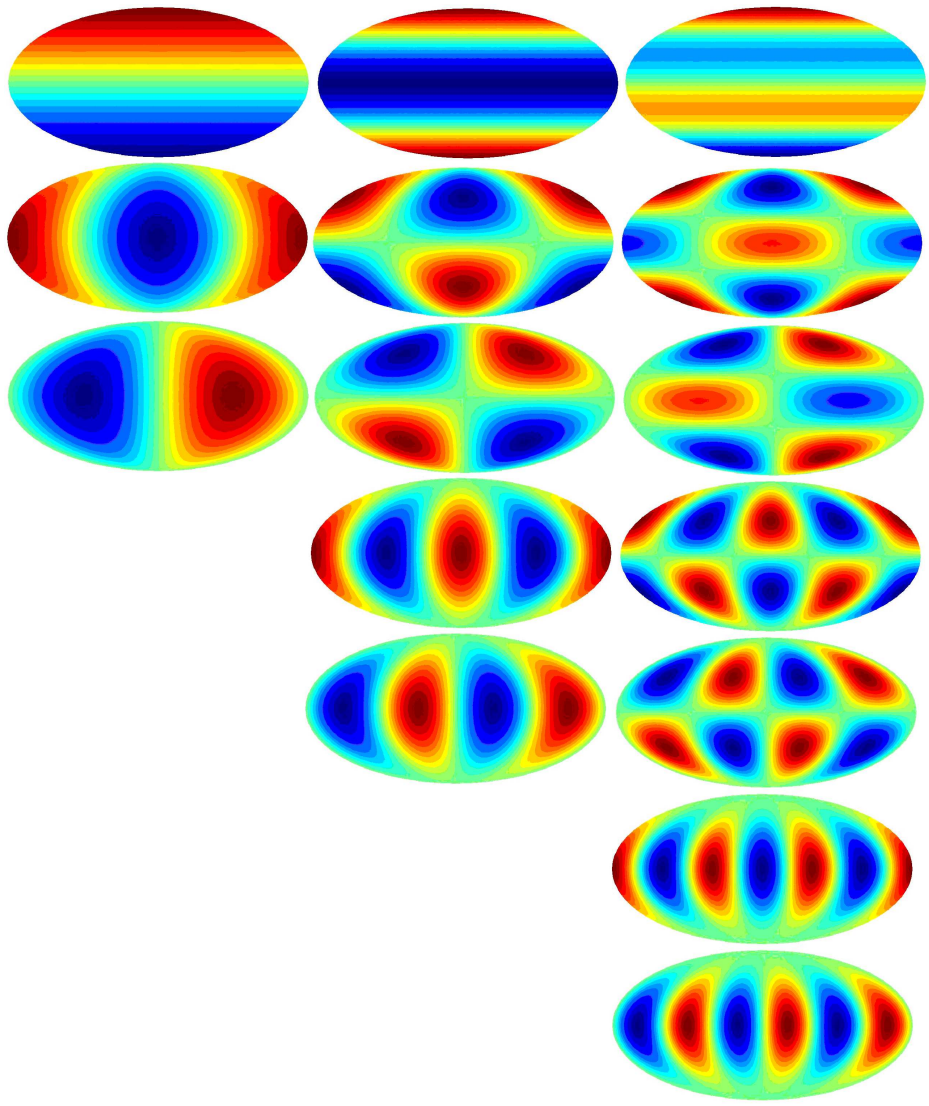


Figure 10: The three lowest multipoles $\ell = 1, 2, 3$ of spherical harmonics. Left column: Y_{10}, $\mathrm{Re}\, Y_{11}$, $\mathrm{Im}\, Y_{11}$. Middle column: Y_{20}, $\mathrm{Re}\, Y_{21}$, $\mathrm{Im}\, Y_{21}$, $\mathrm{Re}\, Y_{22}$, $\mathrm{Im}\, Y_{22}$. Right column: Y_{30}, $\mathrm{Re}\, Y_{31}$, $\mathrm{Im}\, Y_{31}$, $\mathrm{Re}\, Y_{32}$, $\mathrm{Im}\, Y_{32}$, $\mathrm{Re}\, Y_{33}$, $\mathrm{Im}\, Y_{33}$. Figure by Ville Heikkilä.

Just like the 3D density power spectrum $\mathcal{P}(k)$ gives the contribution of scale k to the density variance $\langle\delta(\mathbf{x})^2\rangle$, the angular power spectrum C_ℓ is related to the contribution of multipole ℓ to the temperature variance,

$$\begin{aligned}\left\langle\left(\frac{\delta T(\theta,\phi)}{T}\right)^2\right\rangle &= \left\langle\sum_{\ell m} a_{\ell m}Y_{\ell m}(\theta,\phi)\sum_{\ell' m'} a^*_{\ell' m'}Y^*_{\ell' m'}(\theta,\phi)\right\rangle \\ &= \sum_{\ell\ell'}\sum_{mm'} Y_{\ell m}(\theta,\phi)Y^*_{\ell' m'}(\theta,\phi)\langle a_{\ell m}a^*_{\ell' m'}\rangle \\ &= \sum_\ell C_\ell\sum_m |Y_{\ell m}(\theta,\phi)|^2 \;=\; \sum_\ell \frac{2\ell+1}{4\pi}C_\ell\,, \end{aligned} \tag{16}$$

where we used (15) and (9).

Thus, if we plot $(2\ell+1)C_\ell/4\pi$ on a linear ℓ scale, or $\ell(2\ell+1)C_\ell/4\pi$ on a logarithmic ℓ scale, the area under the curve gives the temperature variance, i.e., the expectation value for the squared deviation from the average temperature. It has become customary to plot the angular power spectrum as $\ell(\ell+1)C_\ell/2\pi$, which is neither of these, but for large ℓ approximates the second case. The reason for this custom is explained later.

Equation (16) represents the expectation value from theory and thus it is the same for all directions θ,ϕ. The actual, "realized", value of course varies from one direction θ,ϕ to another. We can imagine an ensemble of universes, otherwise like our own, but representing a different realization of the same random process of producing the primordial perturbations. Then $\langle\ \rangle$ represents the average over such an ensemble.

Equation(16) can be generalized to the angular correlation function (**exercise**)

$$C(\vartheta) \;\equiv\; \left\langle\frac{\delta T(\hat{\mathbf{n}})}{T}\frac{\delta T(\hat{\mathbf{n}}')}{T}\right\rangle \;=\; \frac{1}{4\pi}\sum_\ell (2\ell+1)C_\ell P_\ell(\cos\vartheta),\,, \tag{17}$$

where ϑ is the angle between $\hat{\mathbf{n}}$ and $\hat{\mathbf{n}}'$.

9.2.3 Observed angular power spectrum

Theory predicts expectation values $\langle|a_{\ell m}|^2\rangle$ from the random process responsible for the CMB anisotropy, but we can observe only one realization of this random process, the set $\{a_{\ell m}\}$ of our CMB sky. We define the *observed* angular power spectrum as the average

$$\widehat{C}_\ell = \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2 \tag{18}$$

of these observed values.

The variance of the observed temperature anisotropy is the average of $\left(\frac{\delta T(\theta,\phi)}{T}\right)^2$ over the celestial sphere,

$$\begin{aligned}\frac{1}{4\pi}\int\left[\frac{\delta T(\theta,\phi)}{T}\right]^2 d\Omega &= \frac{1}{4\pi}\int d\Omega\sum_{\ell m} a_{\ell m}Y_{\ell m}(\theta,\phi)\sum_{\ell' m'} a^*_{\ell' m'}Y^*_{\ell' m'}(\theta,\phi) \\ &= \frac{1}{4\pi}\sum_{\ell m}\sum_{\ell' m'} a_{\ell m}a^*_{\ell' m'}\underbrace{\int Y_{\ell m}(\theta,\phi)Y^*_{\ell' m'}(\theta,\phi)d\Omega}_{\delta_{\ell\ell'}\delta_{mm'}} \\ &= \frac{1}{4\pi}\sum_\ell\underbrace{\sum_m |a_{\ell m}|^2}_{(2\ell+1)\widehat{C}_\ell} \;=\; \sum_\ell\frac{2\ell+1}{4\pi}\widehat{C}_\ell\,. \end{aligned} \tag{19}$$

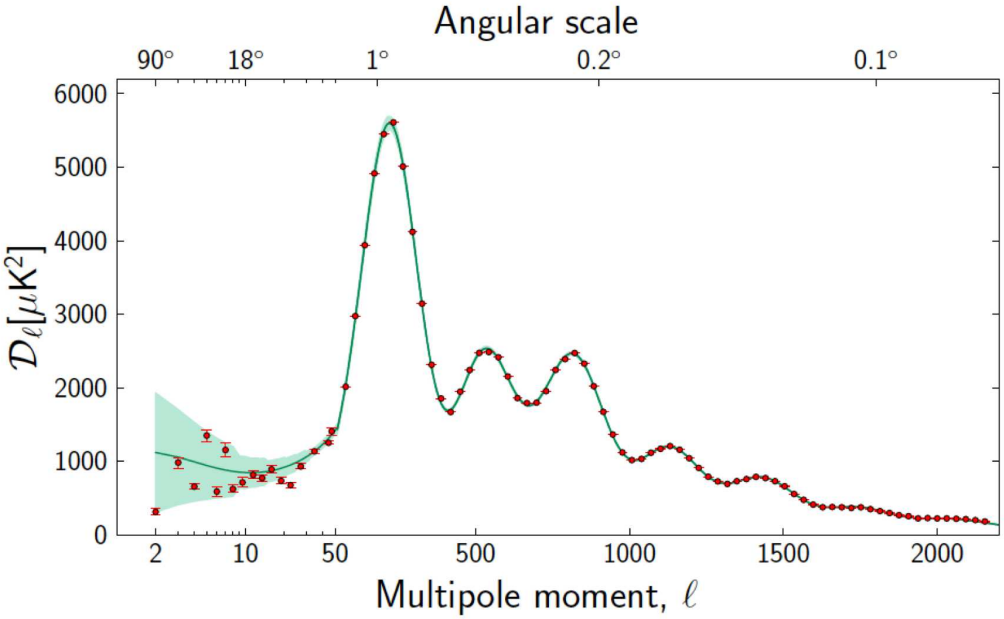


Figure 11: The angular power spectrum $\widehat{C}_\ell$ as observed by Planck. The observational results are the red data points with small error bars. The green curve is the theoretical C_ℓ from a best-fit model, and the light green band around it represents the cosmic variance corresponding to this C_ℓ. The quantity plotted is actually $\mathcal{D}_\ell \equiv T_0^2[\ell(\ell+1)/(2\pi)]C_\ell$. Note that the ℓ-axis is logarithmic until 50 and linear after that. (This is Fig. 21 of [1].)

Contrast this with (16), which gives the variance of $\delta T/T$ at an arbitrary location on the sky over different realizations of the random process which produced the primordial perturbations; whereas (19) gives the variance of $\delta T/T$ of our given sky over the celestial sphere.

9.2.4 Cosmic Variance

The expectation value of the observed spectrum $\widehat{C}_\ell$ is equal to C_ℓ, the *theoretical* spectrum of Eq. (14), i.e.,

$$\langle \widehat{C}_\ell \rangle = C_\ell \quad \Rightarrow \quad \langle \widehat{C}_\ell - C_\ell \rangle = 0\,, \tag{20}$$

but its actual, realized, value is not, although we expect it to be close. The expected squared difference between $\widehat{C}_\ell$ and C_ℓ is called the *cosmic variance*. We can calculate it using the properties of (complex) Gaussian random variables (**exercise**). The answer is

$$\langle (\widehat{C}_\ell - C_\ell)^2 \rangle = \frac{2}{2\ell+1} C_\ell^2\,. \tag{21}$$

We see that the expected relative difference between $\widehat{C}_\ell$ and C_ℓ is smaller for higher ℓ. This is because we have a larger (size $2\ell+1$) statistical sample of $a_{\ell m}$ available for calculating the $\widehat{C}_\ell$.

The cosmic variance limits the accuracy of comparison of CMB observations with theory, especially for large scales (low ℓ). See Fig. 11.

9.3 Multipoles and scales

9.3.1 Rough correspondence

The different multipole numbers ℓ correspond to different angular scales, low ℓ to large scales and high ℓ to small scales. Examination of the functions $Y_{\ell m}(\theta, \phi)$ reveals that they have an

oscillatory pattern on the sphere, so that there are typically ℓ "wavelengths" of oscillation around a full great circle of the sphere. See Figs. 10 and 12.

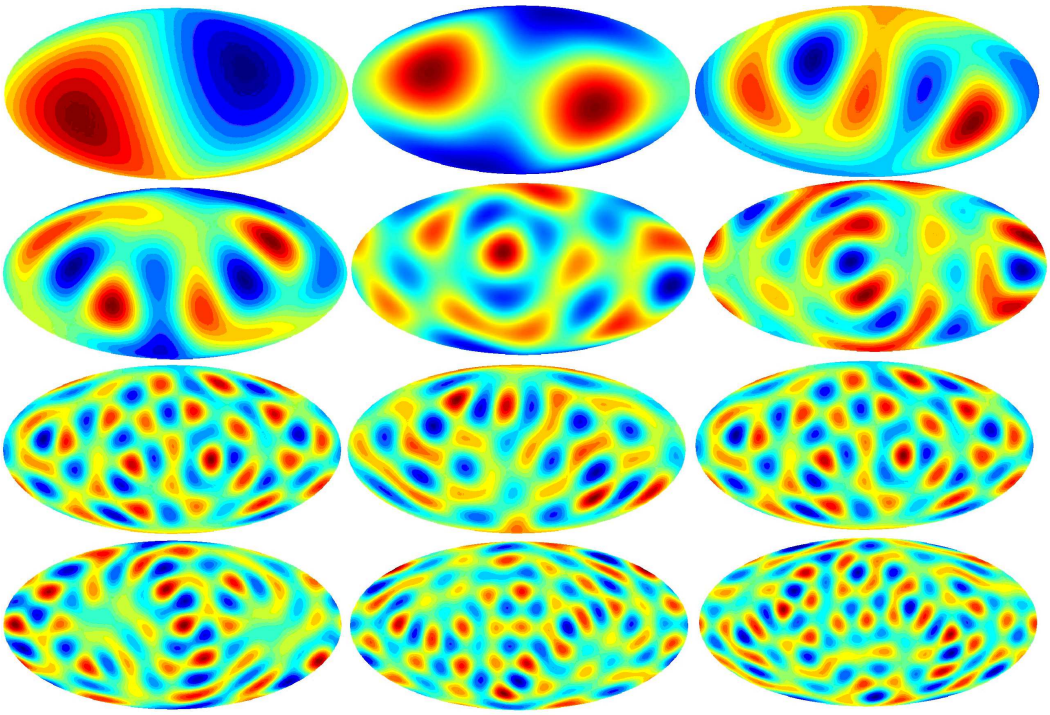


Figure 12: Randomly generated skies containing only a single multipole ℓ. Starting from top left: $\ell = 1$ (dipole only), 2 (quadrupole only), 3 (octupole only), 4, 5, 6, 7, 8, 9, 10, 11, 12. Figure by Ville Heikkilä.

Thus the angle corresponding to this wavelength is

$$\vartheta_\lambda = \frac{2\pi}{\ell} = \frac{360^\circ}{\ell}\,. \tag{22}$$

See Fig. 13. The angle corresponding to a "half-wavelength", i.e., the separation between a neighboring minimum and maximum is then

$$\vartheta_{\rm res} = \frac{\pi}{\ell} = \frac{180^\circ}{\ell}\,. \tag{23}$$

This is the angular resolution required of the microwave detector for it to be able to resolve the angular power spectrum up to this ℓ.

For example, COBE had an angular resolution of 7° allowing a measurement up to $\ell = 180/7 = 26$, WMAP had resolution 0.23° reaching to $\ell = 180/0.23 = 783$, and Planck had resolution 5′, allowing the measurement of C_ℓ up to $\ell = 2160$.[4]

The angles on the sky are related to actual physical distances via the *angular diameter distance* d_A, defined as the ratio of the physical length (transverse to the line of sight) and the angle it covers (see Chapter 3),

$$d_A \equiv \frac{\lambda_{\rm phys}}{\vartheta}\,. \tag{24}$$

Likewise, we defined the *comoving angular diameter distance* d_A^c by

$$d_A^c \equiv \frac{\lambda^c}{\vartheta} \tag{25}$$

[4]In reality, there is no sharp cut-off at a particular ℓ, the observational error bars just blow up rapidly around this value of ℓ.

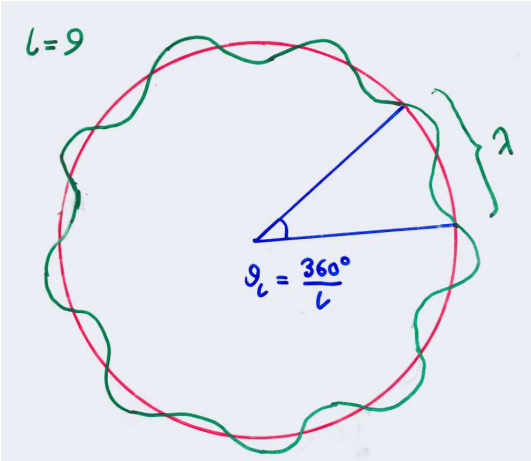


Figure 13: The rough correspondence between multipoles ℓ and angles.

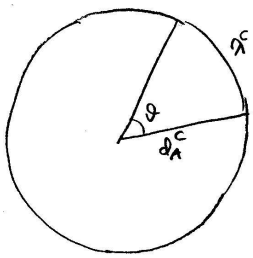


Figure 14: The comoving angular diameter distance relates the comoving size of an object and the angle in which we see it.

where $\lambda^c = a^{-1}\lambda_{\rm phys} = (1+z)\lambda_{\rm phys}$ is the corresponding comoving length. Thus $d_A^c = a^{-1}d_A = (1+z)d_A$. See Fig. 14.

Consider now the Fourier modes of our earlier perturbation theory discussion. A mode with comoving wavenumber k has comoving wavelength $\lambda^c = 2\pi/k$. Thus this mode should show up as a pattern on the CMB sky with angular size

$$\vartheta_\lambda = \frac{\lambda^c}{d_A^c} = \frac{2\pi}{kd_A^c} = \frac{2\pi}{\ell}\,. \tag{26}$$

For the last equality we used the relation (22). From it we get that the modes with wavenumber k contribute mostly to multipoles around

$$\ell = kd_A^c\,. \tag{27}$$

9.3.2 Exact treatment

The above matching of wavenumbers with multipoles was of course rather naive, for two reasons:

1. The description of a spherical harmonic $Y_{\ell m}$ having an "angular wavelength" of $2\pi/\ell$ is just a crude characterization. See Fig. 12.

2. The modes $\mathbf{k}$ are not wrapped around the sphere of last scattering, but the wave vector forms a different angle with the sphere at different places.

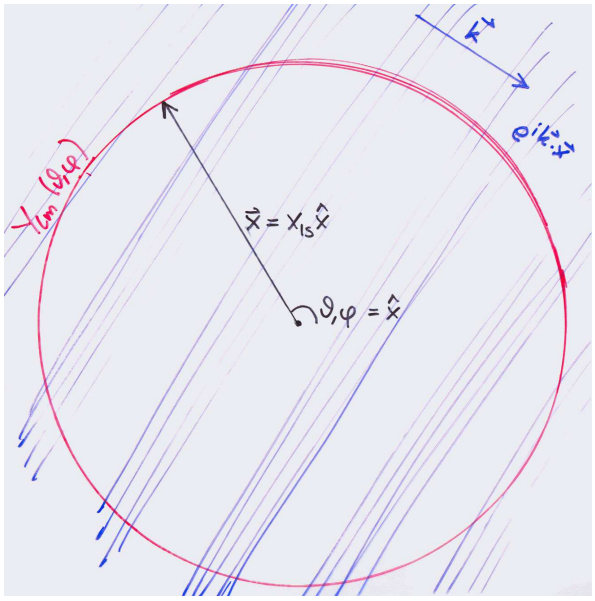


Figure 15: A plane wave intersecting the last scattering sphere.

The following precise discussion applies only for the case of a flat universe ($K = 0$ Friedmann model as the background), where one can Fourier expand functions on a time slice. We start from the expansion of the plane wave in terms of spherical harmonics, for which we have the result, Eq. (10),

$$e^{i\mathbf{k}\cdot\mathbf{x}} = 4\pi \sum_{\ell m} i^\ell j_\ell(kx) Y_{\ell m}(\hat{\mathbf{x}}) Y^*_{\ell m}(\hat{\mathbf{k}})\,, \tag{28}$$

where j_ℓ is the spherical Bessel function.

Consider now some function

$$f(\mathbf{x}) = \sum_{\mathbf{k}} f_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} \tag{29}$$

on the $t = t_{\rm dec}$ time slice. We want the multipole expansion of the values of this function on the last scattering sphere. See Fig. 15. These are the values $f(x\hat{\mathbf{x}})$, where $x \equiv |\mathbf{x}|$ has a constant value, the (comoving) radius of this sphere. Thus

$$\begin{aligned} a_{\ell m} &= \int d\Omega_x Y^*_{\ell m}(\hat{\mathbf{x}}) f(x\hat{\mathbf{x}}) \\ &= \sum_{\mathbf{k}} \int d\Omega_x Y^*_{\ell m}(\hat{\mathbf{x}}) f_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} \\ &= 4\pi \sum_{\mathbf{k}} \sum_{\ell' m'} \int d\Omega_x f_{\mathbf{k}} Y^*_{\ell m}(\hat{\mathbf{x}}) i^{\ell'} j_{\ell'}(kx) Y_{\ell' m'}(\hat{\mathbf{x}}) Y^*_{\ell' m'}(\hat{\mathbf{k}}) \\ &= 4\pi i^\ell \sum_{\mathbf{k}} f_{\mathbf{k}} j_\ell(kx) Y^*_{\ell m}(\hat{\mathbf{k}})\,, \end{aligned} \tag{30}$$

where we used the orthonormality of the spherical harmonics. The corresponding result for a Fourier transform $f(\mathbf{k})$ is

$$a_{\ell m} = \frac{4\pi i^\ell}{(2\pi)^3} \int d^3k f(\mathbf{k}) j_\ell(kx) Y^*_{\ell m}(\hat{\mathbf{k}})\,. \tag{31}$$

The j_ℓ are oscillating functions with decreasing amplitude. For large values of ℓ the position of the first (and largest) maximum is near $kx = \ell$ (see Fig. 16).

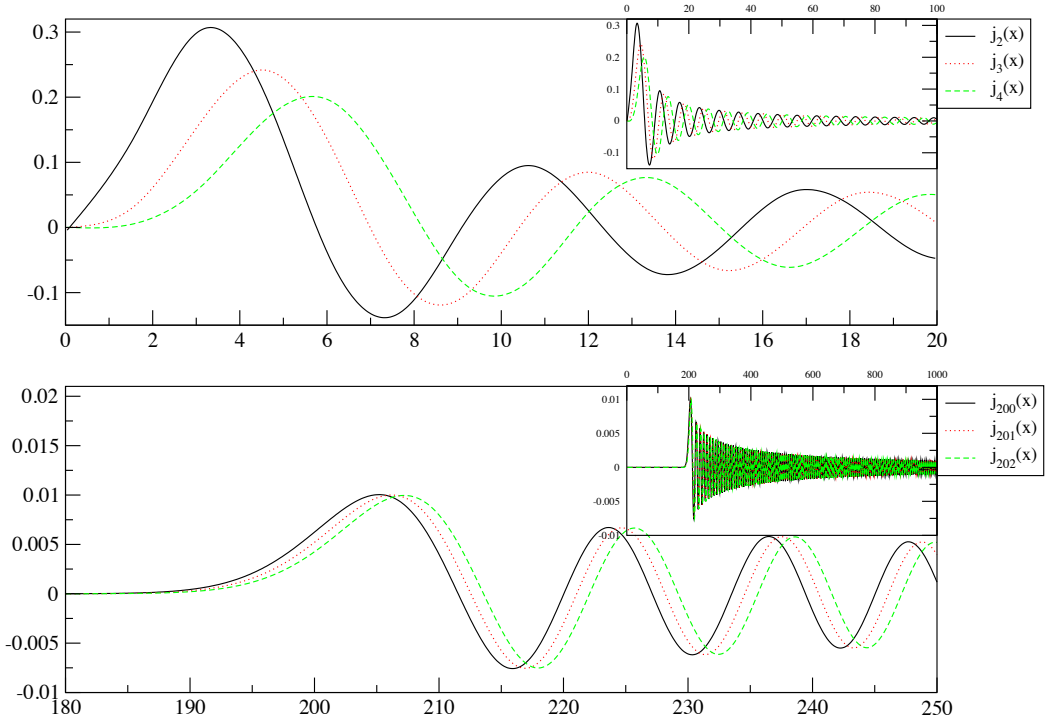


Figure 16: Spherical Bessel functions $j_\ell(x)$ for $\ell = 2$, 3, 4, 200, 201, and 202. Note how the first and largest peak is near $x = \ell$ (but to be precise, at a slightly larger value). Figure by R. Keskitalo.

Thus the $a_{\ell m}$ pick a large contribution from those Fourier modes $\mathbf{k}$ where

$$kx \sim \ell\,. \tag{32}$$

In a flat universe the comoving distance x (from our location to the sphere of last scattering) and the comoving angular diameter distance d_A^c are equal, so we can write this result as

$$kd_A^c \sim \ell\,. \tag{33}$$

The conclusion is that a given multipole ℓ acquires a contribution from modes with a range of wavenumbers, but most of the contribution comes from near the value given by Eq. (27). This concentration is tighter for larger ℓ.

We shall use Eq. (27) for qualitative purposes in the following discussion.

9.4 Important distance scales on the last scattering surface

9.4.1 Angular diameter distance to last scattering

In Chapter 3 we derived the formula for the comoving distance to redshift z,

$$d^c(z) = H_0^{-1} \int_{\frac{1}{1+z}}^{1} \frac{da}{\sqrt{\Omega_0(a-a^2) - \Omega_\Lambda(a-a^4) + a^2}} \tag{34}$$

(where we have approximated $\Omega_0 \approx \Omega_m + \Omega_\Lambda$) and the corresponding comoving angular diameter distance

$$d_A^c(z) = f_K\left(d^c(z)\right)\,, \tag{35}$$

where

$$f_K(x) \equiv \begin{cases} K^{-1/2}\sin(K^{1/2}x)\,, & K>0 \\ x\,, & K=0 \\ |K|^{-1/2}\sinh(|K|^{1/2}x)\,, & K<0\,. \end{cases} \tag{36}$$

We also define

$$f_k(x) \equiv \begin{cases} \sin x\,, & k=1 \\ x\,, & k=0 \\ \sinh x\,, & k=-1\,. \end{cases} \tag{37}$$

For the flat universe ($K=k=0$, $\Omega_0=1$), the comoving angular diameter distance is equal to the comoving distance,

$$d_A^c(z) = d^c(z) \qquad (K=0)\,. \tag{38}$$

For the open ($K<0$, $\Omega_0<1$) and closed ($K>0$, $\Omega_0>1$) cases we can write Eq. (35) as

$$\begin{aligned} d_A^c(z) &= \frac{H_0^{-1}}{\sqrt{|\Omega_k|}}\, f_k\left(\frac{\sqrt{|\Omega_k|}}{H_0^{-1}}\, d^c(z)\right) \\ &= H_0^{-1}\frac{1}{\sqrt{|\Omega_k|}}\, f_k\left(\sqrt{|\Omega_k|}\int_{\frac{1}{1+z}}^{1}\frac{da}{\sqrt{\Omega_0(a-a^2)-\Omega_\Lambda(a-a^4)+a^2}}\right)\,. \end{aligned} \tag{39}$$

Thus $d_A^c(z) \propto H_0^{-1}$, and has some more complicated dependence on Ω_0 and Ω_Λ (or on Ω_m and Ω_Λ).

We are now interested in the distance to the last scattering sphere, i.e., $d_A^c(z_{\rm dec})$, where $z_{\rm dec}\approx 1090$.

For the simplest case, $\Omega_\Lambda=0$, $\Omega_m=1$, the integral gives

$$d_A^c(z_{\rm dec}) = H_0^{-1}\int_{\frac{1}{1+z}}^{1}\frac{dx}{\sqrt{x}} = 2H_0^{-1}\left(1-\frac{1}{\sqrt{1+z_{\rm dec}}}\right) = 1.94H_0^{-1} \approx 2H_0^{-1}\,, \tag{40}$$

where the last approximation corresponds to ignoring the contribution from the lower limit.

We shall consider two more general cases, of which the above is a special case of both:

a) Open universe with no dark energy: $\Omega_\Lambda=0$ and $\Omega_m=\Omega_0<1$. Now the integral gives

$$\begin{aligned} d_A^c(z_{\rm dec}) &= \frac{H_0^{-1}}{\sqrt{1-\Omega_m}}\sinh\left(\sqrt{1-\Omega_m}\int_{\frac{1}{1+z}}^{1}\frac{dx}{\sqrt{(1-\Omega_m)x^2+\Omega_m x}}\right) \\ &= \frac{H_0^{-1}}{\sqrt{1-\Omega_m}}\sinh\left(\int_{\frac{1}{1+z}}^{1}\frac{dx}{\sqrt{x^2+\frac{\Omega_m}{1-\Omega_m}x}}\right) \\ &= \frac{H_0^{-1}}{\sqrt{1-\Omega_m}}\sinh\left(2\,\mathrm{arsinh}\sqrt{\frac{1-\Omega_m}{\Omega_m}}-2\,\mathrm{arsinh}\sqrt{\frac{1-\Omega_m}{\Omega_m}\frac{1}{1+z_{\rm dec}}}\right) \\ &\approx \frac{H_0^{-1}}{\sqrt{1-\Omega_m}}\sinh\left(2\,\mathrm{arsinh}\sqrt{\frac{1-\Omega_m}{\Omega_m}}\right) \;=\; 2\frac{H_0^{-1}}{\Omega_m}\,, \end{aligned} \tag{41}$$

where again the approximation ignores the contribution from the lower limit (i.e., it actually gives the angular diameter distance to the horizon, $d_A^c(z=\infty)$, in a model where we ignore the effect of other energy density components besides matter). In the last step we used $\sinh 2x = 2\sinh x\cosh x = 2\sinh x\sqrt{1+\sinh^2 x}$. We show this result (together with $d^c(z=\infty)$) in Fig. 17.

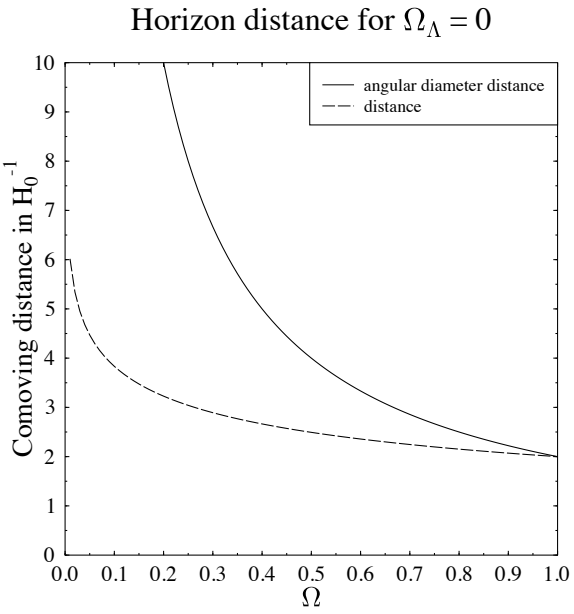


Figure 17: The comoving distance $d^c(z=\infty)$ (dashed) and the comoving angular diameter distance $d_A^c(z=\infty)$ (solid) to the horizon in matter-only open universe. The vertical axis is the distance in units of Hubble distance H_0^{-1} and the horizontal axis is the density parameter $\Omega_0 = \Omega_m$. The distances to last scattering, $d^c(z_{\rm dec})$ and $d_A^c(z_{\rm dec})$ are a few per cent less.

b) Flat universe with vacuum energy, $\Omega_\Lambda + \Omega_m = 1$. Here the integral does not give an elementary function, but a reasonable approximation, which we shall use in the following, is

$$d_A^c(z_{\rm dec}) = d^c(z_{\rm dec}) \approx \frac{2}{\Omega_m^{0.4}} H_0^{-1}\,. \tag{42}$$

The comoving distance $d_c(z_{\rm dec})$ depends on the expansion history of the universe. The longer it takes for the universe to cool from $T_{\rm dec}$ to T_0 (i.e., to expand by the factor $1+z_{\rm dec}$), the longer distance the photons have time to travel. When a larger part of this time is spent at small values of the scale factor, this distance gets a bigger boost from converting it to a comoving distance. For open/closed universes the angular diameter distance gets an additional effect from the geometry of the universe (the f_K), which acts like a "lens" to make the distant CMB pattern at the last scattering sphere to look smaller or larger (see Fig. 18).

9.4.2 Hubble scale and the matter-radiation equality scale

Subhorizon ($k \gg \mathcal{H}$) and superhorizon ($k \ll \mathcal{H}$) scales behave differently. Thus we want to know which of the structures we see on the last scattering surface are subhorizon and which are superhorizon. For that we need to know the comoving Hubble scale $\mathcal{H}$ at $t_{\rm dec}$. This was discussed in Sec. 8.3.1. At that time both matter and radiation are contributing to the energy density and the Hubble parameter. The scale which is just entering at $t = t_{\rm dec}$ is

$$\begin{aligned} k_{\rm dec}^{-1} &\equiv \mathcal{H}_{\rm dec}^{-1} = (1+z_{\rm dec})H_{\rm dec}^{-1} = (1+z_{\rm dec})^{-1/2} H_0^{-1}\Omega_m^{-1/2}\left[1+\frac{\Omega_r}{\Omega_m}(1+z_{\rm dec})\right]^{-1/2} \\ &= \Omega_m^{-1/2}(1+0.046\,\omega_m^{-1})^{-1/2}\,91\,h^{-1}{\rm Mpc} \end{aligned} \tag{43}$$

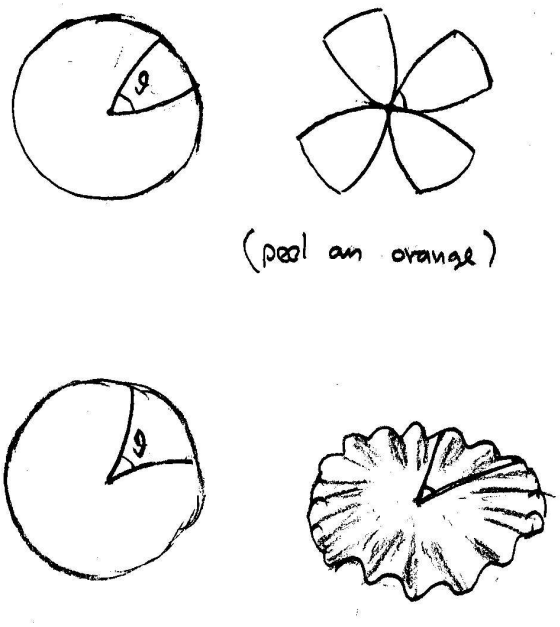


Figure 18: The geometry effect in a closed (top) or an open (bottom) universe affects the angle at which we see a structure of given size at the last scattering surface, and thus its angular diameter distance.

(using $z_{\rm dec} = 1090$; here $0.046\,\omega_m^{-1}$ is ρ_r/ρ_m at $t_{\rm dec}$) and the corresponding multipole number on the last scattering sphere is

$$\begin{aligned} \ell_H &\equiv k_{\rm dec} d_A^c \qquad (44) \\ &= (1+z_{\rm dec})^{1/2}\Omega_m^{-1/2}\left[1+\frac{\Omega_r}{\Omega_m}(1+z_{\rm dec})\right]^{1/2} \times \begin{cases} 2/\Omega_m &= 66\,\Omega_m^{-0.5}\sqrt{1+0.046\omega_m^{-1}} \quad (\Omega_\Lambda = 0) \\ 2/\Omega_m^{0.4} &\approx 66\,\Omega_m^{0.1}\sqrt{1+0.046\omega_m^{-1}} \quad (\Omega_0 = 1) \end{cases} \end{aligned}$$

The angle subtended by a half-wavelength π/k of this mode on the last scattering sphere is

$$\vartheta_H \equiv \frac{\pi}{\ell_H} = \frac{180^\circ}{\ell_H} = \sqrt{1+0.046\,\omega_m^{-1}} \times \begin{cases} 2.7^\circ\Omega_m^{0.5} \\ 2.7^\circ\Omega_m^{-0.1} \end{cases} \qquad (45)$$

For $\Omega_m \sim 0.3$, $\Omega_\Lambda \sim 0.7$, $h \sim 0.7$, $\ell_H \approx 67$ and $\vartheta_H \approx 3.5^\circ$.

Another important scale is $k_{\rm eq}$, the scale which enters at the time of matter-radiation equality $t_{\rm eq}$, since the transfer function $T(k)$ is bent at that point. Perturbations for scales $k \ll k_{\rm eq}$ maintain essentially their primordial spectrum, whereas scales $k \gg k_{\rm eq}$ have lost relative power between their horizon entry and $t_{\rm eq}$. This scale is

$$k_{\rm eq}^{-1} = \mathcal{H}_{\rm eq}^{-1} \sim 13.7\,\Omega_m^{-1}h^{-2}\,{\rm Mpc} = 4.6\times10^{-3}\,\Omega_m^{-1}h^{-1}H_0^{-1} \qquad (46)$$

and the corresponding multipole number of these scales seen on the last scattering sphere is

$$l_{\rm eq} = k_{\rm eq}d_A^c = 219\,\Omega_m h \times \begin{cases} 2/\Omega_m &= 440\,h \quad (\Omega_\Lambda = 0) \\ 2/\Omega_m^{0.4} &\approx 440\,h\,\Omega_m^{0.6} \quad (\Omega_0 = 1) \end{cases} \qquad (47)$$

9.5 CMB anisotropy from perturbation theory

We began this chapter with the observation, Eq. (1), that the CMB temperature anisotropy is a sum of two parts,

$$\left(\frac{\delta T}{T}\right)_{\rm obs} = \left(\frac{\delta T}{T}\right)_{\rm intr} + \left(\frac{\delta T}{T}\right)_{\rm jour}, \qquad (48)$$

and that this separation is gauge dependent. We shall consider this in the conformal-Newtonian gauge, since the second part, $\left(\frac{\delta T}{T}\right)_{\rm jour}$, the integrated redshift perturbation along the line of sight, is easiest to calculate in this gauge. (However, we won't do the calculation here.[5])

The result of this calculation is

$$\begin{aligned}\left(\frac{\delta T}{T}\right)_{\rm jour} &= -\int d\Phi + \int (\dot\Phi + \dot\Psi) dt + \mathbf{v}_{\rm obs}\cdot\hat{\mathbf{n}} \\ &= \Phi(t_{\rm dec},\mathbf{x}_{\rm ls}) - \Phi(t_0,\mathbf{0}) + \int (\dot\Phi + \dot\Psi) dt + \mathbf{v}_{\rm obs}\cdot\hat{\mathbf{n}} \\ &\overset{\Psi\approx\Phi}{\approx} \Phi(t_{\rm dec},\mathbf{x}_{\rm ls}) - \Phi(t_0,\mathbf{0}) + 2\int \dot\Phi dt + \mathbf{v}_{\rm obs}\cdot\hat{\mathbf{n}} \end{aligned} \tag{49}$$

where the integral is from $(t_{\rm dec},\mathbf{x}_{\rm ls})$ to $(t_0,\mathbf{0})$ along the path of the photon (a null geodesic). The origin $\mathbf{0}$ is located where the observer is. The last term, $\mathbf{v}_{\rm obs}\cdot\hat{\mathbf{n}}$, is the Doppler effect from observer motion (assumed nonrelativistic), $\mathbf{v}_{\rm obs}$ being the observer velocity and $\hat{\mathbf{n}}$ the direction we are looking at. The $_{\rm ls}$ in $\mathbf{x}_{\rm ls}$ is just to remind us that $\mathbf{x}$ lies somewhere on the last scattering sphere. In the matter-dominated universe the Newtonian potential remains constant in time, $\dot\Phi = 0$, so we get a contribution from the integral only from epochs when radiation or dark energy contributions to the total energy density, or the effect of curvature, cannot be ignored. We can understand the above result as follows. If the potential is constant in time, the blueshift the photon acquires when falling into a potential well is canceled by the redshift from climbing up the well. Thus the net redshift/blueshift caused by gravitational potential perturbations is just the difference between the values of Φ at the beginning and in the end. However, if the potential is changing while the photon is traversing the well, this cancelation is not exact, and we get the integral term to account for this effect.

The value of the potential perturbation at the observing site, $\Phi(t_0,\mathbf{0})$ is the same for photons coming from all directions. Thus it does not contribute to the observed anisotropy. It just produces an overall shift in the observed average temperature. This is included in the observed value $T_0 = 2.725\,{\rm K}$, and there is no way for us to separate it from the "correct" unperturbed value. Thus we can ignore it. The observer motion $\mathbf{v}_{\rm obs}$ causes a dipole ($\ell = 1$) pattern in the CMB anisotropy, and likewise, there is no way for us to separate from it the cosmological dipole on the last scattering sphere. Therefore the dipole is usually removed from the CMB map before analyzing it for cosmological purposes. Accordingly, we shall ignore this term also, and our final result is

$$\left(\frac{\delta T}{T}\right)_{\rm jour} = \Phi(t_{\rm dec},\mathbf{x}_{\rm ls}) + 2\int \dot\Phi dt\,. \tag{50}$$

The other part, $\left(\frac{\delta T}{T}\right)_{\rm intr}$, comes from the local temperature perturbation at $t = t_{\rm dec}$ and the Doppler effect, $-\mathbf{v}\cdot\hat{\mathbf{n}}$, from the local (baryon+photon) fluid motion at that time. Since

$$\rho_\gamma = \frac{\pi^2}{15}T^4\,, \tag{51}$$

the local temperature perturbation is directly related to the relative perturbation in the photon energy density,

$$\left(\frac{\delta T}{T}\right)_{\rm intr} = \frac{1}{4}\delta_\gamma - \mathbf{v}\cdot\hat{\mathbf{n}}\,. \tag{52}$$

We can now write the observed temperature anisotropy as

$$\boxed{\left(\frac{\delta T}{T}\right)_{\rm obs} = \frac{1}{4}\delta_\gamma^N - \mathbf{v}^N\cdot\hat{\mathbf{n}} + \Phi(t_{\rm dec},\mathbf{x}_{\rm ls}) + 2\int \dot\Phi dt\,.} \tag{53}$$

[5] It is done in my course on Cosmological Perturbation Theory, Sec. 25.

(note that both the density perturbation δ_γ and the fluid velocity $\mathbf{v}$ are gauge dependent).

To make further progress we now

1. consider adiabatic primordial perturbations only (like we did in Chapter 8), and
2. make the (crude) approximation that the universe is already matter dominated at $t = t_{\rm dec}$.

For adiabatic perturbations

$$\delta_b = \delta_c \equiv \delta_m = \frac{3}{4}\delta_\gamma \,. \tag{54}$$

The perturbations stay adiabatic only at superhorizon scales. Once the perturbation has entered horizon, different physics can begin to act on different matter components, so that the adiabatic relation between their density perturbations is broken. In particular, the baryon+photon perturbation is affected by photon pressure, which will damp their growth and cause them to oscillate, whereas the CDM perturbation is unaffected and keeps growing. Since the baryon and photon components see the same pressure they still evolve together and maintain their adiabatic relation until photon decoupling. Thus, after horizon entry, but before decoupling,

$$\delta_c \neq \delta_b = \frac{3}{4}\delta_\gamma \,. \tag{55}$$

At decoupling, the equality holds for scales larger than the photon mean free path at $t_{\rm dec}$.

After decoupling, this connection between the photons and baryons is broken, and the baryon density perturbation begins to approach the CDM density perturbation,

$$\delta_c \leftarrow \delta_b \neq \frac{3}{4}\delta_\gamma \,. \tag{56}$$

We shall return to these issues as we discuss the shorter scales in Sections 9.7 and 9.8. But let us first discuss the scales which are still superhorizon at $t_{\rm dec}$, so that Eq. (54) still applies.

9.6 Large scales: Sachs–Wolfe part of the spectrum

Consider now the scales $k \ll k_{\rm dec}$, or $\ell \ll \ell_H$, which are still superhorizon at decoupling. We can now use the adiabatic condition (54), so that

$$\frac{1}{4}\delta_\gamma = \frac{1}{3}\delta_m \approx \frac{1}{3}\delta \,, \tag{57}$$

where the latter (approximate) equality comes from taking the universe to be matter dominated at $t_{\rm dec}$, so that we can identify $\delta \approx \delta_m$. For these scales the Doppler effect from fluid motion is subdominant, and we can ignore it (the fluid is set into motion by gradients in the pressure and gravitational potential, but the timescale of getting into motion is longer than the Hubble time for superhorizon scale gradients).

Thus Eq. (53) becomes

$$\left(\frac{\delta T}{T}\right)_{\rm obs} = \frac{1}{3}\delta^N + \Phi(t_{\rm dec}, \mathbf{x}_{\rm ls}) + 2\int \dot{\Phi} dt \,. \tag{58}$$

The Newtonian relation

$$\delta = \frac{1}{4\pi G\bar{\rho}a^2}\nabla^2\Phi = \frac{2}{3}\left(\frac{1}{aH}\right)^2 \nabla^2\Phi$$

(here ∇ is with respect to the comoving coordinates, hence the a^{-2}) or

$$\delta_{\mathbf{k}} = -\frac{2}{3}\left(\frac{k}{\mathcal{H}}\right)^2 \Phi_{\mathbf{k}}$$

does not hold at superhorizon scales (where δ is gauge dependent). A GR calculation using the Newtonian gauge gives the result[6]

$$\boxed{\delta^N_{\mathbf{k}} \;=\; -\left[2+\frac{2}{3}\left(\frac{k}{\mathcal{H}}\right)^2\right]\Phi_{\mathbf{k}}} \tag{59}$$

for perturbations in a matter-dominated universe. Thus for superhorizon scales we can approximate

$$\delta^N \approx -2\Phi \tag{60}$$

and Eq. (58) becomes

$$\begin{aligned}\left(\frac{\delta T}{T}\right)_{\rm obs} &= -\frac{2}{3}\Phi(t_{\rm dec},\mathbf{x}_{\rm ls})+\Phi(t_{\rm dec},\mathbf{x}_{\rm ls})+2\int\dot\Phi dt \\ &= \frac{1}{3}\Phi(t_{\rm dec},\mathbf{x}_{\rm ls})+2\int\dot\Phi dt\,.\end{aligned} \tag{61}$$

This explains the "mysterious" factor 1/3 in this relation between the potential Φ and the temperature perturbation.

This result is called the *Sachs–Wolfe effect*. The first part, $\frac{1}{3}\Phi(t_{\rm dec},\mathbf{x}_{\rm ls})$, is called the *ordinary* Sachs–Wolfe effect, and the second part, $2\int\dot\Phi dt$, the *integrated Sachs-Wolfe effect* (ISW), since it involves integrating along the line of sight. Note that the approximation of matter domination at $t=t_{\rm dec}$, making $\dot\Phi=0$, does not eliminate the ISW, since it only applies to the "early part" of the integral. At times closer to t_0, dark energy becomes important, causing Φ to evolve again. This ISW caused by dark energy (or curvature of the background universe, if $k\neq 0$) is called the *late Sachs–Wolfe effect* (LSW) and it shows up as a rise in the smallest ℓ of the angular power spectrum C_ℓ. Correspondingly, the contribution to the ISW from the evolution of Φ near $t_{\rm dec}$ due to the radiation contribution to the expansion law (which we ignored in our approximation) is called the *early Sachs–Wolfe effect* (ESW). The ESW shows up as a rise in C_ℓ for larger ℓ, near ℓ_H.

We shall now forget for a while the ISW, which for $\ell\ll\ell_H$ is expected to be smaller than the ordinary Sachs–Wolfe effect.

9.6.1 Angular power spectrum from the ordinary Sachs–Wolfe effect

We now calculate the contribution from the ordinary Sachs–Wolfe effect,

$$\left(\frac{\delta T}{T}\right)_{\rm SW} \;=\; \frac{1}{3}\Phi(t_{\rm dec},\mathbf{x}_{\rm ls})\,, \tag{62}$$

to the angular power spectrum C_ℓ. This is the dominant effect for $\ell\ll\ell_H$.

Since Φ is evaluated at the last scattering sphere, we have, from Eq. (30),

$$a_{\ell m} \;=\; 4\pi i^\ell\sum_{\mathbf{k}}\frac{1}{3}\Phi_{\mathbf{k}}j_\ell(kx)Y^*_{\ell m}(\hat{\mathbf{k}})\,, \tag{63}$$

In the matter-dominated epoch,

$$\Phi=-\frac{3}{5}\mathcal{R}\,, \tag{64}$$

so that

$$a_{\ell m}=-\frac{4\pi}{5}i^\ell\sum_{\mathbf{k}}\mathcal{R}_{\mathbf{k}}j_\ell(kx)Y^*_{\ell m}(\hat{\mathbf{k}})\,. \tag{65}$$

[6]Cosmological Perturbation Theory, Sec. 13.

The coefficient $a_{\ell m}$ is thus a linear combination of the independent random variables $\mathcal{R}_{\mathbf{k}}$, i.e., it is of the form

$$\sum_{\mathbf{k}} b_{\mathbf{k}} \mathcal{R}_{\mathbf{k}} \,, \tag{66}$$

For any such linear combination, the expectation value of its absolute value squared is

$$\begin{aligned} \left\langle \left| \sum_{\mathbf{k}} b_{\mathbf{k}} \mathcal{R}_{\mathbf{k}} \right|^2 \right\rangle &= \sum_{\mathbf{k}} \sum_{\mathbf{k}'} b_{\mathbf{k}} b^*_{\mathbf{k}'} \left\langle \mathcal{R}_{\mathbf{k}} \mathcal{R}^*_{\mathbf{k}'} \right\rangle \\ &= \left(\frac{2\pi}{L} \right)^3 \sum_{\mathbf{k}} \frac{1}{4\pi k^3} \mathcal{P}_{\mathcal{R}}(k) \left| b_{\mathbf{k}} \right|^2 \,, \end{aligned} \tag{67}$$

where we used

$$\left\langle \mathcal{R}_{\mathbf{k}} \mathcal{R}^*_{\mathbf{k}'} \right\rangle = \delta_{\mathbf{k}\mathbf{k}'} \left(\frac{2\pi}{L} \right)^3 \frac{1}{4\pi k^3} \mathcal{P}_{\mathcal{R}}(k) \tag{68}$$

(the independence of the random variables $\mathcal{R}_{\mathbf{k}}$ and the definition of the power spectrum $\mathcal{P}(k)$).

Thus

$$\begin{aligned} C_\ell &\equiv \frac{1}{2\ell+1} \sum_m \langle |a_{\ell m}|^2 \rangle \\ &= \frac{16\pi^2}{25} \frac{1}{2\ell+1} \sum_m \left(\frac{2\pi}{L} \right)^3 \sum_{\mathbf{k}} \frac{1}{4\pi k^3} \mathcal{P}_{\mathcal{R}}(k) j_\ell(kx)^2 \left| Y^*_{\ell m}(\hat{\mathbf{k}}) \right|^2 \\ &= \frac{1}{25} \left(\frac{2\pi}{L} \right)^3 \sum_{\mathbf{k}} \frac{1}{k^3} \mathcal{P}_{\mathcal{R}}(k) j_\ell(kx)^2 \,. \end{aligned} \tag{69}$$

(Although all $\langle |a_{\ell m}|^2 \rangle$ are equal for the same ℓ, we used the sum over m, so that we could use Eq. (9).) Replacing the sum with an integral, we get

$$\begin{aligned} C_\ell &= \frac{1}{25} \int \frac{d^3k}{k^3} \mathcal{P}_{\mathcal{R}}(k) j_\ell(kx)^2 \\ &= \frac{4\pi}{25} \int_0^\infty \frac{dk}{k} \mathcal{P}_{\mathcal{R}}(k) j_\ell(kx)^2 \,, \end{aligned} \tag{70}$$

the final result for an arbitrary primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$.

The integral can be done for a power-law power spectrum, $\mathcal{P}_{\mathcal{R}}(k) = A_s^2 k^{n-1}$. In particular, for a scale-invariant ($n = 1$) primordial power spectrum,

$$\mathcal{P}_{\mathcal{R}}(k) = \text{const.} = A_s^2 \,, \tag{71}$$

we have

$$C_\ell = A_s^2 \frac{4\pi}{25} \int_0^\infty \frac{dk}{k} j_\ell(kx)^2 = \frac{A_s^2}{25} \frac{2\pi}{\ell(\ell+1)} \,, \tag{72}$$

since

$$\int_0^\infty \frac{dk}{k} j_\ell(kx)^2 = \frac{1}{2\ell(\ell+1)} \,. \tag{73}$$

We can write this as

$$\frac{\ell(\ell+1)}{2\pi} C_\ell = \frac{A_s^2}{25} = \text{const. (independent of } \ell) \tag{74}$$

This is the reason why the angular power spectrum is customarily plotted as $\ell(\ell+1)C_\ell/2\pi$; it makes the ordinary Sachs–Wolfe part of the C_ℓ flat for a scale-invariant primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$.

Present data is consistent with an almost scale-invariant primordial power spectrum (actually it favors a small red tilt, $n < 1$). The constant A_s can be determined from the ordinary Sachs–Wolfe part of the observed $\widehat{C}_\ell$. From Fig. 11 we see that at low ℓ

$$\frac{\ell(\ell+1)}{2\pi}\widehat{C}_\ell \;\sim\; \frac{800\,\mu\mathrm{K}^2}{(2.725\,K)^2} \;\sim\; 10^{-10} \tag{75}$$

on the average. This gives the amplitude of the primordial power spectrum as

$$\mathcal{P}_\mathcal{R}(k) = A_s^2 \sim 25\times 10^{-10} = \left(5\times 10^{-5}\right)^2 . \tag{76}$$

We already used this result in Chapter 8 as a constraint on the energy scale of inflation.

Exercise: Find the C_ℓ of the ordinary Sachs-Wolfe effect due to a power-law power spectrum $\mathcal{P}_\mathcal{R}(k) = A_s^2 k^{n-1}$. Help:

$$\int_0^\infty dx\, x^{n-2} j_\ell^2(x) \;=\; 2^{n-4}\pi \frac{\Gamma(\ell+\frac{n}{2}-\frac{1}{2})\Gamma(3-n)}{\Gamma(\ell+\frac{5}{2}-\frac{n}{2})\Gamma(2-\frac{n}{2})^2}\,. \tag{77}$$

Take $A_s = 4.62\times 10^{-5}$ and $n = 0.968$ (Planck 2015 central values). Give the numerical values for C_2 and C_{20}.

9.7 Acoustic oscillations

Consider now the scales $k \gg k_{\rm dec}$, or $\ell \gg \ell_H$, which are subhorizon at decoupling. The observed temperature anisotropy is, from Eq. (53)

$$\left(\frac{\delta T}{T}\right)_{\rm obs} = \frac{1}{4}\delta_\gamma(t_{\rm dec}, \mathbf{x}_{\rm ls}) + \Phi(t_{\rm dec}, \mathbf{x}_{\rm ls}) - \mathbf{v}_\gamma \cdot \hat{\mathbf{n}}(t_{\rm dec}, \mathbf{x}_{\rm ls}) + 2\int \dot{\Phi} dt\,. \tag{78}$$

Since we are considering subhorizon scales, we dropped the reference to the Newtonian gauge. We shall concentrate on the three first terms, which correspond to the situation at the point $(t_{\rm dec}, \mathbf{x}_{\rm ls})$ we are looking at on the last scattering sphere.

Before decoupling photons are coupled to baryons. Perturbations in the baryon-photon fluid are oscillating, whereas CDM perturbations grow (slowly during the radiation-dominated epoch, and then faster during the matter-dominated epoch). Therefore CDM perturbations begin to dominate the total density perturbation $\delta\rho$ and thus also Φ already before the universe becomes matter dominated, and CDM begins to dominate the background energy density. Thus we make the approximation that Φ is given by the CDM perturbation. The baryon-photon fluid oscillates in these potential wells caused by the CDM. The potential Φ evolves at first but then becomes constant as the universe becomes matter dominated.

We shall not attempt an exact calculation of the $\delta_{b\gamma}$ oscillations in the expanding universe. One reason is that $\rho_{b\gamma}$ is a relativistic fluid, and we have derived the perturbation equations for a nonrelativistic fluid only. From Sec. 8.2.7 we have that the nonrelativistic perturbation equation for a fluid component i is

$$\ddot{\delta}_{\mathbf{k}i} + 2H\dot{\delta}_{\mathbf{k}i} = -\frac{k^2}{a^2}\left(\frac{\delta p_{\mathbf{k}i}}{\bar{\rho}_i} + \Phi_{\mathbf{k}}\right). \tag{79}$$

The generalization of the (subhorizon) perturbation equations to the case of a relativistic fluid is considerably easier if we ignore the expansion of the universe. Then Eq. (79) becomes

$$\ddot{\delta}_{\mathbf{k}i} + k^2\left(\frac{\delta p_{\mathbf{k}i}}{\bar{\rho}_i} + \Phi_{\mathbf{k}}\right) = 0\,. \tag{80}$$

According to GR, the density of "passive gravitational mass" is $\rho + p = (1+w)\rho$, not just ρ as in Newtonian gravity. Therefore the force on a fluid element of the fluid component i is proportional to $(\rho_i + p_i)\nabla\Phi = (1+w_i)\rho_i\nabla\Phi$ instead of just $\rho_i\nabla\Phi$, and Eq. (80) generalizes to the case of a relativistic fluid as[7]

$$\ddot{\delta}_{\mathbf{k}i} + k^2\left[\frac{\delta p_{\mathbf{k}i}}{\bar{\rho}_i} + (1+w_i)\Phi_{\mathbf{k}}\right] = 0\,. \tag{81}$$

In the present application the fluid component ρ_i is the baryon-photon fluid $\rho_{b\gamma}$ and the gravitational potential Φ is caused by the CDM. Before decoupling, the adiabatic relation $\delta_b = \frac{3}{4}\delta_\gamma$ still holds between photons and baryons, and we have the adiabatic relation between pressure and density perturbations,

$$\delta p_{b\gamma} = c_s^2 \delta\rho_{b\gamma} \tag{82}$$

where

$$c_s^2 = \frac{\delta p_{b\gamma}}{\delta\rho_{b\gamma}} \approx \frac{\delta p_\gamma}{\delta\rho_{b\gamma}} = \frac{1}{3}\frac{\delta\rho_\gamma}{\delta\rho_\gamma + \delta\rho_b} = \frac{1}{3}\frac{\bar{\rho}_\gamma\delta_\gamma}{\bar{\rho}_\gamma\delta_\gamma + \bar{\rho}_b\delta_b} = \frac{1}{3}\frac{1}{1+\frac{3}{4}\frac{\bar{\rho}_b}{\bar{\rho}_\gamma}} \equiv \frac{1}{3}\frac{1}{1+R} \tag{83}$$

[7] Actually the derivation is more complicated, since also the density of "inertial mass" is $\rho_i + p_i$ and the energy continuity equation is modified by a work-done-by-pressure term. The more detailed derivation of Eq. (81) was given in Sec. 8.2.8.

gives the speed of sound c_s of the baryon-photon fluid. We defined

$$R \equiv \frac{3}{4}\frac{\bar{\rho}_b}{\bar{\rho}_\gamma}\,. \tag{84}$$

We can now write the perturbation equation (81) for the baryon-photon fluid as

$$\ddot{\delta}_{b\gamma\mathbf{k}} + k^2\left[c_s^2\delta_{b\gamma\mathbf{k}} + (1+w_{b\gamma})\Phi_\mathbf{k}\right] = 0\,. \tag{85}$$

The equation-of-state parameter for the baryon-photon fluid is

$$w_{b\gamma} \equiv \frac{\bar{p}_{b\gamma}}{\bar{\rho}_{b\gamma}} = \frac{\frac{1}{3}\bar{\rho}_\gamma}{\bar{\rho}_\gamma + \bar{\rho}_b} = \frac{1}{3}\frac{1}{1+\frac{4}{3}R}\,, \tag{86}$$

so that

$$1 + w_{b\gamma} = \frac{\frac{4}{3}(1+R)}{1+\frac{4}{3}R} \tag{87}$$

and we can write Eq. (85) as

$$\ddot{\delta}_{b\gamma\mathbf{k}} + k^2\left[\frac{1}{3}\frac{1}{1+R}\delta_{b\gamma\mathbf{k}} + \frac{\frac{4}{3}(1+R)}{1+\frac{4}{3}R}\Phi_\mathbf{k}\right] = 0\,. \tag{88}$$

For the CMB anisotropy we are interested in[8]

$$\Theta_0 \equiv \frac{1}{4}\delta_\gamma\,, \tag{89}$$

which gives the local temperature perturbation, not in $\delta_{b\gamma}$. These two are related by

$$\delta_{b\gamma} = \frac{\delta\rho_{b\gamma}}{\bar{\rho}_{b\gamma}} = \frac{\delta\rho_\gamma + \delta\rho_b}{\bar{\rho}_\gamma + \bar{\rho}_b} = \frac{\bar{\rho}_\gamma\delta_\gamma + \bar{\rho}_b\delta_b}{\bar{\rho}_\gamma + \bar{\rho}_b} = \frac{1+R}{1+\frac{4}{3}R}\delta_\gamma\,. \tag{90}$$

Thus we can write Eq. (85) as

$$\ddot{\delta}_{\gamma\mathbf{k}} + k^2\left[\frac{1}{3}\frac{1}{1+R}\delta_{\gamma\mathbf{k}} + \frac{4}{3}\Phi_\mathbf{k}\right] = 0\,, \tag{91}$$

or

$$\ddot{\Theta}_{0\mathbf{k}} + k^2\left[\frac{1}{3}\frac{1}{1+R}\Theta_{0\mathbf{k}} + \frac{1}{3}\Phi_\mathbf{k}\right] = 0\,, \tag{92}$$

or

$$\ddot{\Theta}_{0\mathbf{k}} + c_s^2k^2\left[\Theta_{0\mathbf{k}} + (1+R)\Phi_\mathbf{k}\right] = 0\,, \tag{93}$$

If we now take R and Φ_k to be constant, this is the harmonic oscillator equation for the quantity $\Theta_{0\mathbf{k}} + (1+R)\Phi_\mathbf{k}$ with the general solution

$$\Theta_{0\mathbf{k}} + (1+R)\Phi_\mathbf{k} = A_\mathbf{k}\cos c_skt + B_\mathbf{k}\sin c_skt\,, \tag{94}$$

or

$$\Theta_{0\mathbf{k}} + \Phi_\mathbf{k} = -R\Phi_\mathbf{k} + A_\mathbf{k}\cos c_skt + B_\mathbf{k}\sin c_skt\,, \tag{95}$$

or

$$\Theta_{0\mathbf{k}} = -(1+R)\Phi_\mathbf{k} + A_\mathbf{k}\cos c_skt + B_\mathbf{k}\sin c_skt\,. \tag{96}$$

[8]The subscript 0 refers to the monopole ($\ell = 0$) of the *local* photon distribution. Likewise, the dipole ($\ell = 1$) of the local photon distribution corresponds to the velocity of the photon fluid, $\Theta_1 \equiv v_\gamma/3$.

We are interested in the quantity $\Theta_0+\Phi = \frac{1}{4}\delta_\gamma+\Phi$, called the *effective temperature perturbation*, since this combination appears in Eq. (78). It is the local temperature perturbation minus the redshift photons suffer when climbing from the potential well of the perturbation (negative Φ for a CDM overdensity). We see that this quantity oscillates in time, and the effect of baryons (via R) is to shift the equilibrium point of the oscillation by $-R\Phi_{\mathbf{k}}$.

In the preceding we ignored the effect of the expansion of the universe. The expansion affects the preceding in a number of ways. For example, c_s, $w_{b\gamma}$ and R change with time. The potential Φ also evolves, especially at the earlier times when radiation dominates the expansion law. However, the qualitative result of an oscillation of $\Theta_0 + \Phi$, and the shift of its equilibrium point by baryons, remains. The time t in the solution (95) gets replaced by conformal time η, and since c_s changes with time, $c_s\eta$ is replaced by

$$r_s(t) \equiv \int_0^\eta c_s d\eta = \int_0^t \frac{c_s(t)}{a(t)} dt\,. \tag{97}$$

We call this quantity $r_s(t)$ the *sound horizon* at time t, since it represents the comoving distance sound has traveled by time t.

The relative weight of the cosine and sine solutions (i.e., the constants $A_{\mathbf{k}}$ and $B_{\mathbf{k}}$ in Eq. (94) depends on the initial conditions. Since the perturbations are initially at superhorizon scales, the initial conditions are determined there, and the present discussion does not really apply. However, using the Newtonian gauge superhorizon initial conditions gives the correct qualitative result for the phase of the oscillation.

We had that for adiabatic primordial perturbations, initially $\Phi = -\frac{3}{5}\mathcal{R}$ and $\frac{1}{4}\delta_\gamma^N = -\frac{2}{3}\Phi = \frac{2}{5}\mathcal{R}$, giving us an initial condition $\Theta_0 + \Phi = \frac{1}{3}\Phi = -\frac{1}{5}\mathcal{R} = \text{const}$. (At these early times $R \ll 1$, so we don't write the $1 + R$.) Thus adiabatic primordial perturbations correspond essentially to the cosine solution. (There are effects at the horizon scale which affect the amplitude of the oscillations—the main effect being the decay of Φ as it enters the horizon—so we can't use the preceding discussion to determine the amplitude, but we get the right result about the initial phase of the $\Theta_0 + \Phi$ oscillations.)

Thus we have that, qualitatively, the effective temperature behaves at subhorizon scales as

$$\Theta_{0\mathbf{k}} + (1 + R)\Phi_{\mathbf{k}} \;\propto\; \cos kr_s(t)\,, \tag{98}$$

Consider a region which corresponds to a positive primordial curvature perturbation $\mathcal{R}$. It begins with an initial overdensity (of all components, photons, baryons, CDM and neutrinos), and a negative gravitational potential Φ. For the scales of interest for CMB anisotropy, the potential stays negative, since the CDM begins to dominate the potential early enough and the CDM perturbations do not oscillate, they just grow. The effective temperature perturbation $\Theta_0 + \Phi$, which is the oscillating quantity, begins with a negative value. After half an oscillation period it is at its positive extreme value. This increase of $\Theta_0+\Phi$ corresponds to an increase in δ_γ; from its initial positive value it has grown to a larger positive value. Thus the oscillation begins by the, already initially overdense, baryon-photon fluid falling deeper into the potential well, and reaching its maximum compression after half a period. After this maximum compression the photon pressure pushes the baryon-photon fluid out from the potential well, and after a full period, the fluid reaches its maximum decompression in the potential well. Since the potential Φ has meanwhile decayed (horizon entry and the resulting potential decay always happens during the first oscillation period, since the sound horizon and the Hubble length are close to each other, as the sound speed is close to the speed of light), the decompression does not bring the $\delta_{b\gamma}$ back to its initial value (which was overdense), but the photon-baryon fluid actually becomes underdense in the potential well (and overdense in the neighboring potential "hill"). And so the oscillation goes on until photon decoupling.

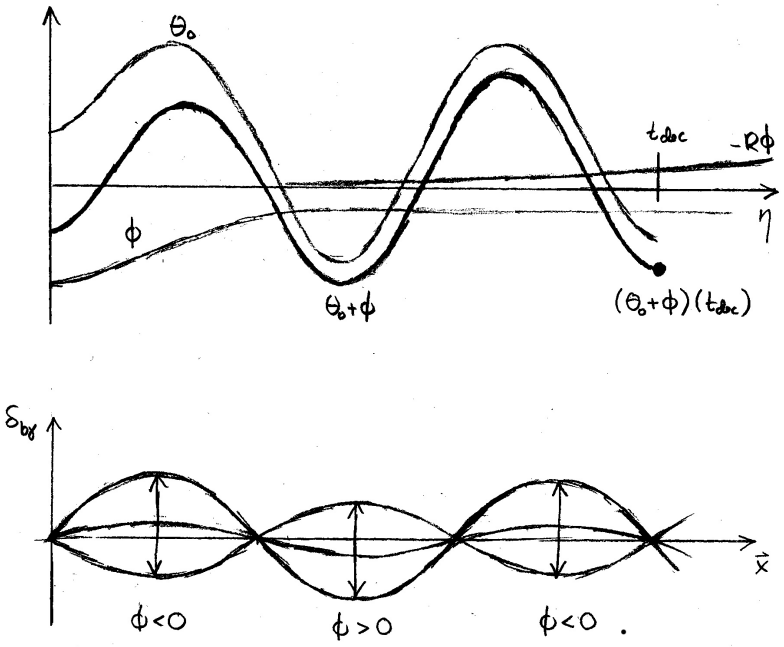


Figure 19: Acoustic oscillations. The top panel shows the time evolution of the Fourier amplitudes $\Theta_{0\mathbf{k}}$, $\Phi_{\mathbf{k}}$, and the effective temperature $\Theta_{0\mathbf{k}} + \Phi_{\mathbf{k}}$. The Fourier mode shown corresponds to the fourth acoustic peak of the C_ℓ spectrum. The bottom panel shows $\delta_{b\gamma}(\mathbf{x})$ for one Fourier mode as a function of position at various times (maximum compression, equilibrium level, and maximum decompression).

These are standing waves and they are called *acoustic oscillations*. See Fig. 19. Because of the potential decay at horizon entry, the amplitude of the oscillation is larger than Φ, and thus also Θ_0 changes sign in the oscillation.

These oscillations end at photon decoupling, when the photons are liberated. The CMB shows these standing waves as a snapshot[9] at their final moment $t = t_{\rm dec}$.

At photon decoupling we have

$$\Theta_{0\mathbf{k}} + (1+R)\Phi_{\mathbf{k}} \propto \cos kr_s(t_{\rm dec})\,. \tag{99}$$

At this moment oscillations for scales k which have

$$kr_s(t_{\rm dec}) = m\pi \tag{100}$$

($m = 1, 2, 3, \ldots$) are at their extreme values (maximum compression or maximum decompression). Therefore we see strong structure in the CMB anisotropy at the multipoles

$$\ell = kd_A^c(t_{\rm dec}) = m\pi\frac{d_A^c(t_{\rm dec})}{r_s(t_{\rm dec})} \equiv m\ell_A \tag{101}$$

corresponding to these scales. Here

$$\ell_A \equiv \pi\frac{d_A^c(t_{\rm dec})}{r_s(t_{\rm dec})} \equiv \frac{\pi}{\vartheta_s} \tag{102}$$

is the *acoustic scale* in multipole space and

$$\vartheta_s \equiv \frac{r_s(t_{\rm dec})}{d_A^c(t_{\rm dec})} \tag{103}$$

[9]Actually, photon decoupling takes quite a long time. Therefore this "snapshot" has a rather long "exposure time" causing it to be "blurred". This prevents us from seeing very small scales in the CMB anisotropy.

is the *sound horizon angle*, i.e., the angle at which we see the sound horizon on the last scattering surface.

Because of these acoustic oscillations, the CMB angular power spectrum C_ℓ has a structure of *acoustic peaks* at subhorizon scales. The centers of these peaks are located approximately at $\ell_m \approx m\ell_A$. An exact calculation shows that they will actually lie at somewhat smaller ℓ due to a number of effects. The separation of neighboring peaks is closer to ℓ_A than the positions of the peaks are to $m\ell_A$.

These acoustic oscillations involve motion of the baryon-photon fluid. When the oscillation of one Fourier mode is at its extreme, e.g., at the maximal compression in the potential well, the fluid is momentarily at rest, but then it begins flowing out of the well until the other extreme, the maximal decompression, is reached. Therefore those Fourier modes $\mathbf{k}$ which have the maximum effect on the CMB anisotropy via the $\frac{1}{4}\delta_\gamma(t_{\rm dec}, \mathbf{x}_{\rm ls}) + \Phi(t_{\rm dec}, \mathbf{x}_{\rm ls})$ term (the effective temperature effect) in Eq. (78) have the minimum effect via the $-\mathbf{v}\cdot\hat{\mathbf{n}}(t_{\rm dec}, \mathbf{x}_{\rm ls})$ term (the Doppler effect) and vice versa. Therefore the Doppler effect also contributes a peak structure to the C_ℓ spectrum, but the peaks are in the locations where the effective temperature contribution has troughs.

The Doppler effect is subdominant to the effective temperature effect, and therefore the peak positions in the C_ℓ spectrum are determined by the effective temperature effect, according to Eq. (101). The Doppler effect just partially fills the troughs between the peaks, weakening the peak structure of C_ℓ. See Fig. 22.

Fig. 20 shows the values of the effective temperature perturbation $\Theta_0 + \Phi$ (as well as Θ_0 and Φ separately) and the magnitude of the velocity perturbation ($\Theta_1 \sim v/3$) at $t_{\rm dec}$ as a function of the scale k. This is a result of a numerical calculation which includes the effect of the expansion of the universe, but not diffusion damping (Sec. 9.8).

9.8 Diffusion damping

For small enough scales the effect of photon diffusion and the finite thickness of the last scattering surface ($\sim$ the photon mean free path just before last scattering) smooth out the photon distribution and the CMB anisotropy.

This effect can be characterized by the damping scale $k_D^{-1} \sim$ photon diffusion length $\sim$ geometric mean of the Hubble time and photon mean free path λ_γ. Actually λ_γ is increasing rapidly during recombination, so a calculation of the diffusion scale involves an integral over time which includes this effect.

A calculation, that we shall not do here,[10] gives that photon density and velocity perturbations at scale k are damped at $t_{\rm dec}$ by

$$e^{-k^2/k_D^2}\,, \tag{104}$$

where the diffusion scale is

$$k_D^{-1} \sim \frac{1}{\text{few}}\frac{1}{a}\sqrt{\frac{\lambda_\gamma(t_{\rm dec})}{H_{\rm dec}}}\,. \tag{105}$$

Accordingly, the C_ℓ spectrum is also damped as

$$e^{-\ell^2/\ell_D^2} \tag{106}$$

where

$$\ell_D \sim k_D d_A^c(t_{\rm dec})\,. \tag{107}$$

For typical values of cosmological parameters $\ell_D \sim 1500$. See Fig. 21 for a result of a numerical calculation with and without diffusion damping.

[10] See, e.g., Dodelson [9], Chapter 8.

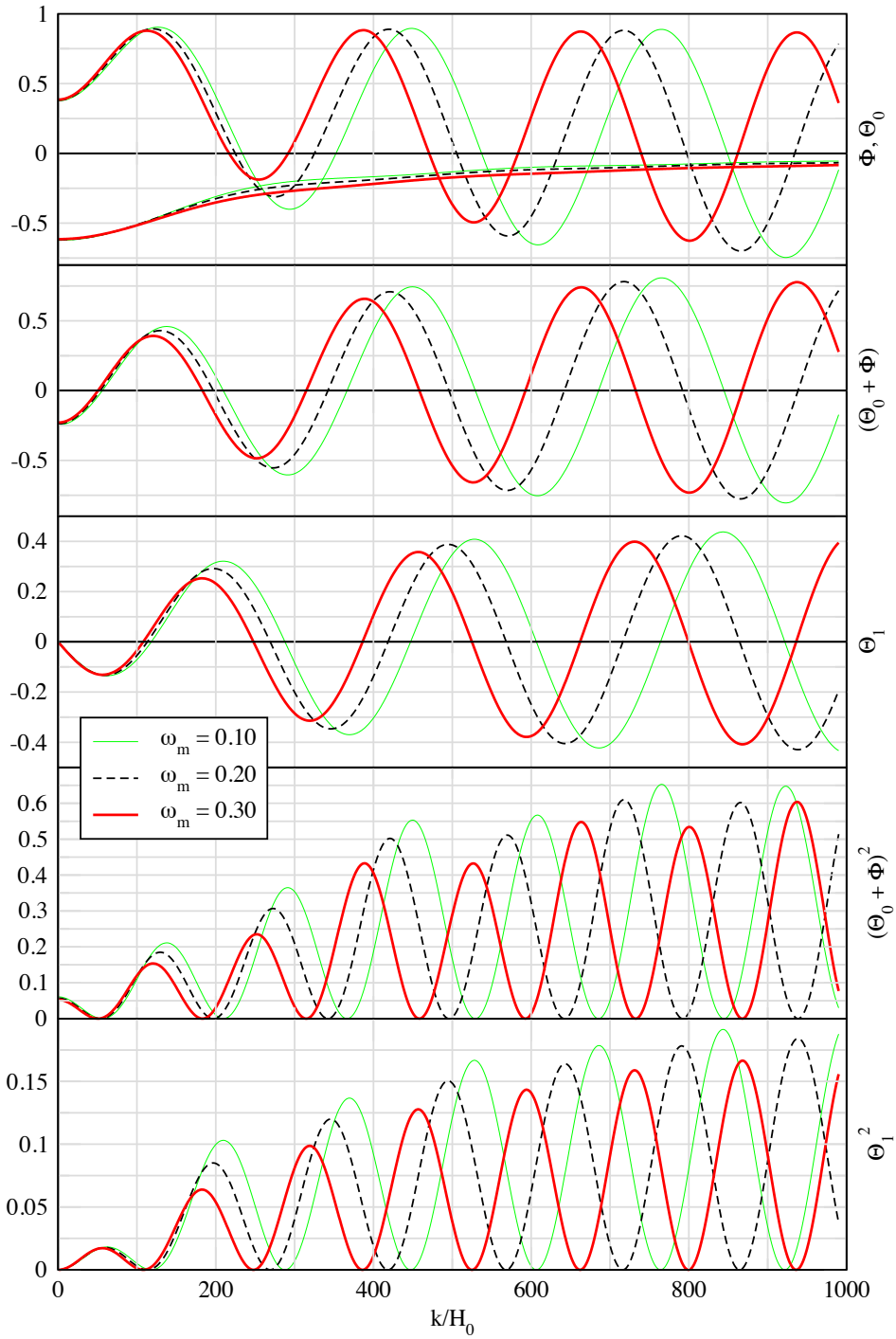


Figure 20: Values of oscillating quantities (normalized to an initial value $\mathcal{R}_{\mathbf{k}} = 1$) at the time of decoupling as a function of the scale k, for three different values of ω_m, and for $\omega_b = 0.01$. Θ_1 represents the velocity perturbation. The effect of diffusion damping is neglected. Figure and calculation by R. Keskitalo.

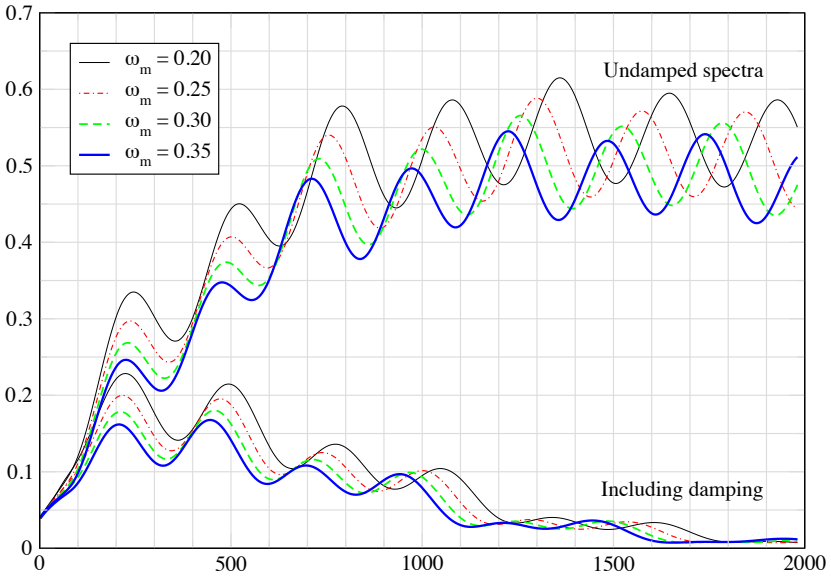


Figure 21: The angular power spectrum C_ℓ, calculated both with and without the effect of diffusion damping. The spectrum is given for four different values of ω_m, with $\omega_b = 0.01$. (This is a rather low value of ω_b, so $\ell_D < 1500$ and damping is quite strong.) Figure and calculation by R. Keskitalo.

Of the cosmological parameters, the damping scale is the most strongly dependent on ω_b, since increasing the baryon density shortens the photon mean free path before decoupling. Thus for larger ω_b the damping moves to shorter scales, i.e., ℓ_D becomes higher (there is less damping).

(Of course, decoupling only happens as the photon mean free path becomes comparable to the Hubble length, so one might think that λ_γ at $t_{\rm dec}$ should be independent of ω_b. However there is a distinction here between whether a photon will not scatter again after a particular scattering and what was the mean free path between the second-to-last and the last scattering. And k_D depends on an integral over the past history of the photon mean free path, not just the last one. The factor 1/few in Eq. (105) comes from that integration, and actually depends on ω_b. For small ω_b the λ_γ has already become quite large through the slow dilution of the baryon density by the expansion of the universe, and relies less on the fast reduction of free electron density due to recombination. Thus the time evolution of λ_γ before decoupling is different for different ω_b and we get a different diffusion scale.)

9.9 The complete C_ℓ spectrum

As we have discussed the CMB anisotropy has three contributions (see Eq. 78), the effective temperature effect,

$$\frac{1}{4}\delta_\gamma(t_{\rm dec}, \mathbf{x}_{\rm ls}) + \Phi(t_{\rm dec}, \mathbf{x}_{\rm ls})\,, \tag{108}$$

the Doppler effect,

$$-\mathbf{v}\cdot\hat{\mathbf{n}}(t_{\rm dec}, \mathbf{x}_{\rm ls})\,, \tag{109}$$

and the integrated Sachs–Wolfe effect,

$$2\int_{t_{\rm dec}}^{t_0} \dot{\Phi}(t, \mathbf{x}(t))dt\,. \tag{110}$$

Since the C_ℓ is a quadratic quantity, it also includes cross terms between these three effects.

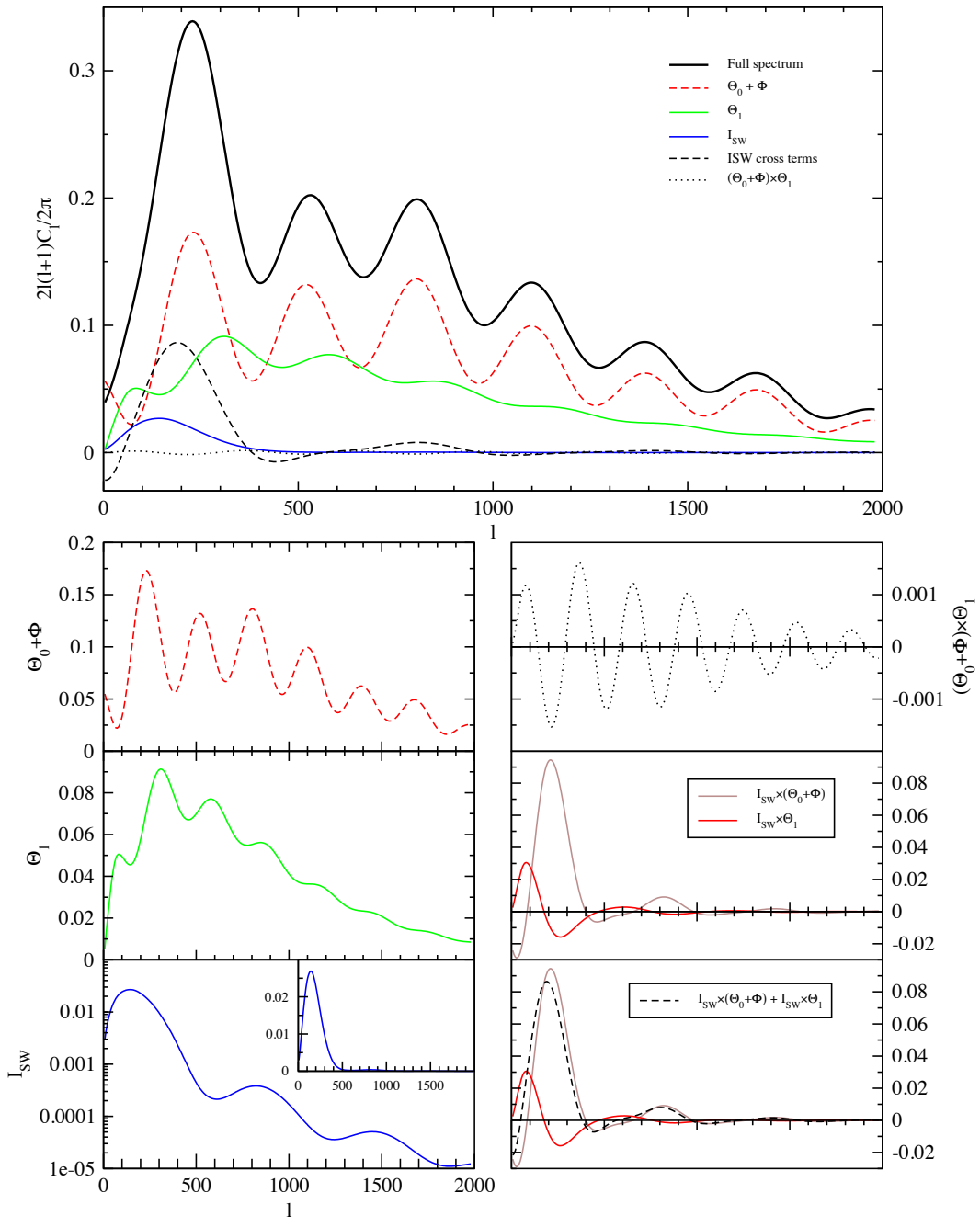


Figure 22: The full C_ℓ spectrum calculated for the cosmological model $\Omega_0 = 1$, $\Omega_\Lambda = 0$, $\omega_m = 0.2$, $\omega_b = 0.03$, $A_s = 1$, $n_s = 1$, and the different contributions to it. (The calculation involves some approximations which allow the description of C_ℓ as just a sum of these contributions and is not as accurate as a CMBFAST or CAMB calculation.) Here Θ_1 denotes the Doppler effect. Figure and calculation by R. Keskitalo.

The calculation of the full C_ℓ proceeds much as the calculation of just the ordinary Sachs–Wolfe part (which the effective temperature effect becomes at superhorizon scales) in Sec. 9.6.1, but now with the full $\delta T/T$. Since all perturbations are proportional to the primordial perturbations, the C_ℓ spectrum is proportional to the primordial perturbation spectrum $\mathcal{P}_\mathcal{R}(k)$ (with integrals over the spherical Bessel functions $j_\ell(kx)$, like in Eq. (70), to get from k to ℓ).

The difference is that instead of the constant proportionality factor $(\delta T/T)_{SW} = -(1/5)\mathcal{R}$, we have a k-dependent proportionality resulting from the evolution (including, e.g., the acoustic oscillations) of the perturbations.

In Fig. 22 we show the full C_ℓ spectrum and the different contributions to it.

Because the Doppler effect and the effective temperature effect are almost completely off-phase, their cross term gives a negligible contribution.

Since the ISW effect is relatively weak, it contributes more via its cross terms with the Doppler effect and effective temperature than directly. The cosmological model used for Fig. 22 has $\Omega_\Lambda = 0$, so there is no late ISW effect (which would contribute at the very lowest ℓ), and the ISW effect shown is the early ISW effect due to radiation contribution to the expansion law. This effect contributes mainly to the first peak and to the left of it, explaining why the first peak is so much higher than the other peaks. It also shifts the first peak position slightly to the left and changes its shape.

9.10 Cosmological parameters and CMB anisotropy

Let us finally consider the total effect of the various cosmological parameters on the C_ℓ spectrum. The C_ℓ provides the most important single observational data set for determining (or constraining) cosmological parameters, since it has a rich structure which we can measure with an accuracy that other cosmological observations cannot match, and because it depends on so many different cosmological parameters in many ways. The latter is both a strength and weakness: the number of cosmological parameters we can determine is large, but on the other hand, some feature in C_ℓ may depend on more than one parameter, so that we may only be able to constrain some combination of such parameters, not the parameters individually. We say that such parameters are *degenerate* in the CMB data. Other cosmological observations are then needed to break such degeneracies.

We shall consider 7 "standard parameters":

- Ω_0 total density parameter
- Ω_Λ cosmological constant (or vacuum energy) density parameter
- A_s amplitude of primordial scalar perturbations (at some pivot scale k_p)
- n_s spectral index of primordial scalar perturbations
- τ optical depth due to reionization (discussed in Sec. 9.10.6)
- $\omega_b \equiv \Omega_b h^2$ "physical" baryon density parameter
- $\omega_m \equiv \Omega_m h^2$ "physical" matter density parameter

There are other possible cosmological parameters ("additional parameters") which might affect the C_ℓ spectrum, e.g.,

- m_{ν_i} neutrino masses
- w dark energy equation-of-state parameter
- $\frac{dn_s}{d\ln k}$ scale dependence of the spectral index
- r, n_T relative amplitude and spectral index of tensor perturbations
- B, $n_{\rm iso}$ amplitudes and spectral indices of primordial isocurvature perturbations
- $A_{\rm cor}$, $n_{\rm cor}$ and their correlation with primordial curvature perturbations

We assume here that these additional parameters have no impact, i.e., they have the "standard" values

$$r = \frac{dn}{d\ln k} = B = A_{\rm cor} = 0\,, \quad w = -1\,, \quad \text{and} \quad \sum m_{\nu_i} = 0.06\,\text{meV}\,, \tag{111}$$

to the accuracy which matters for C_ℓ observations. This is both observationally and theoretically reasonable. There is no sign in the present-day CMB data for deviations from these values. On the other hand, significant deviations can be consistent with the current data, and may be discovered by more accurate future observations. The primordial isocurvature perturbations refer to the possibility that the primordial scalar perturbations are not adiabatic, and therefore are not completely determined by the comoving curvature perturbation $\mathcal{R}$.

The assumption that these additional parameters have no impact, leads to a determination of the standard parameters with an accuracy that may be too optimistic, since the standard parameters may have degeneracies with the additional parameters.

9.10.1 Independent vs. dependent parameters

The above is our choice of independent cosmological parameters. Ω_m, Ω_b and H_0 (or h) are then dependent (or "derived") parameters, since they are determined by

$$\Omega_0 = \Omega_m + \Omega_\Lambda \quad \Rightarrow \quad \Omega_m = \Omega_0 - \Omega_\Lambda \tag{112}$$

$$h = \sqrt{\frac{\omega_m}{\Omega_m}} = \sqrt{\frac{\omega_m}{\Omega_0 - \Omega_\Lambda}} \tag{113}$$

$$\Omega_b = \frac{\omega_b}{h^2} = \frac{\omega_b}{\omega_m}(\Omega_0 - \Omega_\Lambda) \tag{114}$$

Note that the Hubble constant $H_0 \equiv h \cdot 100$ km/s/Mpc is now a dependent parameter! We cannot vary it independently, but rather the varying of ω_m, Ω_0, or Ω_Λ also causes H_0 to change.

Different choices of independent parameters are possible within our 7-dimensional parameter space (e.g., we could have chosen H_0 to be an independent parameter and let Ω_Λ to be a dependent parameter instead). They can be thought of as different coordinate systems[11] in this 7D space. *It is not meaningful to discuss the effect of one parameter without specifying what is your set of independent parameters!*

Some choices of independent parameters are better than others. The above choice represents standard practice in cosmology today.[12] The independent parameters have been chosen so that they correspond as directly as possible to physics affecting the C_ℓ spectrum and thus to observable features in it. We want the effects of our independent parameters on the observables to be as different ("orthogonal") as possible in order to avoid parameter degeneracy.

In particular,

- ω_m (not Ω_m) determines $z_{\rm eq}$ and $k_{\rm eq}$, and thus, e.g., the magnitude of the early ISW effect and which scales enter during matter- or radiation-dominated epochs.
- ω_b (not Ω_b) determines the baryon/photon ratio and thus, e.g., the relative heights of the odd and even peaks.
- Ω_Λ (not $\Omega_\Lambda h^2$) determines the late ISW effect.

There are many effects on the C_ℓ spectrum, and parameters act on them in different combinations. Thus there is no perfectly "clean" way of choosing independent parameters. Especially having the Hubble constant as a dependent parameter takes some getting used to.

In the following CAMB[13] plots we see the effect of these parameters on C_ℓ by varying one parameter at a time around a *reference model*, whose parameters have the following values.

Independent parameters:

$$\Omega_0 = 1 \qquad \Omega_\Lambda = 0.7$$
$$A_s = 1 \qquad \omega_m = 0.147$$
$$n_s = 1 \qquad \omega_b = 0.022$$
$$\tau = 0.1$$

[11] The situation is analogous to the choice of independent thermodynamic variables in thermodynamics.

[12] There are other choices in use, which are even more geared to minimizing parameter degeneracy. For example, the sound horizon angle ϑ_s may be used instead of Ω_Λ as an independent parameter, since it is directly determined by the acoustic peak separation. However, since the determination of the dependent parameters from it is complicated, such use is more directed towards technical data analysis than pedagogical discussion.

[13] CAMB is a publicly available code for precise calculation of the C_ℓ spectrum. See http://camb.info

which give for the dependent parameters

$$\begin{aligned} \Omega_m &= 0.3 & h &= 0.7 \\ \Omega_c &= 0.2551 & \omega_c &= 0.125 \\ \Omega_b &= 0.0449 \end{aligned}$$

The meaning of setting $A_s = 1$ is just that the resulting C_ℓ still need to be multiplied by the true value of A_s^2. (In this model the true value should be about $A_s = 5 \times 10^{-5}$ to agree with observations.) If we really had $A_s = 1$, perturbation theory of course would not be valid! This is a relatively common practice, since the effect of changing A_s is so trivial that it makes not much sense to plot C_ℓ separately for different values of A_s.

9.10.2 Sound horizon angle

The positions of the acoustic peaks of the C_ℓ spectrum provide us with a measurement of the sound horizon angle

$$\vartheta_s \equiv \frac{r_s(t_{\rm dec})}{d_A^c(t_{\rm dec})}$$

We can use this in the determination of the values of the cosmological parameters, once we have calculated how this angle depends on those parameters. It is the ratio of two quantities, the sound horizon at photon decoupling, $r_s(t_{\rm dec})$, and the angular diameter distance to the last scattering, $d_A^c(t_{\rm dec})$.

Angular diameter distance to last scattering

The angular diameter distance $d_A^c(t_{\rm dec})$ to the last scattering surface we have already calculated and it is given by Eq. (39) as

$$d_A^c(t_{\rm dec}) = H_0^{-1} \frac{1}{\sqrt{|\Omega_0 - 1|}}\, f_k \left(\sqrt{|\Omega_0 - 1|} \int_{\frac{1}{1+z_{\rm dec}}}^{1} \frac{da}{\sqrt{\Omega_0(a - a^2) - \Omega_\Lambda(a - a^4) + a^2}} \right), \quad (115)$$

from which we see that it depends on the three cosmological parameters H_0, Ω_0 and Ω_Λ. Here $\Omega_0 = \Omega_m + \Omega_\Lambda$, so we could also say that it depends on H_0, Ω_m, and Ω_Λ, but it is easier to discuss the effects of these different parameters if we keep Ω_0 as an independent parameter, instead of Ω_m, since the "geometry effect" of the curvature of space, which determines the relation between the comoving angular diameter distance d_A^c and the comoving distance d^c, is determined by Ω_0.

1. The comoving angular diameter distance is inversely proportional to H_0 (directly proportional to the Hubble distance H_0^{-1}).

2. Increasing Ω_0 decreases $d_A^c(t_{\rm dec})$ in relation to $d^c(t_{\rm dec})$ because of the geometry effect.

3. With a fixed Ω_Λ, increasing Ω_0 decreases $d^c(t_{\rm dec})$, since it means increasing Ω_m, which has a decelerating effect on the expansion. With a fixed present expansion rate H_0, deceleration means that expansion was faster earlier $\Rightarrow$ universe is younger $\Rightarrow$ there is less time for photons to travel as the universe cools from $T_{\rm dec}$ to T_0 $\Rightarrow$ last scattering surface is closer to us.

4. Increasing Ω_Λ (with a fixed Ω_0) increases $d^c(t_{\rm dec})$, since it means a larger part of the energy density is in dark energy, which has an accelerating effect on the expansion. With fixed H_0, this means that expansion was slower in the past $\Rightarrow$ universe is older $\Rightarrow$ more time for photons $\Rightarrow$ last scattering surface is further out. $\therefore$ Ω_Λ increases $d_A^c(t_{\rm dec})$.

Here 2 and 3 work in the same direction: increasing Ω_0 decreases $d_A^c(t_{\rm dec})$, but the geometry effect (2) is stronger. See Fig. 17 for the case $\Omega_\Lambda = 0$, where the dashed line (the comoving distance) shows effect (3) and the solid line (the comoving angular diameter distance) the combined effect (2) and (3).

However, now we have to take into account that, in our chosen parametrization, H_0 is not an independent parameter, but

$$H_0^{-1} \propto \sqrt{\frac{\Omega_0 - \Omega_\Lambda}{\omega_m}}\,,$$

so that via H_0^{-1}, Ω_0 increases and Ω_Λ decreases $d_A^c(t_{\rm dec})$, which are the opposite effects to those discussed above. For Ω_Λ this opposite effect wins. See Fig. 25.

Sound horizon

To calculate the sound horizon,

$$r_s(t_{\rm dec}) \;=\; \int_0^{t_{\rm dec}} \frac{c_s(t)}{a(t)}dt \;=\; \int_0^{a_{\rm dec}} \frac{da}{a\cdot(da/dt)}c_s(a)\,, \tag{116}$$

we need the speed of sound, from Eq. (83),

$$c_s^2(x) = \frac{1}{3}\frac{1}{1+\frac{3}{4}\frac{\bar\rho_b}{\bar\rho_\gamma}} = \frac{1}{3}\frac{1}{1+\frac{3}{4}\frac{\omega_b}{\omega_\gamma}a}\,, \tag{117}$$

where the upper limit of the integral is $a_{\rm dec} = 1/(1+z_{\rm dec})$.

The other element in the integrand of Eq. (116) is the expansion law $a(t)$ before decoupling. From Chapter 3 we have that

$$a\frac{da}{dt} = H_0\sqrt{\Omega_r + \Omega_m a + (1-\Omega_0)a^2 + \Omega_\Lambda a^4}\,. \tag{118}$$

In the integral (115) we dropped the Ω_r, since it is important only at early times, and the integral from $a_{\rm dec}$ to 1 is dominated by late times. Integral (116), on the other hand, includes only early times, and now we can instead drop the Ω_Λ and $1-\Omega_0$ terms (i.e., we can ignore the effect of curvature and dark energy in the early universe, before photon decoupling), so that

$$a\frac{da}{dt} \approx H_0\sqrt{\Omega_m a + \Omega_r} = H_{100}\sqrt{\omega_m a + \omega_r} = \frac{\sqrt{\omega_m a + \omega_r}}{2998\,{\rm Mpc}}\,, \tag{119}$$

where we have written

$$H_0 \equiv h\cdot 100\frac{\rm km/s}{\rm Mpc} \equiv h\cdot H_{100} = \frac{h}{2997.92\,{\rm Mpc}}\,. \tag{120}$$

Thus the sound horizon is given by

$$\begin{aligned} r_s(a) &= 2998\,{\rm Mpc}\int_0^a \frac{c_s(x)dx}{\sqrt{\omega_m x + \omega_r}} \\ &= 2998\,{\rm Mpc}\cdot\frac{1}{\sqrt{3\omega_r}}\int_0^a \frac{dx}{\sqrt{\left(1+\frac{\omega_m}{\omega_r}x\right)\left(1+\frac{3}{4}\frac{\omega_b}{\omega_\gamma}x\right)}}\,. \end{aligned} \tag{121}$$

Here

$$\omega_\gamma = 2.4702\times 10^{-5} \qquad \text{and} \tag{122}$$

$$\omega_r = \left[1+\frac{7}{8}N_\nu\left(\frac{4}{11}\right)^{4/3}\right]\omega_\gamma = 1.6904\,\omega_\gamma = 4.1756\times 10^{-5} \tag{123}$$

are accurately known from the CMB temperature $T_0 = 2.725\,\mathrm{K}$ (and therefore we do not consider them as cosmological parameters in the sense of something to be determined from the C_ℓ spectrum).

Thus the sound horizon depends on the two cosmological parameters ω_m and ω_b,

$$r_s(t_{\rm dec}) = r_s(\omega_m, \omega_b)$$

From Eq. (121) we see that increasing either ω_m or ω_b *makes the sound horizon at decoupling,* $r_s(a_{\rm dec})$*, shorter*:

- ω_b slows the sound down
- ω_m speeds up the expansion at a given temperature, so the universe cools to $T_{\rm dec}$ in less time.

The integral (121) can be done and it gives

$$r_s(t_{\rm dec}) = \frac{2998\,\mathrm{Mpc}}{\sqrt{1+z_{\rm dec}}}\frac{2}{\sqrt{3\omega_m R_*}}\ln\frac{\sqrt{1+R_*}+\sqrt{R_*+r_*R_*}}{1+\sqrt{r_*R_*}}\,, \tag{124}$$

where

$$r_* \equiv \frac{\bar\rho_r(t_{\rm dec})}{\bar\rho_m(t_{\rm dec})} = \frac{\omega_r}{\omega_m}(1+z_{\rm dec}) = 0.0456\frac{1}{\omega_m}\left(\frac{1+z_{\rm dec}}{1091}\right) \tag{125}$$

$$R_* \equiv \frac{3\bar\rho_b(t_{\rm dec})}{4\bar\rho_\gamma(t_{\rm dec})} = \frac{3\omega_b}{4\omega_\gamma}\frac{1}{1+z_{\rm dec}} = 27.8\,\omega_b\left(\frac{1091}{1+z_{\rm dec}}\right)\,. \tag{126}$$

For our reference values $\omega_m = 0.147$, $\omega_b = 0.022$, and $1 + z_{\rm dec} = 1091$[14] we get $r_* = 0.310$ and $R_* = 0.614$ and $r_s(t_{\rm dec}) = 144\,\mathrm{Mpc}$ for the sound horizon at decoupling.

Summary

The angular diameter distance $d_A^c(t_{\rm dec})$ is the most naturally discussed in terms of H_0, Ω_0, and Ω_Λ, but since these are not the most convenient choice of independent parameters for other purposes, we shall trade H_0 for ω_m according to Eq. (113). Thus we have that the sound horizon angle depends on 4 parameters,

$$\vartheta_s \equiv \frac{r_s(\omega_m,\omega_b)}{d_A^c(\Omega_0,\Omega_\Lambda,\omega_m)} = \vartheta_s(\Omega_0,\Omega_\Lambda,\omega_m,\omega_b) \tag{127}$$

9.10.3 Acoustic peak heights

There are a number of effects affecting the heights of the acoustic peaks:

1. **The early ISW effect.** The early ISW effect raises the first peak. It is caused by the evolution of Φ because of the effect of the radiation contribution on the expansion law after $t_{\rm dec}$. This depends on the radiation-matter ratio at that time; decreasing ω_m makes the early ISW effect stronger.
2. **Shift of oscillation equilibrium by baryons.** (Baryon drag.) This makes the odd peaks (which correspond to compression of the baryon-photon fluid in the potential wells, decompression on potential hills) higher, and the even peaks (decompression at potential wells, compression on top of potential hills) lower.

[14] Photon decoupling temperature, and thus $1+z_{\rm dec}$, depends somewhat on ω_b, but since this dependence is not easy to calculate (recombination and photon decoupling were discussed in Chapter 4), we have mostly ignored this dependence and used the fixed value $1 + z_{\rm dec} = 1091$.

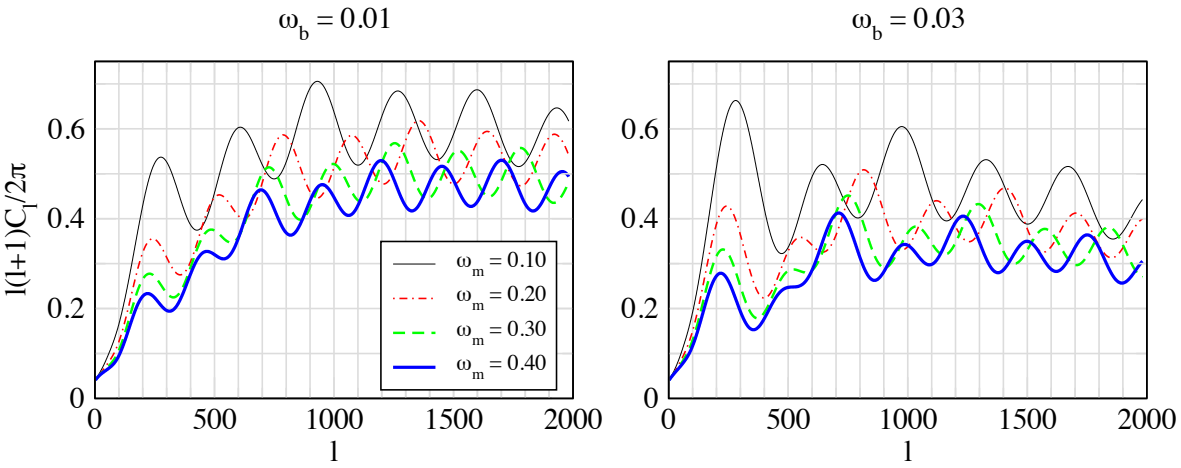


Figure 23: The effect of ω_m. The angular power spectrum C_ℓ is here calculated without the effect of diffusion damping, so that the other effects on peak heights could be seen more clearly. Notice how reducing ω_m raises all peaks, but the effect on the first few peaks is stronger in relative terms, as the radiation driving effect is extended towards larger scales (smaller ℓ). The first peak is raised mainly because the ISW effect becomes stronger. Figure and calculation by R. Keskitalo.

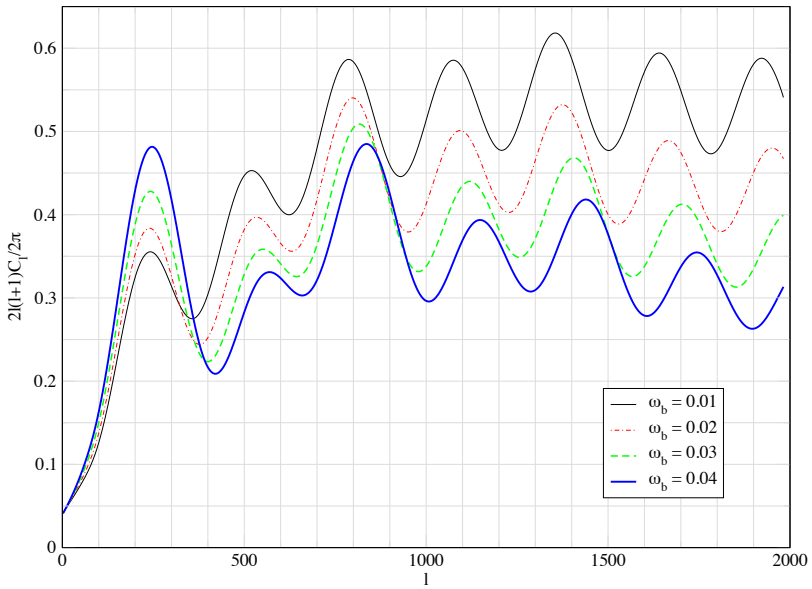


Figure 24: The effect of ω_b. The angular power spectrum C_ℓ is here calculated without the effect of diffusion damping, so that the other effects on peak heights could be seen more clearly. Notice how increasing ω_b raises odd peaks relative to the even peaks. Because of baryon damping there is a general trend downwards with increasing ω_b. This figure is for $\omega_m = 0.20$. Figure and calculation by R. Keskitalo.

3. **Baryon damping.** The time evolution of $R \equiv 3\bar{\rho}_b/4\bar{\rho}_\gamma$ causes the amplitude of the acoustic oscillations to be damped in time roughly as $(1+R)^{-1/4}$. This reduces the amplitudes of all peaks.

4. **Radiation driving.**[15] This is an effect related to horizon scale physics that we have not tried to properly calculate. For scales k which enter during the radiation-dominated epoch, or near matter-radiation equality, the potential Φ decays around the time when the scale enters. The potential keeps changing as long as the radiation contribution is important, but the largest change in Φ is around horizon entry. Because the sound horizon and Hubble length are comparable, horizon entry and the corresponding potential decay always happen during the first oscillation period. This means that the baryon-photon fluid is falling into a deep potential well, and therefore is compressed by gravity by a large factor, before the resulting overpressure is able to push it out. Meanwhile the potential has decayed, so it is less able to resist the decompression phase, and the overpressure is able to kick the fluid further out of the well. This increases the amplitude of the acoustic oscillations. The effect is stronger for the smaller scales which enter when the universe is more radiation-dominated, and therefore raises the peaks with a larger peak number m more. Reducing ω_m makes the universe more radiation dominated, making this effect stronger and extending it towards the peaks with lower peak number m.

5. **Diffusion damping.** Diffusion damping lowers the heights of the peaks. It acts in the opposite direction than the radiation driving effect, lowering the peaks with a larger peak number m more. Because the diffusion damping effect is exponential in ℓ, it wins for large ℓ.

Effects 1 and 4 depend on ω_m, effects 2, 3, and 5 on ω_b. See Figs. 23 and 24 for the effects of ω_m and ω_b on peak heights.

[15] This is also called gravitational driving, which is perhaps more appropriate, since the effect is due to the change in the gravitational potential.

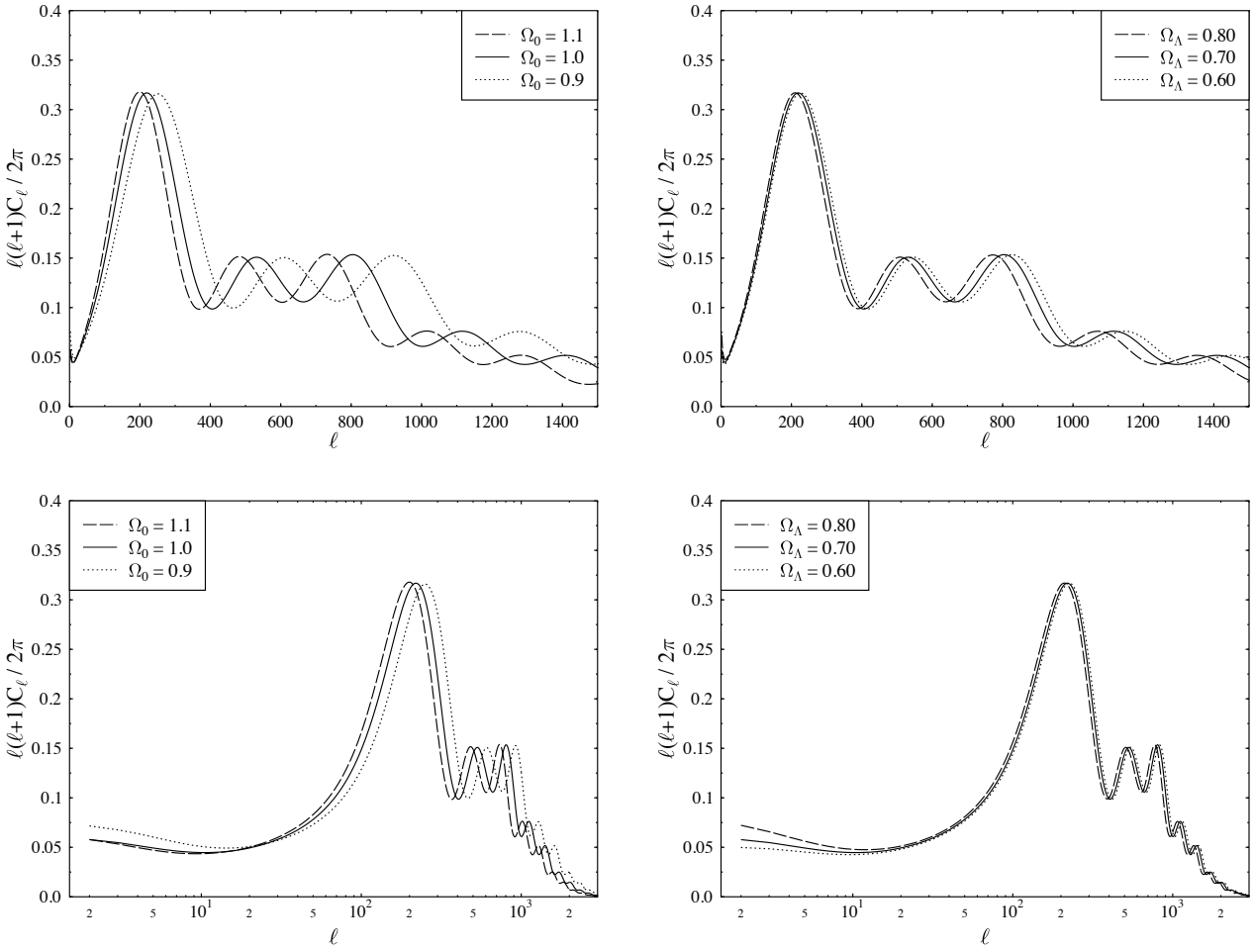


Figure 25: The effect of changing Ω_0 or Ω_Λ from their reference values $\Omega_0 = 1$ and $\Omega_\Lambda = 0.7$. The top panels shows the C_ℓ spectrum with a linear ℓ scale so that details at larger ℓ where cosmic variance effects are smaller can be better seen. The bottom plot has a logarithmic ℓ scale so that the integrated Sachs-Wolfe effect at small ℓ can be better seen. The logarithmic scale also makes clear that the effect of the change in sound horizon angle is to stretch the spectrum by a constant factor in ℓ space.

9.10.4 Effect of Ω_0 and Ω_Λ

These two parameters have only two effects:

1. they affect the sound horizon angle and thus the positions of the acoustic peaks
2. they affect the late ISW effect

See Fig. 25. Since the late ISW effect is in the region of the C_ℓ spectrum where the cosmic variance is large, it is difficult to detect. Thus we can in practice only use ϑ_s to determine Ω_0 and Ω_Λ. Since ω_b and ω_m can be determined quite accurately from C_ℓ acoustic peak heights, peak separation, i.e., ϑ_s, can then indeed be used for the determination of Ω_0 and Ω_Λ. Since one number cannot be used to determine two, the parameters Ω_0 and Ω_Λ are degenerate. CMB observations alone cannot be used to determine them both. Other cosmological observations (like the power spectrum $P_\delta(k)$ from large scale structure, or the SNIa redshift-distance relationship) are needed to break this degeneracy.

A fixed ϑ_s together with fixed ω_b and ω_m determine a line on the (Ω_0, Ω_Λ) -plane. See Fig. 26. Derived parameters, e.g., h, vary along that line. As you can see from Fig. 25, changing Ω_0 (around the reference model) affects ϑ_s much more strongly than changing Ω_Λ. This means

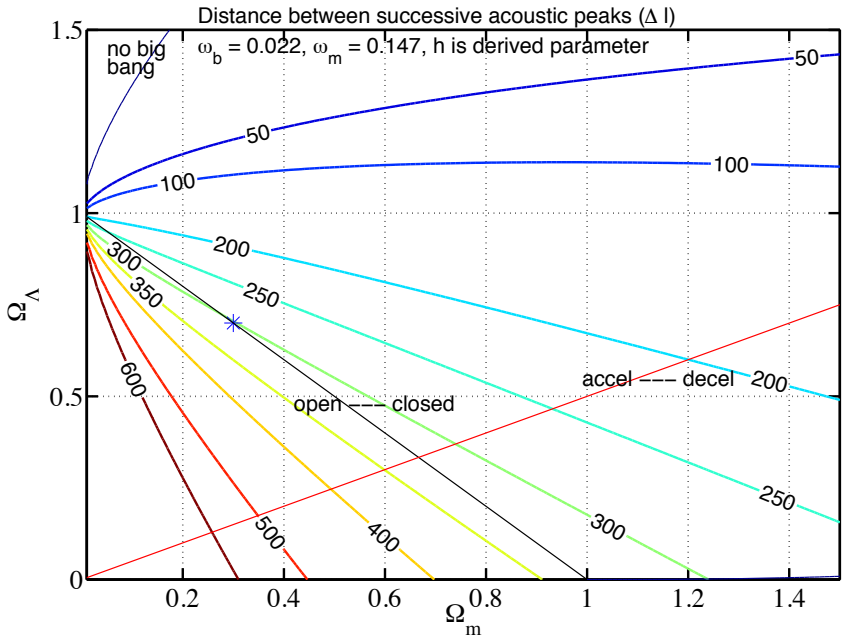


Figure 26: The lines of constant sound horizon angle ϑ_s on the $(\Omega_m,\Omega_\Lambda)$ plane for fixed ω_b and ω_m. The numbers on the lines refer to the corresponding acoustic scale $\ell_A \equiv \pi/\vartheta_s$ ($\sim$ peak separation) in multipole space. Figure by J. Väliviita. See his PhD thesis[10], p.70, for an improved version including the HST constraint on h.

that the orientation of the line is such that Ω_Λ varies more rapidly along that line than Ω_0. Therefore using additional constraints from other cosmological observations, e.g., the Hubble Space Telescope determination of h based on the distance ladder, which select a short section from that line, gives us a fairly good determination of Ω_0, leaving the allowed range for Ω_Λ still quite large.

Therefore it is often said that CMB measurements have determined that $\Omega_0 \sim 1$. But as explained above, this determination necessary requires the use of some auxiliary cosmological data besides the CMB.

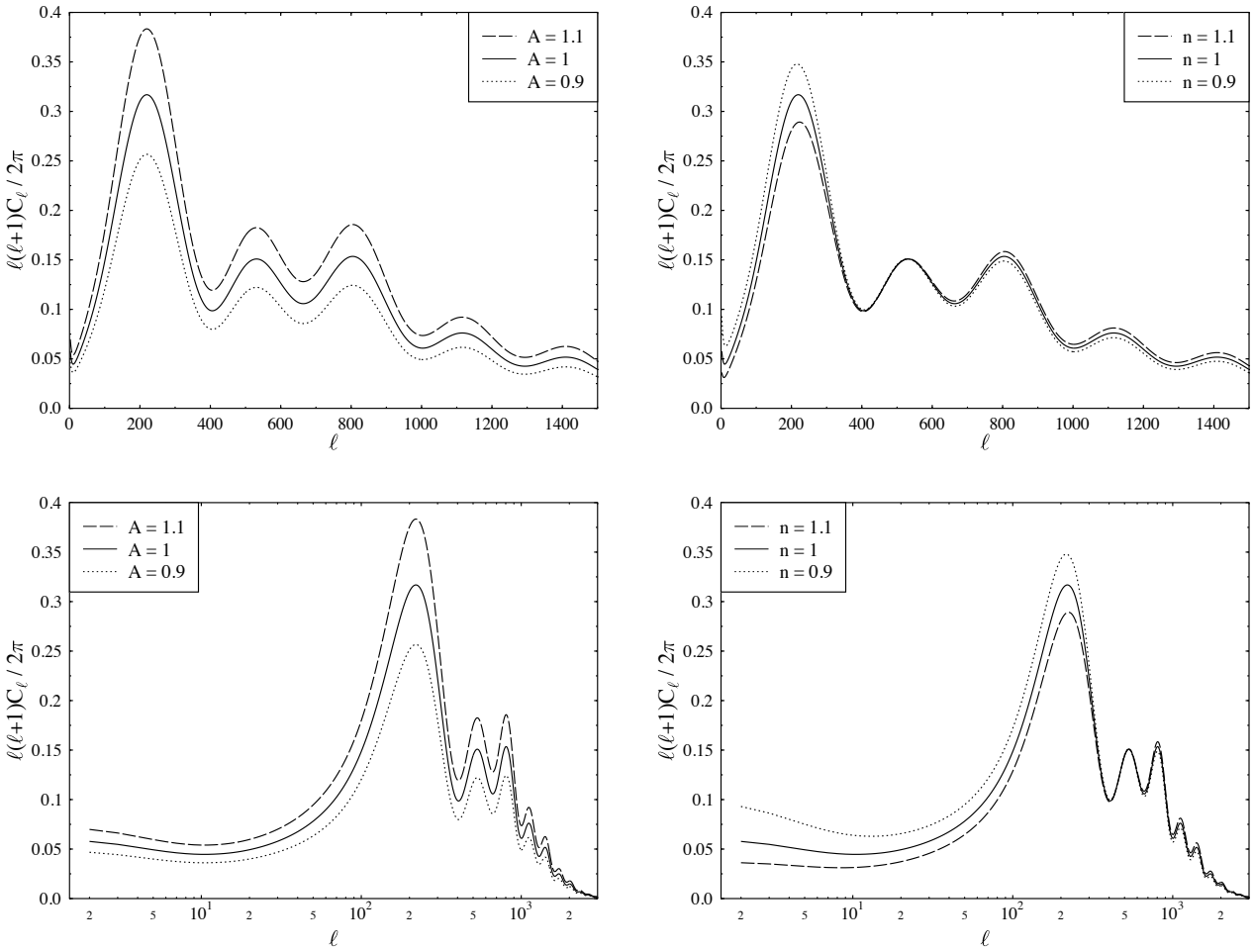


Figure 27: The effect of changing the primordial amplitude and spectral index from their reference values $A_s = 1$ and $n_s = 1$.

9.10.5 Effect of the primordial spectrum

The effect of the primordial spectrum is simple: raising the amplitude A_s raises the C_ℓ also, and tilting the primordial spectrum tilts the C_ℓ also. See Fig. 27.

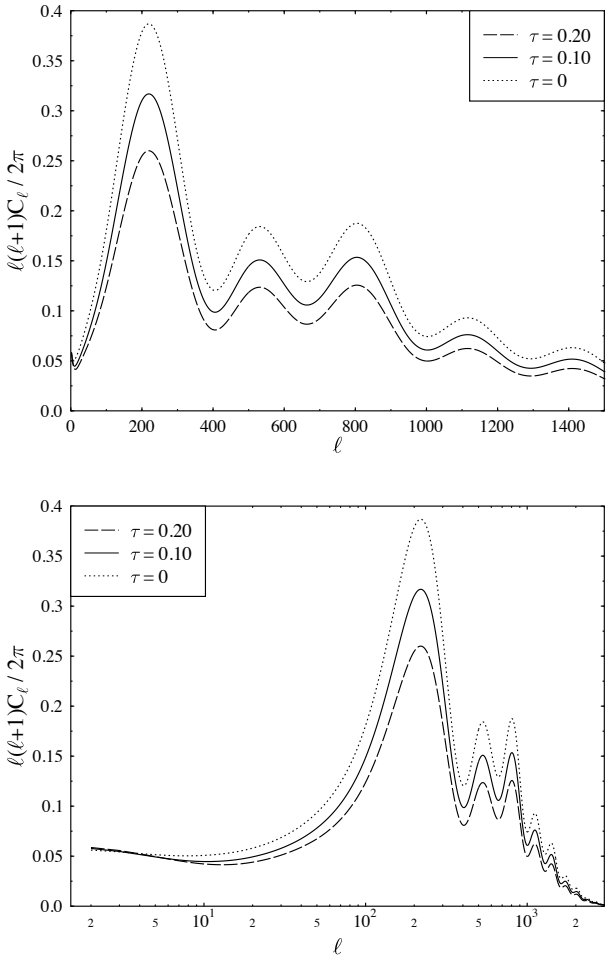


Figure 28: The effect of changing the optical depth from its reference value $\tau = 0.1$.

9.10.6 Optical depth due to reionization

When radiation from the first stars reionizes the intergalactic gas, CMB photons may scatter from the resulting free electrons. The optical depth τ due to reionization is the expectation number of such scatterings per CMB photon. It is expected to be about 0.1 or less, i.e., most CMB photons do not scatter at all. This rescattering causes additional polarization[16] of the CMB, and CMB polarization measurements are actually the best way to determine τ.

The optical depth is thus directly related to the reionization redshift $z_{\rm reion}$. A smaller τ corresponds to later reionization and thus means that the first stars formed later.

Because of this scattering, not all the CMB photons come from the location on the last scattering surface they seem to come from. The effect of the rescattered photons is to mix up signals from different directions and therefore to reduce the CMB anisotropy. The reduction factor on $\delta T/T$ is $e^{-\tau}$ and on the C_ℓ spectrum $e^{-2\tau}$. However, this does not affect the largest scales, scales larger than the area from which the rescattered photons reaching us from a certain direction, originally came from. Such a large-scale anisotropy has affected all such photons the same way, and thus is not lost in the mixing. See Fig. 28.

[16]Due to time constraints, CMB polarization is not discussed in these lectures.

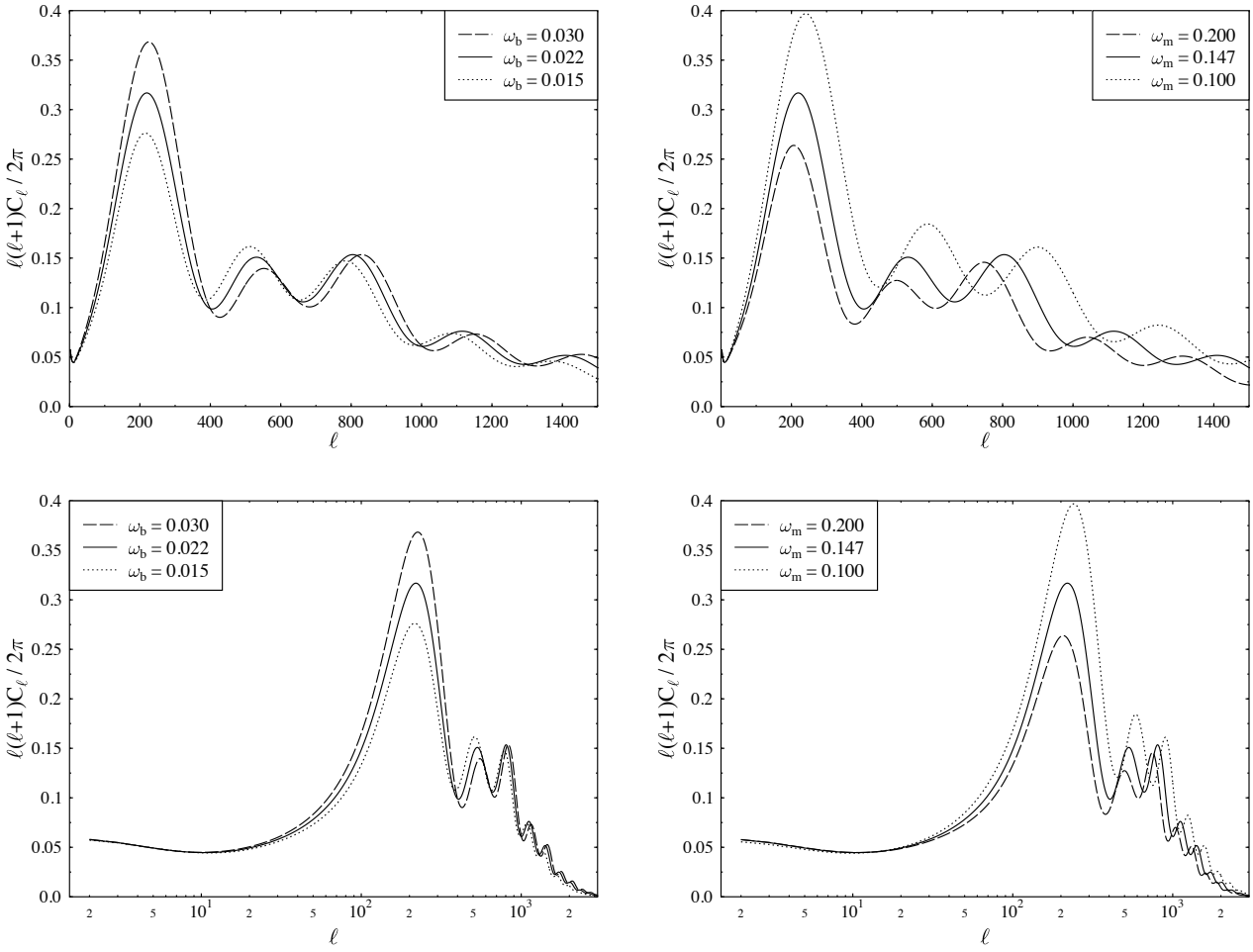


Figure 29: The effect of changing the physical baryon density and matter density parameters from their reference values $\omega_b = 0.022$ and $\omega_m = 0.147$.

9.10.7 Effect of ω_b and ω_m

These parameters affect both the positions of the acoustic peaks (through ϑ_s) and the heights of the different peaks. The latter effect is the more important, since both parameters have their own signature on the peak heights, allowing an accurate determination of these parameters, whereas the effect on ϑ_s is degenerate with Ω_0 and Ω_Λ.

Especially ω_b has a characteristic effect on peak heights: Increasing ω_b raises the odd peaks and reduces the even peaks, because it shifts the balance of the acoustic oscillations (the $-R\Phi$ effect). This shows the most clearly at the first and second peaks. Raising ω_b also shortens the damping scale k_D^{-1} due to photon diffusion, moving the corresponding damping scale ℓ_D of the C_ℓ spectrum towards larger ℓ. This has the effect of raising C_ℓ at large ℓ. See Fig. 29.

There is also an overall "baryon damping effect" on the acoustic oscillations which we have not calculated. It is due to the time dependence of $R \equiv 3\bar{\rho}_b/4\bar{\rho}_m$, which reduces the amplitude of the oscillation by about $(1+R)^{-1/4}$. This explains why the third peak in Fig. 29 is no higher for $\omega_b = 0.030$ than it is for $\omega_b = 0.022$.

Increasing ω_m makes the universe more matter dominated at $t_{\rm dec}$ and therefore it reduces the early ISW effect, making the first peak lower. This also affects the shape of the first peak.

The "radiation driving" effect is most clear at the second to fourth peaks. Reducing ω_m makes these peaks higher by making the universe more radiation-dominated at the time the corresponding scales enter, strengthening this radiation driving. The fifth and further peaks

Parameters for the ΛCDM model	
	Planck 2018
ω_b	0.02237 ± 0.00015
ω_m	0.1424 ± 0.0012
Ω_Λ	0.685 ± 0.007
τ	0.054 ± 0.007
A_s	$4.58 \pm 0.04 \times 10^{-5}$
n_s	0.965 ± 0.004
H_0	$67.36 \pm 0.54\,\mathrm{km/s/Mpc}$
ω_c	0.1200 ± 0.0012
Ω_m	0.315 ± 0.007
z_{eq}	3402 ± 26
k_{eq}^{-1}	$96.3 \pm 0.8\,\mathrm{Mpc}$
z_{dec}	1089.92 ± 0.25
k_D^{-1}	$7.10 \pm 0.02\,\mathrm{Mpc}$
z_{reion}	7.7 ± 0.7
ϑ_s	$0.5965^\circ \pm 0.0002^\circ$
t_0	$13.797 \pm 0.023 \times 10^9\,\mathrm{a}$

Table 2: These parameter values are based on the CMB temperature power spectrum C_ℓ, CMB polarization, and gravitational lensing of the CMB, as observed by the Planck satellite [5]. The first six parameters, above the line, are independent parameters, and the parameters below the line are quantities that can be derived from them in the ΛCDM model. The error estimates are 68% confidence limits. Note that here Ω_m includes the contribution from neutrinos with $\sum m_\nu = 0.06\,\mathrm{meV}$ ($\Omega_\nu = 0.0014$) whereas ω_m does not.

correspond to scales that have anyway essentially the full effect, and for the first peak this effect is anyway weak. (We see instead the ISW effect in the first peak.) See Fig. 29.

9.11 Current best estimates for the cosmological parameters

The most important data set for determining cosmological parameters is the Planck data [5] on the CMB anisotropy. We give the parameter values determined by Planck for the ΛCDM model in Table 2. Note that all independent parameters of the model are fit simultaneously to the same data. The determination is based on the assumption that the model, here ΛCDM, is correct. One can judge this assumption based on how well the model fits the data. In the case of Planck and ΛCDM the fit is good; adding more parameters to the model does not improve the fit significantly.

This model agrees reasonable well with most of the other available cosmological data, with the exception of the distance-ladder determination of the Hubble constant, based on Cepheids and Type Ia supernovae, which gives $H_0 = 73.5 \pm 1.6\,\mathrm{km/s/Mpc}$ [11, 12]. This is called the *local* measurement of H_0, since these measurements are from nearby parts of the Universe, in contrast to the *global* determination from the CMB, where the CMB has traversed the entire observable Universe. This discrepancy has been evident in the data for some time, but it has gradually become more serious as the error bars on H_0 from both CMB and local measurements have become tighter without the central values changing much. One may suspect systematic errors in the distance ladder data or that the ΛCDM model is a too simple model for the universe.

In the ΛCDM model the universe is flat, $\Omega_0 = 1$. We can also fit *extended models*, with additional independent parameters. Such 7-parameter models, with one extra parameter in addition to the ΛCDM parameters, are fit to Planck data in Table 3. Since the ΛCDM model

Constraints for extended models

	ΛCDM	Planck 2018	Planck+ext
Ω_0	1.0	1.011 ± 0.013	0.9993 ± 0.0037
r	0	< 0.101	< 0.065
$dn_s/d\ln k$	0	-0.005 ± 0.013	-0.004 ± 0.013
w	-1	-1.6 ± 0.5	-1.04 ± 0.10
$\sum m_\nu$	0.06 eV	< 0.241 eV	< 0.120 eV
$N_{\rm eff}$	3.046	2.89 ± 0.38	2.99 ± 0.34

Table 3: Each row is a different model and we show limits only to the "additional" parameter. As is customary with limits, the ranges are given as 95% confidence limits. $N_{\rm eff}$, the "effective number of neutrino species", refers to relativistic energy density (in addition to photons) near the time of photon decoupling. The ΛCDM value corresponds to the N_ν defined in Chapter 4.

is a good fit, the estimates for these extra parameters are consistent with their values in the ΛCDM model. Instead of the central value, we therefore concentrate on the estimated probable range, i.e., *limits* to the deviation from the ΛCDM model. Note that in these extended models the ranges for the 6 ΛCDM parameters will be different from Table 2; they will be wider and the central values will be slightly different. One could of course consider models with more independent parameters, e.g., the 12-parameter model, where all the 6 parameters of Table 3 were added to ΛCDM. In such a model the allowed ranges for all these parameters would be wider than in Tables 2 and 3. The argument against such a model is *Occam's razor*: if there are many models that fit the data, one should prefer the simplest one; a corollary to this is that the models one should consider next are those that are almost as simple. Of course, there is no guarantee against all these parameters having a significant effect on the CMB. These one-parameter extensions to ΛCDM do not relieve the tension with the local determination of H_0 much, but by adding sufficiently many additional parameters one can get rid of the tension.

The parameter $N_{\rm eff}$ corresponds to making ω_r a free parameter, i.e., replacing N_ν in (123) with $N_{\rm eff}$. From the discussion in Sec. 9.10.2 we see that we are constraining relativistic energy density at or before $t_{\rm dec}$.

Because of degeneracies of cosmological parameters in the CMB data, most importantly the *geometrical degeneracy* between parameters, like Ω_0, Ω_Λ, and the dark energy equation-of-state parameter w, whose main effect on CMB is via their effect on the angular diameter distance to the last scattering sphere, some parameters of these extended models are only weakly constrained by Planck data. To break these degeneracies, additional cosmological data (BAO and BICEP2/Keck, see below) has been used in the fourth column of Table 3 (ext = external to Planck). The impressive accuracy in this column is, however, mainly due to the accuracy of Planck. The parameter values allowed by Planck form a narrow but long region in the 7-parameter space, and the external data allows a region that is wider, but oriented differently; the intersection of these regions is then a shorter segment of the region allowed by Planck alone, see Fig. 30.

Large scale structure surveys, i.e., the measurement of the 3-dimensional matter power spectrum $P_\delta(k)$ from the distribution of galaxies, mainly measure the combination $\Omega_m h$, since this determines where $P_\delta(k)$ turns down. Actually it turns down at $k_{\rm eq}$ which is proportional to $\omega_m \equiv \Omega_m h^2$, but since in these surveys the distances to galaxies are deduced from their redshifts (these surveys are also called galaxy redshift surveys), which give the distances only up to the Hubble constant H_0, these surveys determine $h^{-1}k_{\rm eq}$ instead of $k_{\rm eq}$. This cancels one power of h. Having $\Omega_m h^2$ from CMB and $\Omega_m h$ from the galaxy surveys, gives us both h and $\Omega_m = \Omega_0 - \Omega_\Lambda$, which breaks the Ω_0-Ω_Λ degeneracy.

Measurements of $P_\delta(k)$ are now so accurate that the small residual effect from the acoustic

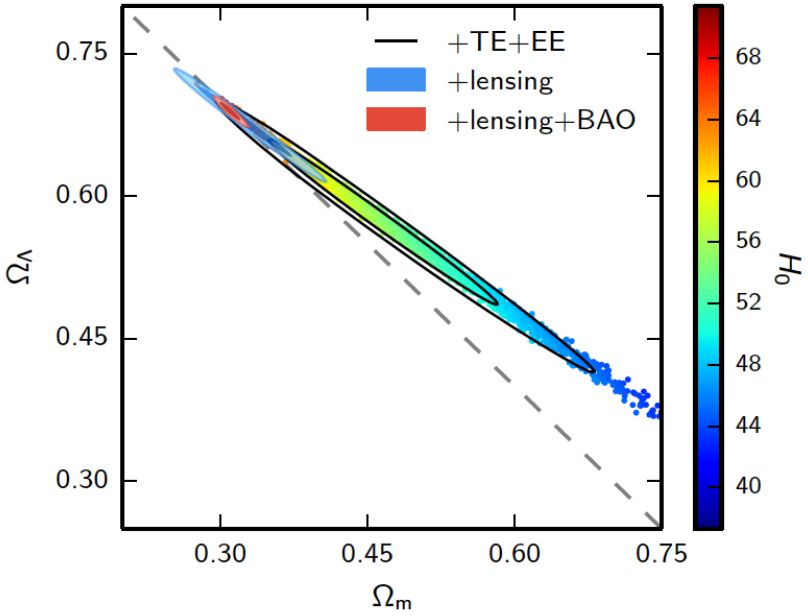


Figure 30: Constraints on Ω_Λ and Ω_m (or $\Omega_0 = \Omega_\Lambda + \Omega_m$) in the ΛCDM+Ω_0 model from Planck and BAO data. The colored dots represent parameter values that fit Planck temperature C_ℓ and large scale polarization data, the color giving the value of H_0 required for the fit. The black contours (inner 68% and outer 95% confidence limits) give the models that remain allowed when Planck small-scale polarization data is also used; and blue contours when Planck CMB lensing data is used instead. The red contours show the effect of adding BAO data to break the Ω_0-Ω_Λ degeneracy. From the colors one can see that also independent H_0 data could be used to break the degeneracy. From [3].

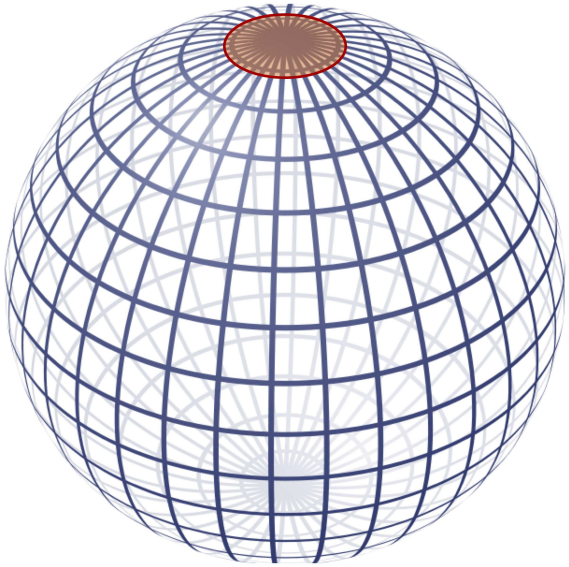


Figure 31: The upper limit $\Omega_0 < 1.003$ means that if we live in a closed universe, its curvature radius $R_{\rm curv} = H^{-1}/\sqrt{|\Omega_k|} > H^{-1}/\sqrt{0.003} = 18.3H^{-1} = 5.9d^c(z = 1090)$ is more than 5 times larger than the distance we can see (to the last scattering sphere).

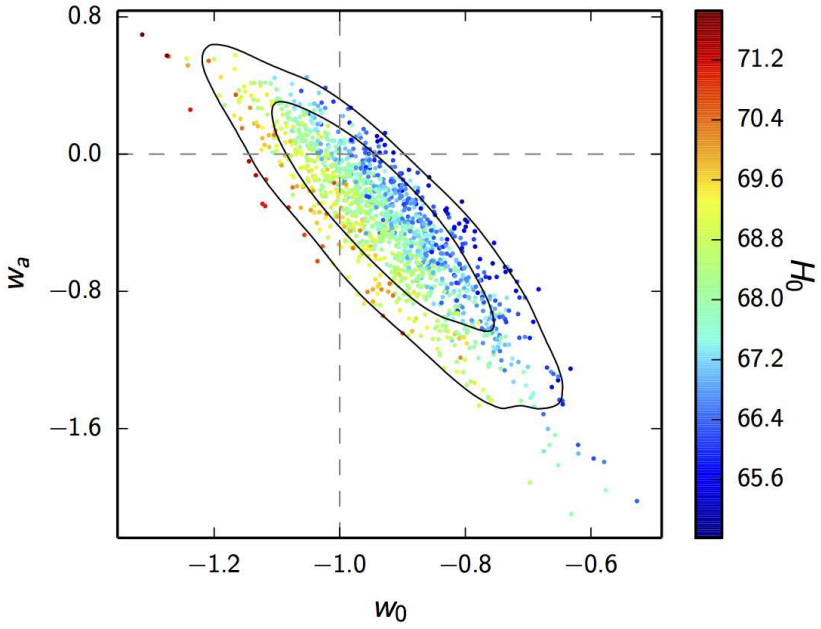


Figure 32: Constraints on the dark energy equation-of-state parameters w_0 and w_a (see text) in the 8-parameter $(w_0 + w_a)$CDM model from Planck, BAO, and SNIa data. From [3].

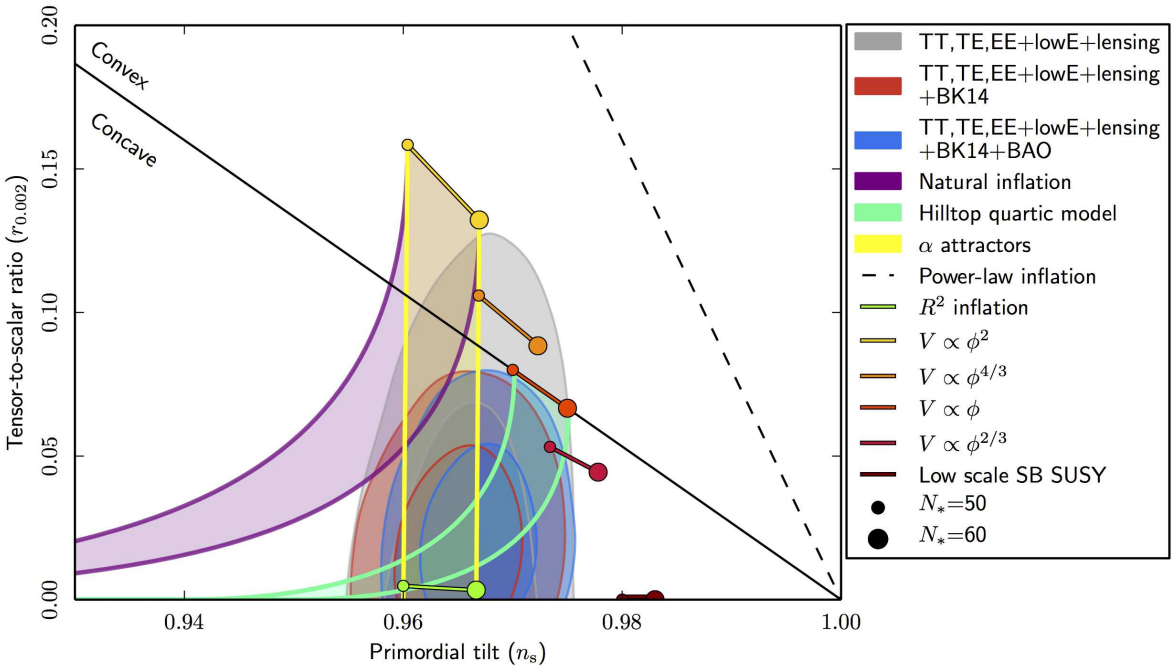


Figure 33: Constraints on the parameters n_s and r, which constrain inflation models, in the 7-parameter ΛCDM+r model from Planck data. Gray contours are based on Planck data only; red and blue contours include external data. Predictions from a selection of inflation models are marked on the plot. From [6].

oscillations before photon decoupling can be seen as a weak wavy pattern [13]. This is the same structure which we see in the C_ℓ but now much fainter, since now the baryons have fallen into the CDM potential wells, and the CDM was only mildly affected by these oscillations in the baryon-photon fluid. In this context these are called *baryon acoustic oscillations* (BAO). The half-wavelength of this pattern, however, corresponds to the same sound horizon distance $r_s(t_{\rm dec})$ in both cases.[17] But now the angular scale on the sky is related to it by the angular diameter distance $d_A^c(z)$ to the much smaller redshifts z of the galaxy survey. This $d_A^c(z)$ has then a different relation to Ω_0, Ω_Λ, and ω_m. Comparing CMB data to galaxy surveys gives us the ratio $d_A^c(z)/d_A^c(t_{\rm dec})$, which gives us independent information on these parameters.

The large scale structure surveys used for the BAO measurements to supplement Planck 2018 data were the 6dF Galaxy Survey (6dFGS) [14] and the Sloan Digital Sky Survey (SDSS) [15, 16].

Another way to break the geometric degeneracy, is to use the redshift-distance relationship from Supernova Type Ia (SNIa) surveys [17], or simply the distance-ladder determination of H_0, where Cepheids and Supernovae are the last two steps of the ladder. These were not used in the Planck 2018 analysis of 6- and 7-parameter models, because of the discrepancy with the local H_0 determination[18], and since the SNIa data adds little statistical power to the CMB+BAO combination, but the SNIa data was used for the following 8-parameter model.

To constrain properties of dark energy, the 7-parameter wCDM model is probably too simplistic, since it assumes that the equation-of-state parameter w stays constant during the epoch when dark energy has a significant effect on the expansion. To stay at a phenomenological level, i.e., not assuming a particular dark-energy model, but just attempting to constrain its equation of state, the next step is a two-parameter equation of state $w(a) = w_0 + w_a(1-a)$, i.e., a first-order Taylor expansion with w_0 the current value of w, and w_a related to its first derivative with respect to the scale factor, leading to an 8-parameter model. From Fig. 32 you can see that the best fits are near the ΛCDM values $w_0 = -1$, $w_a = 0$, but that the equation of state is poorly constrained.

Neutrino masses, i.e., the amount of hot dark matter, have a larger effect on large-scale structure than CMB; the CMB data is mainly needed to determine the other parameters after which the large-scale structure power spectrum $P_\delta(k)$ can be used to determine the sum of the neutrino masses. The value 0.06 eV used for the ΛCDM model is the minimum allowed by neutrino oscillation data.

The polarization pattern of the CMB on the sky can be divided into what are called E and B modes. This is analogous to the division of a vector field into irrotational (curl-free) and rotational (divergence-fee) parts. To first (i.e. linear) order in perturbation theory, only tensor perturbations produce B-mode polarization in the CMB. Only E-mode polarization has so far been detected in the CMB. Upper limits to CMB B-mode polarization provide upper limits to the tensor-scalar ratio r. Planck was not optimized for polarization measurements, so its B-mode measurement is noisy and suffers from instrument systematics. Thus the Planck upper limit to r from B modes is weak, $r < 0.41$, and the Planck constraint $r < 0.101$ comes from the effect of tensor perturbations on the CMB temperature C_ℓ. The ground-based BICEP2/Keck Array [18] at the South Pole can measure polarization more accurately, but it has limited sky coverage and needs to be combined with Planck data to separate the CMB from the foreground. This combination leads to the B-mode upper limit $r < 0.065$.

Since inflation produces tensor perturbations, and many inflation models predict that they should be strong enough to have an observable effect on the CMB, the simplest way to constrain

[17] To be accurate, the $t_{\rm dec}$ value to represent the effect in $P_\delta(k)$, is not exactly the same as for C_ℓ, since photon decoupling was not instantaneous, and in one we are looking at the effect on matter and in the other on photons.

[18] One should not combine discrepant data in parameter fitting. This would lead to artificially tight parameter values with poor fits to both data sets.

inflation is to fit the ΛCDM+r model to CMB data. From Fig. 33 you can see that the $V(\varphi) = \frac{1}{2}m^2\varphi^2$ inflation model, is ruled out by Planck data alone at a 95% confidence level (assuming that ΛCDM+r is the correct model for the universe).

References

[1] Planck Collaboration, *Planck 2013 results. I. Overview of products and scientific results*, arXiv:1303.5062, Astronomy & Astrophysics **571**, A1 (2014)

[2] Planck Collaboration, *Planck 2015 results. I. Overview of products and results*, arXiv:1502.01582, Astronomy & Astrophysics **594**, A1 (2016)

[3] Planck Collaboration, *Planck 2015 results. XIII. Cosmological parameters*, arXiv:1502.01589, Astronomy & Astrophysics **594**, A13 (2016)

[4] Planck Collaboration, *Planck 2018 results. VI. Cosmological parameters*, arXiv:1807.06209v1 (2018)

[5] Planck Collaboration, *Planck 2018 results. X. Constraints on inflation*, arXiv:1807.06211v1 (2018)

[6] A.R. Liddle and D.H. Lyth: Cosmological Inflation and Large-Scale Structure (Cambridge University Press 2000)

[7] S. Dodelson: Modern Cosmology (Academic Press 2003)

[8] J. Väliviita, *PhD thesis*, University of Helsinki 2005

[9] A.G. Riess et al., *New Parallaxes of Galactic Cepheids from Spatially Scanning the Hubble Space Telescope: Implications for the Hubble Constant*, arXiv:1801.01120

[10] A.G. Riess et al., *Milky Way Cepheid Standards for Measuring Cosmic Distances and Application to Gaia DR2: Implications for the Hubble Constant*, arXiv:1804.10655

[11] W.J. Percival et al., *Measuring the Baryon Acoustic Oscillation scale using the Sloan Digital Sky Survey and Galaxy Redshift Survey*

[12] F. Beutler et al., *The 6dF Galaxy Survey: baryon acoustic oscillations and the local Hubble constant*, MNRAS **416**, 3017–3032 (2011)

[13] A.J. Ross et al., *The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample.*, arXiv:1607.03155, MNRAS **470**, 2617 (2017)

[14] S. Alam et al., *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Releases 10 and 11 Galaxy samples*, MNRAS **441**, 24–62 (2014)

[15] M. Betoule et al., *Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples*, Astronomy & Astrophysics **568**, A22 (2014)

[16] Keck Array and BICEP2 Collaborations, *Improved Constraints on Cosmology and Foregrounds from BICEP2 and Keck Array Cosmic Microwave Background Data with Inclusion of 95 GHz Band*, Phys. Rev. Lett. **116**, 031302 (2016)

B Quantum Fluctuations during Inflation

Subhorizon scales during inflation are microscopic and therefore quantum effects are important. Thus we should study the behavior of the inflaton field using quantum field theory. To warm up we first consider quantum field theory of ascalar field in Minkowski space.

B.1 Vacuum fluctuations in Minkowski space

The field equation for a massive free (i.e. $V(\varphi) = \frac{1}{2}m^2\varphi^2$) real scalar field in Minkowski space is

$$\ddot{\varphi} - \nabla^2\varphi + m^2\varphi = 0\,, \tag{1}$$

or

$$\ddot{\varphi}_{\mathbf{k}} + E_k^2\varphi_{\mathbf{k}} = 0\,, \tag{2}$$

where $E_k^2 = k^2 + m^2$, for Fourier components. We recognize Eq. (2) as the equation for a harmonic oscillator. Thus each Fourier component of the field behaves as an independent harmonic oscillator.

In the quantum mechanical treatment of the harmonic oscillator one introduces the creation and annihilation operators, which raise and lower the energy state of the system. We can do the same here.

Now we have a different pair of creation and annihilation operators $\hat{a}^\dagger_{\mathbf{k}}$, $\hat{a}_{\mathbf{k}}$ for each Fourier mode $\mathbf{k}$. We denote the ground state of the system by $|0\rangle$, and call it the *vacuum*. As discussed earlier, *particles* are quanta of the oscillations of the field. The vacuum is a state with no particles. Operating on the vacuum with the creation operator $\hat{a}^\dagger_{\mathbf{k}}$, we add one quantum with momentum $\mathbf{k}$ and energy E_k to the system, i.e., we create one particle. We denote this state with one particle, whose momentum is $\mathbf{k}$ by $|1_{\mathbf{k}}\rangle$. Thus

$$\hat{a}^\dagger_{\mathbf{k}}|0\rangle = |1_{\mathbf{k}}\rangle\,. \tag{3}$$

This particle has a well-defined momentum $\mathbf{k}$, and therefore it is completely unlocalized (Heisenberg's uncertainty principle). The annihilation operator acting on the vacuum gives zero, i.e., not the vacuum state but the zero element of Hilbert space (the space of all quantum states),

$$\hat{a}_{\mathbf{k}}|0\rangle = 0\,. \tag{4}$$

We denote the hermitian conjugate of the vacuum state by $\langle 0|$. Thus

$$\langle 0|\hat{a}_{\mathbf{k}} = \langle 1_{\mathbf{k}}| \qquad \text{and} \qquad \langle 0|\hat{a}^\dagger_{\mathbf{k}} = 0\,. \tag{5}$$

The commutation relations of the creation and annihilation operators are

$$[\hat{a}^\dagger_{\mathbf{k}}, \hat{a}^\dagger_{\mathbf{k}'}] = [\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{k}'}] = 0, \qquad [\hat{a}_{\mathbf{k}}, \hat{a}^\dagger_{\mathbf{k}'}] = \delta_{\mathbf{k}\mathbf{k}'}\,. \tag{6}$$

When going from classical physics to quantum physics, classical observables are replaced by operators. One can then calculate expectation values for these observables using the operators. Here the classical observable

$$\varphi(t, \mathbf{x}) = \sum \varphi_{\mathbf{k}}(t)e^{i\mathbf{k}\cdot\mathbf{x}} \tag{7}$$

is replaced by the *field operator*

$$\hat{\varphi}(t, \mathbf{x}) = \sum \hat{\varphi}_{\mathbf{k}}(t)e^{i\mathbf{k}\cdot\mathbf{x}} \tag{8}$$

where[1]

$$\hat{\varphi}_{\mathbf{k}}(t) = w_k(t)\hat{a}_{\mathbf{k}} + w_k^*(t)\hat{a}^\dagger_{-\mathbf{k}} \tag{9}$$

and

$$w_k(t) = V^{-1/2}\frac{1}{\sqrt{2E_k}}e^{-iE_kt} \tag{10}$$

is the mode function, a normalized solution of the field equation (2). We are using the Heisenberg picture, i.e. we have time-dependent operators; the quantum states are time-independent.

Classically the ground state would be one where $\varphi = \text{const.} = 0$, but we know from the quantum mechanics of a harmonic oscillator that there are oscillations even in the ground state. Likewise, there are fluctuations of the scalar field, *vacuum fluctuations*, even in the vacuum state.

We shall now calculate the *power spectrum* of these vacuum fluctuations. The power spectrum is defined as the expectation value

$$\mathcal{P}_\varphi(k) = V\frac{k^3}{2\pi^2}\langle|\varphi_{\mathbf{k}}|^2\rangle \tag{11}$$

and it gives the variance of $\varphi(\mathbf{x})$ as

$$\langle\varphi(\mathbf{x})^2\rangle = \int_0^\infty \frac{dk}{k}\mathcal{P}_\varphi(k)\,. \tag{12}$$

For the vacuum state $|0\rangle$ the expectation value of $|\varphi_{\mathbf{k}}|^2$ is

$$\begin{aligned}\langle 0|\hat{\varphi}_{\mathbf{k}}\hat{\varphi}^\dagger_{\mathbf{k}}|0\rangle &= |w_k|^2\langle 0|\hat{a}_{\mathbf{k}}\hat{a}^\dagger_{\mathbf{k}}|0\rangle + w_k^2\langle 0|\hat{a}_{\mathbf{k}}\hat{a}_{-\mathbf{k}}|0\rangle + (w_k^*)^2\langle 0|\hat{a}^\dagger_{-\mathbf{k}}\hat{a}^\dagger_{\mathbf{k}}|0\rangle + |w_k|^2\langle 0|\hat{a}^\dagger_{-\mathbf{k}}\hat{a}_{-\mathbf{k}}|0\rangle \\ &= |w_k|^2\langle 1_{\mathbf{k}}|1_{\mathbf{k}}\rangle \;=\; |w_k|^2\,,\end{aligned} \tag{13}$$

since all but the first term give 0, and our states are normalized so that $\langle 1_{\mathbf{k}}|1_{\mathbf{k}'}\rangle = \delta_{\mathbf{k}\mathbf{k}'}$. From Eq. (10) we have $|w_k|^2 = 1/(2VE_k)$. Our main result is that

$$\mathcal{P}_\varphi(k) = V\frac{k^3}{2\pi^2}|w_k|^2 \tag{14}$$

for vacuum fluctuations, which we shall now apply to inflation, where the mode functions $w_k(t)$ are different.

B.2 Vacuum fluctuations during inflation

During inflation the field equation (for inflaton perturbations) is, from Sec. 8.6,

$$\delta\ddot{\varphi}_{\mathbf{k}} + 3H\delta\dot{\varphi}_{\mathbf{k}} + \left[\left(\frac{k}{a}\right)^2 + V''(\bar{\varphi})\right]\delta\varphi_{\mathbf{k}} = 0\,. \tag{15}$$

There are oscillations only in the perturbation $\delta\varphi$, the background $\bar{\varphi}$ is homogeneous and evolving slowly in time. For the particle point of view, the background solution represents the vacuum,[2] i.e., particles are quanta of oscillations around that value.

[1] We skip the detailed derivation of the field operator, which belongs to a course of quantum field theory. See, e.g., Peskin & Schroeder, section 2.3 (note different normalizations of operators and states, related to doing Fourier integrals rather than sums and considerations of Lorentz invariance.)

[2] This is not the vacuum state in the sense of being the ground state of the system. The true ground state has $\bar{\varphi}$ at the minimum of the potential. However there are no particles related to the background evolution $\bar{\varphi}(t)$.

After making the approximations $H = \text{const.}$ and

$$\frac{V''}{H^2} = 3\eta \approx 0 \tag{16}$$

we found that the two independent solutions for $\delta\varphi_{\mathbf{k}}(t)$ are

$$w_k(t) \;=\; V^{-1/2}\frac{H}{\sqrt{2k^3}}\left(i + \frac{k}{aH}\right)\exp\left(\frac{ik}{aH}\right) \tag{17}$$

and its complex conjugate $w_k^*(t)$, where the time dependence is in $a = a(t) \propto e^{Ht}$. The factor $V^{-1/2}H/\sqrt{2k^3}$ is here for normalization purposes ($V = L^3$ being the reference volume, not the inflaton potential).

When the scale k is well inside the horizon, $k \gg aH$, $\delta\varphi_{\mathbf{k}}(t)$ oscillates rapidly compared to the Hubble time H^{-1}. If we consider distance and time scales much smaller than the Hubble scale, spacetime curvature does not matter and things should behave like in Minkowski space. Considering Eq. (17) in this limit, one finds (**exercise**) that $w_k(t)$ indeed becomes equal to the Minkowski space mode function, Eq. (10). (We cleverly chose the normalization in Eq. (17) so that the normalizations would agree.) Therefore the $w_k(t)$ of Eq. (17) is our mode function. We can use it to follow the evolution of the mode functions as the scale approaches and exits the horizon.

The field operator for the inflaton perturbations is

$$\delta\hat{\varphi}_{\mathbf{k}}(t) = w_k(t)\hat{a}_{\mathbf{k}} + w_k^*(t)\hat{a}^\dagger_{-\mathbf{k}}\,, \tag{18}$$

and the power spectrum of inflaton fluctuations is

$$\mathcal{P}_\varphi(k) = V\frac{k^3}{2\pi^2}|w_k|^2\,. \tag{19}$$

Well before horizon exit, $k \gg aH$, observed during timescales $\ll H^{-1}$, the field operator $\delta\hat{\varphi}_{\mathbf{k}}(t)$ becomes the Minkowski space field operator and we have standard vacuum fluctuations in $\delta\varphi$.

Well after horizon exit, $k \ll aH$, the mode function becomes a constant

$$w_k(t) \;\to\; V^{-1/2}\frac{iH}{\sqrt{2k^3}}\,, \tag{20}$$

the vacuum fluctuations "freeze", and the power spectrum acquires the constant value

$$\mathcal{P}_\varphi(k) \;=\; V\frac{k^3}{2\pi^2}|w_k|^2 \;=\; \left(\frac{H}{2\pi}\right)^2\,. \tag{21}$$