Cosmology I

Hannu Kurki-Suonio

Fall 2020

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.

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.

I try to update these notes each time the course is lectured. This version of Cosmology I was updated in August 2019, for the coming fall 2019 course.

– Hannu Kurki-Suonio, August 2019

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 $c = 1$.[9] Velocity is thus a dimensionless quantity, and smaller than one[10] for massive objects. Energy and mass have now the same dimension, and Einstein's

[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 370-000-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] Most cosmological quantities are not known to better than 3-digit accuracy. In these notes I give more digits for many quantities, especially when they are known, so that round-off errors do not accumulate if these quantities are used in calculations.

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

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 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.[11]

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, $k_B = 1.380\,649 \times 10^{-23}$ 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 $1\,\mathrm{K} = 1.380\,649 \times 10^{-23}$ J, or

$$1\,\mathrm{eV} = 11\,604.5\,\mathrm{K} = 1.782\,662 \times 10^{-36}\,\mathrm{kg} = 1.602\,177 \times 10^{-19}\,\mathrm{J}. \tag{1}$$

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. Since $h = 6.626\,070\,15 \times 10^{-34}$ Js in SI units, we have

$$\begin{aligned} 1\,\mathrm{kg} &= 2.842\,788 \times 10^{42}\,\mathrm{m}^{-1} = 8.522\,465 \times 10^{50}\,\mathrm{s}^{-1} \\ 1\,\mathrm{eV} &= 5\,067\,730.7\,\mathrm{m}^{-1} = 1.519\,267 \times 10^{15}\,\mathrm{s}^{-1}. \end{aligned} \tag{2}$$

A useful relation to remember is

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

(more precisely, the number is 197.327), 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, $\mathrm{M}_\odot = 1.988\,48 \pm 9 \times 10^{30}$ kg [1],[12] and a common unit of length is parsec, $1\,\mathrm{pc} = 3.261\,564$ light years $= 3.085\,678 \times 10^{16}$ 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 $1\,\mathrm{Mpc} = 10^6$ pc, which is a typical distance between neighboring galaxies. For angles, 1 degree $(1^\circ) = 60$ arcminutes $(60') = 3600$ arcseconds $(3600'')$.

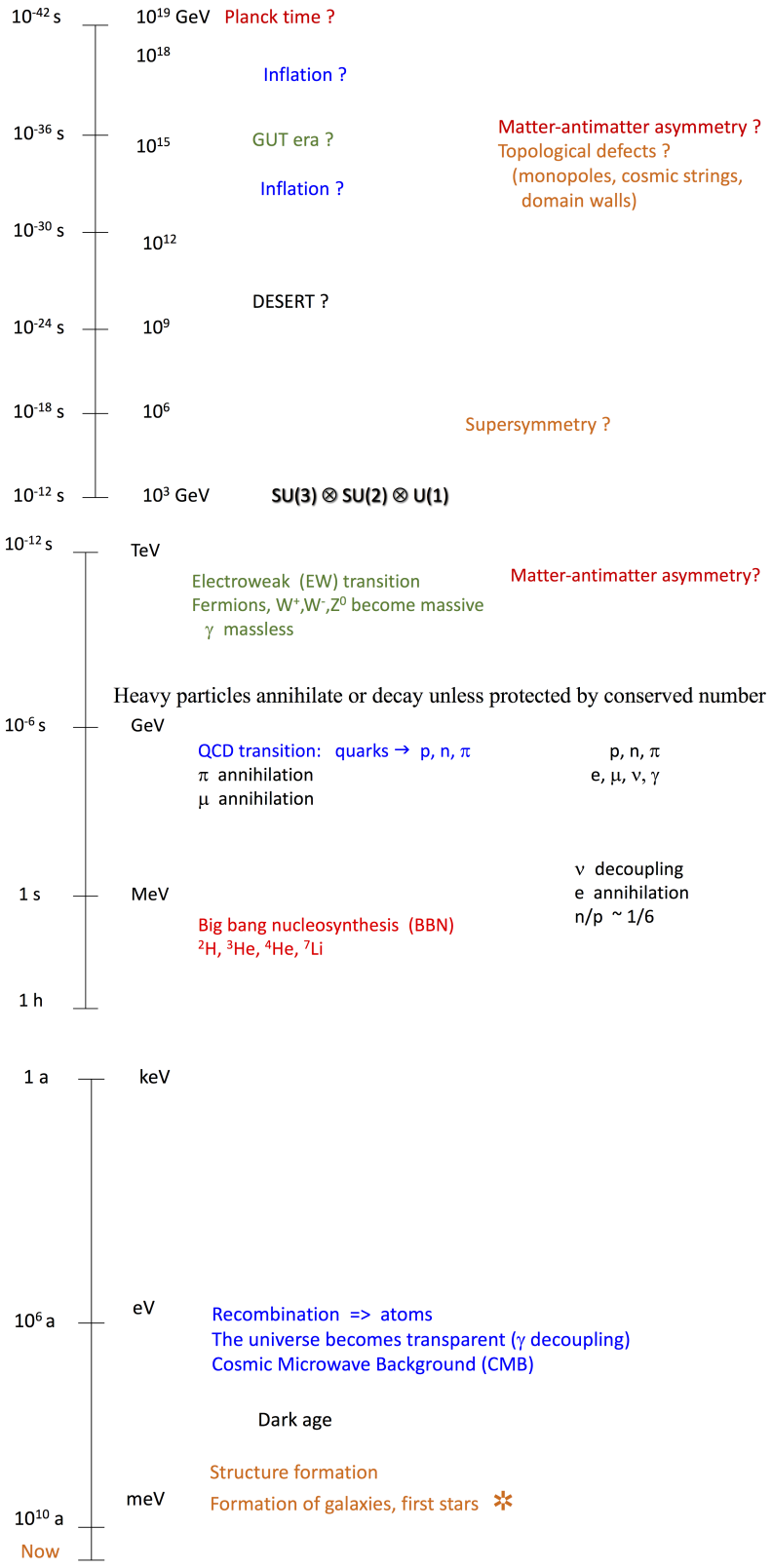

Figure 1: Short history of the universe.

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 GeV (about 10^{15} K), which was the case when the universe was younger than 10^{-11} s. Then the primordial soup of particles consisted of free massless fermions (quarks 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*[13] in the early universe. In the electroweak (EW) 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 transition.[14] The mass is due to the interaction of the particle with the Higgs field. The EW transition took place when the universe cooled below the critical temperature $T_c \sim 100$ GeV of the transition at $t \sim 10^{-11}$ 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 $t \sim 10^{-5}$ s. The critical temperature of the QCD transition is $T_c \sim 150$ 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,[15] so the last annihilation event was the electron-

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

[12] In my notation, uncertainties (here ±9) refer always to the last given digits, so here I mean $\mathrm{M}_\odot = (1.988\,48 \pm 0.000\,09) \times 10^{30}$ kg.

[13] 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.

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

[15] 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.

positron annihilation which took place when $T \sim m_e \sim 0.5\,\mathrm{MeV}$ and $t \sim 1$ 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, $T \sim 100$ 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 370 000 years old. At a temperature $T \sim 3000\,\mathrm{K}$ ($\sim 0.25\,\mathrm{eV}$), 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[16] can be observed. Thus its mass can be estimated and it turns

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

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 is too thin to be observable from here, also contain significant amounts of it; so this 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[17]. 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. 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).[18]

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,

[17] 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.

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

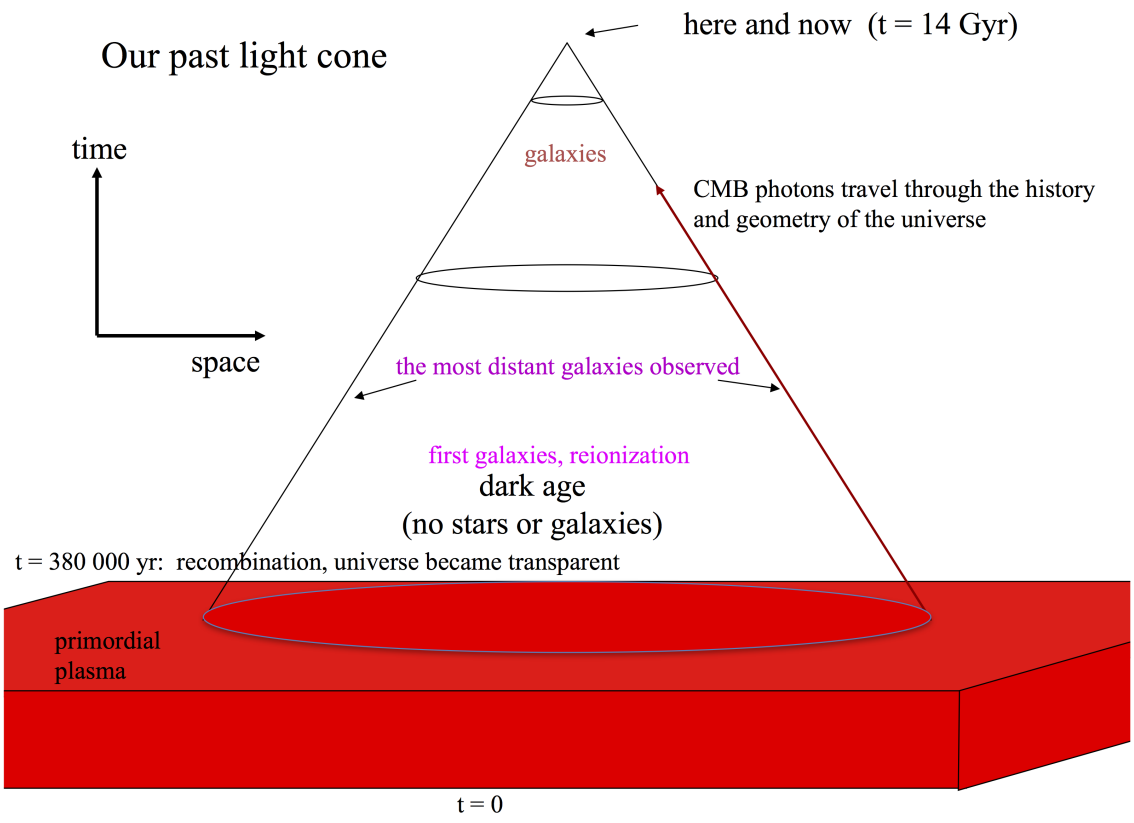

Figure 2: Our past light cone.

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.

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.

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".[19]

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

[19] 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.

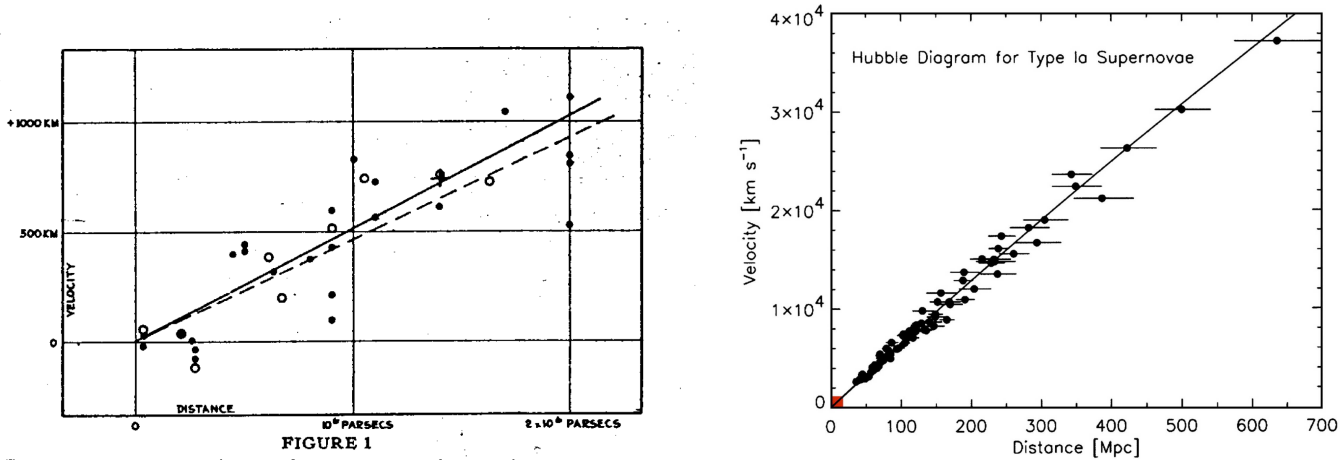

Figure 3: Left: The original Hubble diagram by Hubble. Right: A modern Hubble diagram (R.P. Kishner, PNAS 101, 8 (2004)).

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.

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.[20]

1.7.3 Redshift and the Hubble law

Modern cosmology originated from the observation by Edwin Hubble[21] (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

[20]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.

[21]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.

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 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

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

Still in the 1980s different observers reported values ranging from 50 to 100 km/s/Mpc ($h = 0.5$ to 1).[22]

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 [3]

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

Modern observations have narrowed down the range and some recent results are

$$\begin{aligned} H_0 &= 74.0 \pm 1.4\,\mathrm{km/s/Mpc} \qquad [4] \\ H_0 &= 69.8 \pm 1.9\,\mathrm{km/s/Mpc} \qquad [5] \end{aligned} \tag{9}$$

($h = 0.740 \pm 0.014$ or $h = 0.698 \pm 0.019$). Here the uncertainties (± 1.4 and ± 1.9) 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.) As we see, results from different observers are not all entirely consistent, so that the contribution of systematic effects to the probable error may have been underestimated.[23] 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[24] (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).

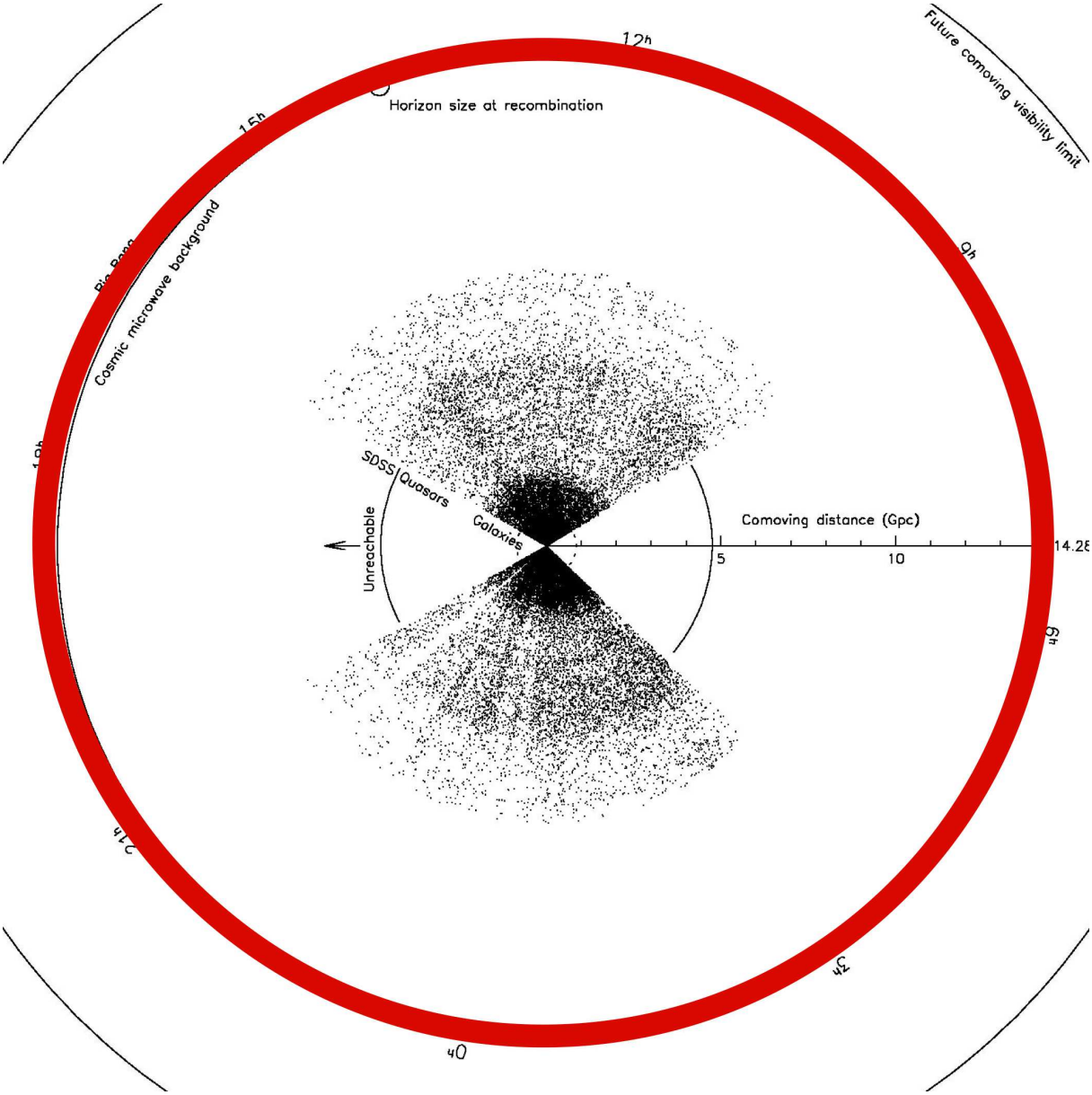

Figure 4: The distribution of galaxies from the Sloan Digital Sky Survey (SDSS) and the horizon. We are at the center of this diagram. Each dot represents an observed galaxy. The empty sectors are regions not surveyed. The figure shows fewer galaxies further out, since only the brightest galaxies can 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) [2].

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 $d_{\rm hor}$ 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., $d_{\rm hor} = d(z = \infty)$. 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 $d(z = 1090)$ 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[25] ; 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

[22] 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.

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

[24] www.jwst.nasa.gov

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

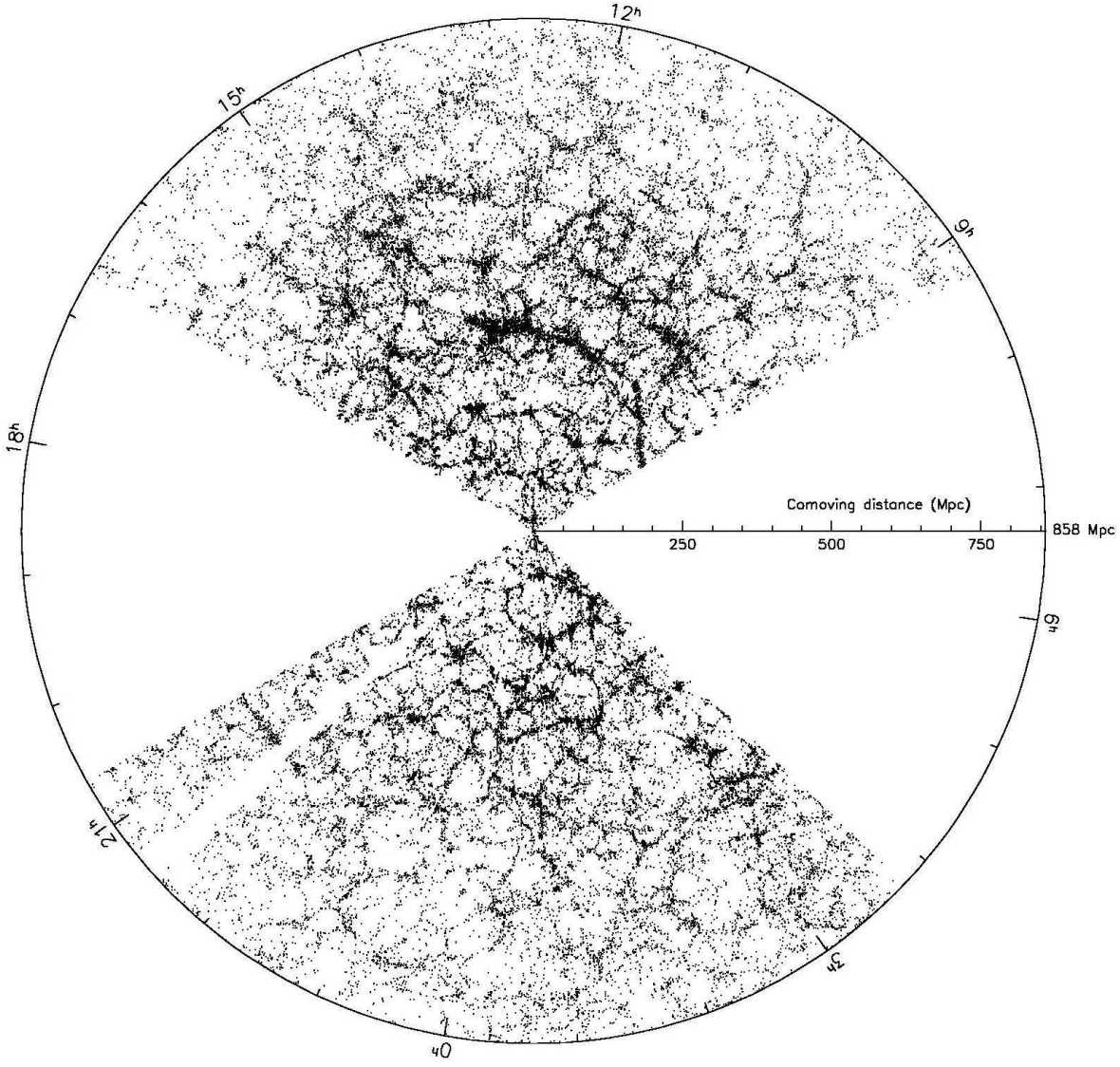

Figure 5: The distribution of galaxies from SDSS. This figure shows observed galaxies that are within 2° of the equator and closer than 858 Mpc. The empty sectors are regions not surveyed. Figure from Gott et al: "Map of the Universe" (2005) [2].

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 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 $5.9\,\mu$eV, 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 $T_0 = 2.7255 \pm 0.0006\,\mathrm{K}$ [7]. 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 $3362.1 \pm 1.0\,\mu$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) [8]. 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

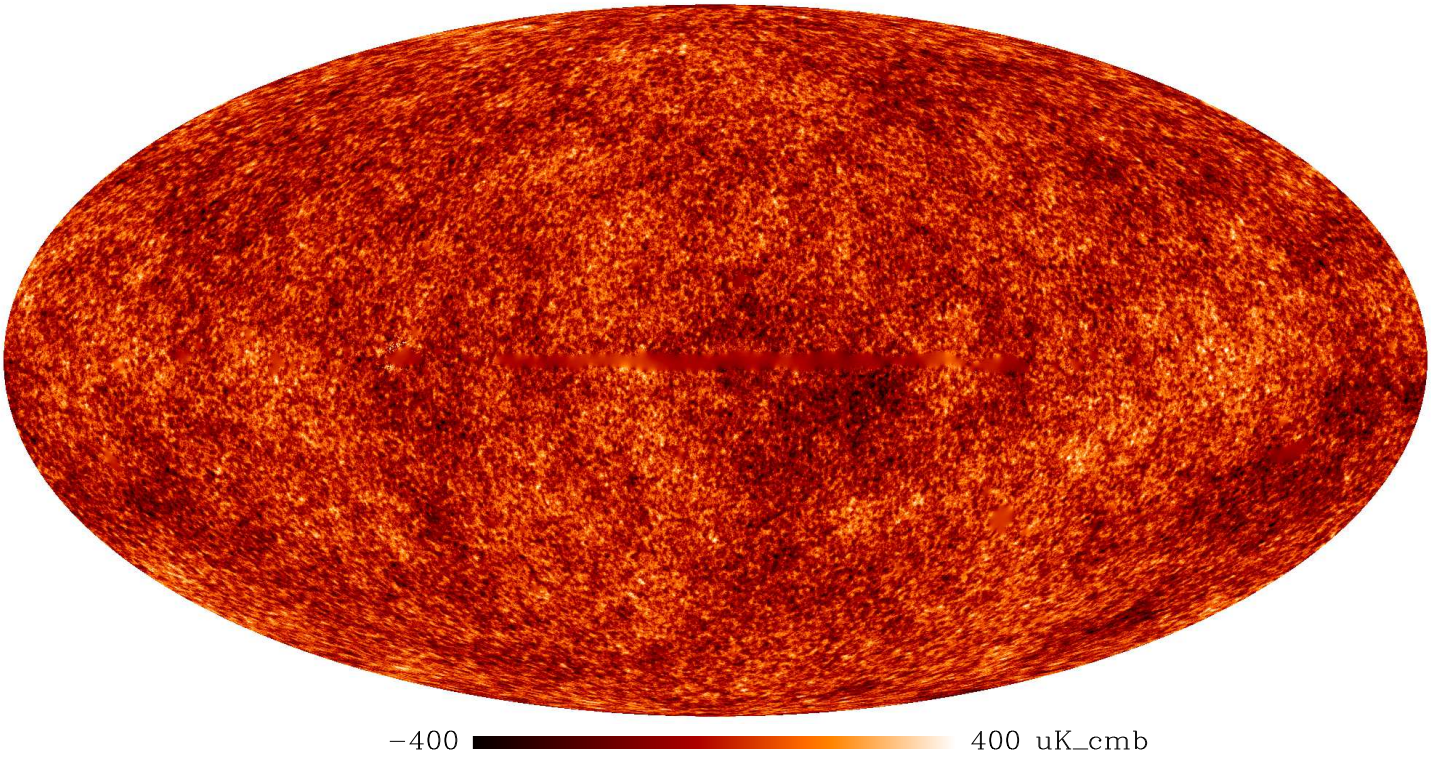

Figure 6: The cosmic microwave background. This figure shows the CMB temperature variations over the entire sky. The color scale shows deviations of -400μK to $+400\mu$K from the average temperature of 2.7255 K. The plane of the milky way is horizontally in the middle. The fuzzy regions are those where the CMB is obscured by our galaxy, or nearby galaxies (can you find the LMC?).

nearby galaxy clusters[26].

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.[27] 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

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

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

[26] 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*.

[27] 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.

$m = M$ (**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 $r_\odot = 7 \times 10^8$ 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 $t_\odot = 4.6 \times 10^9$ a ago. What fraction f of the sky do theycover? What is the energy density of starlight in the universe, in kg/m^3? (The luminosity, or radiated power, of the sun is $L_\odot = 3.85 \times 10^{26}$ W).
4. Calculate $r_{1/2}$ and f for galaxies, using $n_G = 3 \times 10^{-3}\,\mathrm{Mpc}^{-3}$, $r_G = 10$ kpc, and $t_G = 10^{10}$ 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 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 kg/m^3 and the photon density in units of m^{-3}, i) today, when $T = 2.7255\,\mathrm{K}$, 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 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 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] Particle Data Group, *Review of particle physics*, Phys. Rev. D **98**, 030001 (2018), p. 128: 2. Astrophysical Constants and Parameters

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

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

[4] A.G. Riess et al., *Large Magellanic Cloud Cepheid Standards Provide a 1% Foundation for the Determination of the Hubble Constant and Stronger Evidence for Physics beyond ΛCDM*, Astrophys. J. **876**, 85 (2019)

[5] W.L. Freedman et al., *The Carnegie-Chicago Hubble Program. VIII. An Independent Determination of the Hubble Constant Based on the Tip of the Red Giant Branch*, arXiv:1907.05922v1 (2019)

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

[7] D.J. Fixsen, *The Temperature of the Cosmic Microwave Background*, Astrophys. J. **707**, 916 (2009)

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

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.

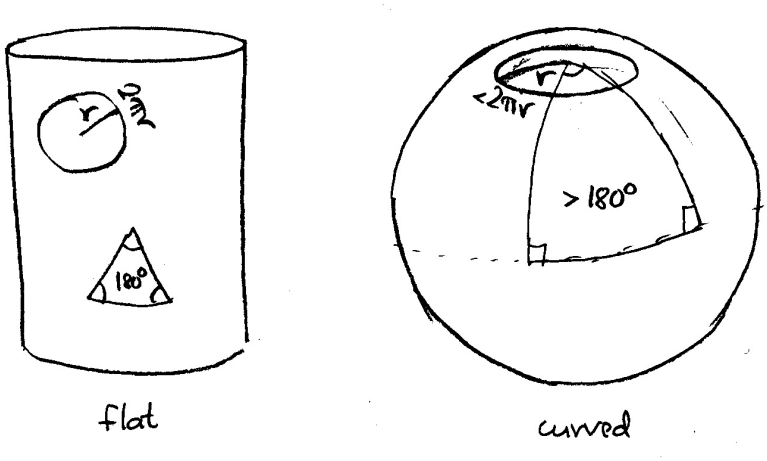

Figure 1: Cylinder and sphere.

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

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

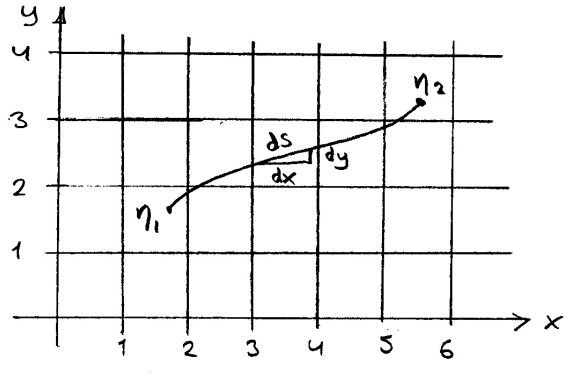

Figure 2: A parameterized curve in Euclidean 2D space with Cartesian coordinates.

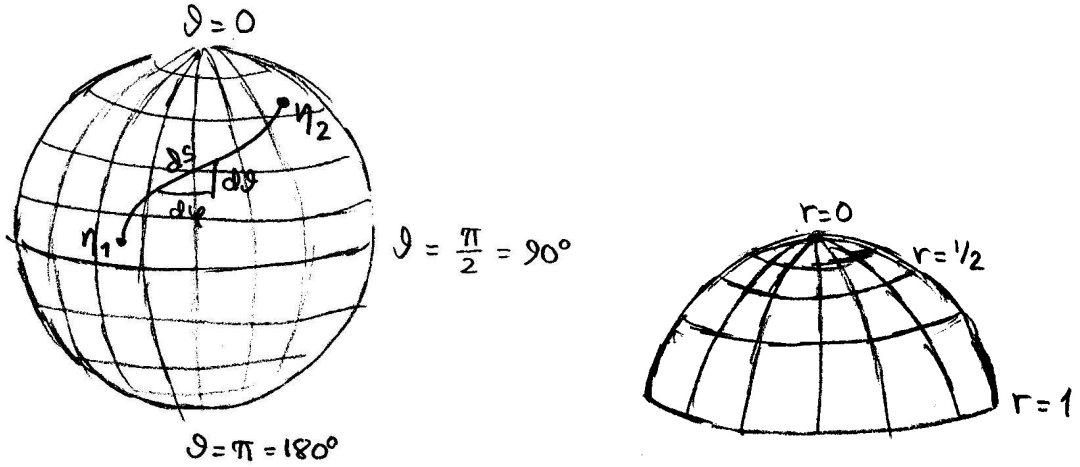

Figure 3: Left: A parameterized curve on a 2D sphere with spherical coordinates. Right: The part of the sphere covered by the coordinates in Eq. (10).

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

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

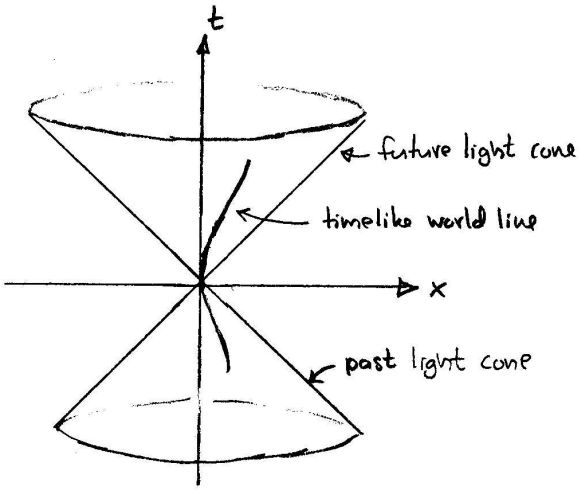

Figure 4: The light cone.

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 (the 0, 1, 2, 3 here are indices, not exponents). 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 e 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).

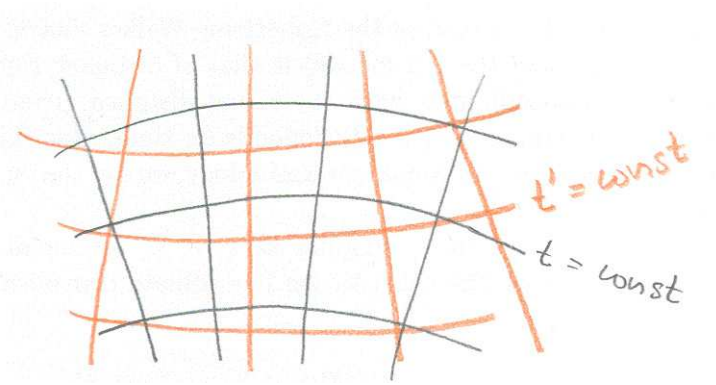

Figure 5: Two coordinate systems with different time slicings.

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

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

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.

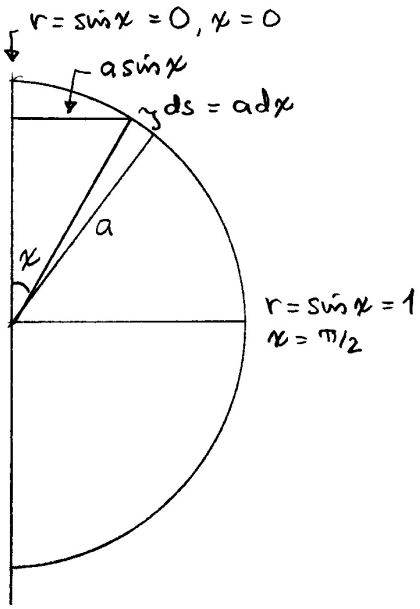

Figure 1: The hypersphere. This figure is for $K = k = 1$. Consider the semicircle in the figure. It corresponds to χ ranging from 0 to π. You get the (2-dimensional) sphere by rotating this semicircle off the paper around the vertical axis by an angle $\Delta\varphi = 2\pi$. You get the (3-dimensional) 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

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

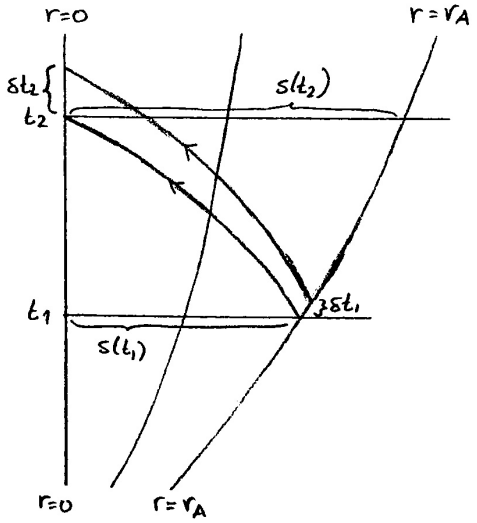

Figure 2: The two light rays to establish the redshift.

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 $t=t_0$. (More properly, z is a coordinate on our past light cone, so it corresponds to both a time and a distance.)

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 $ds^2 = a^2(t)d\chi^2 = a^2(t)\,\dfrac{dr^2}{1-Kr^2}$, 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,

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

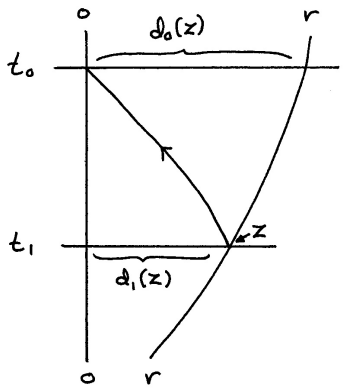

Figure 3: Calculation of the distance-redshift relation.

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 $t =$

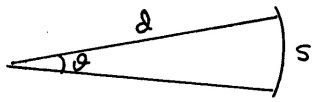

Figure 4: Defining the angular diameter distance.

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 $ar\,d\vartheta \times ar\sin\vartheta\, d\varphi$, so that the proper volume element becomes

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

and the comoving volume is

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

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

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

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 compared 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. $G = 6.674\,30 \pm 15 \times 10^{-11}\,\mathrm{m}^3/\mathrm{kgs}^2$ 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

[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 ρ_{vac}, which has the same effect.

be derived from the Einstein equation. In the present case this becomes the *energy continuity equation*

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

(**Exercise**: Derive this from the Friedmann equations!) Since isotropy requires that the fluid is at rest, there is no corresponding 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 $K/a^2 = 1/R_{\rm curv}^2$ 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

$$\begin{aligned}\rho_{\rm cr0} \;&\equiv\; \rho_{\rm cr}(t_0) \;\equiv\; \frac{3H_0^2}{8\pi G} \;=\; 1.878\,342\times 10^{-26} h^2\,{\rm kg/m^3} \\ &=\; 10.5367 h^2\,{\rm GeV/m^3} \;=\; 2.775\times 10^{11} h^2\,{\rm M_\odot/Mpc^3}\,.\end{aligned} \tag{63}$$

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* if 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 $w_i = const$,

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

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 Universe 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.)

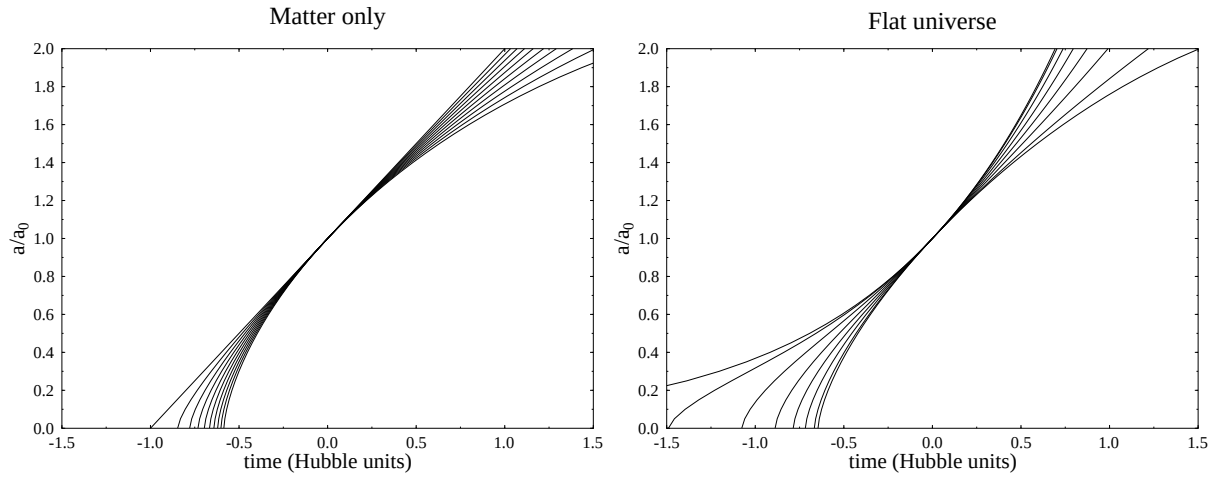

Figure 5: The expansion of the universe $a(t)$ for a) the matter-only universe $\Omega_\Lambda = 0$, $\Omega_m = 0$, 0.2,...,1.8 (from top to bottom) b) the flat universe $\Omega_0 = 1$ ($\Omega_\Lambda = 1 - \Omega_m$), $\Omega_m = 0$, 0.05, 0.2, 0.4, 0.6, 0.8, 1.0, 1.05 (from top to bottom). The time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

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

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

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

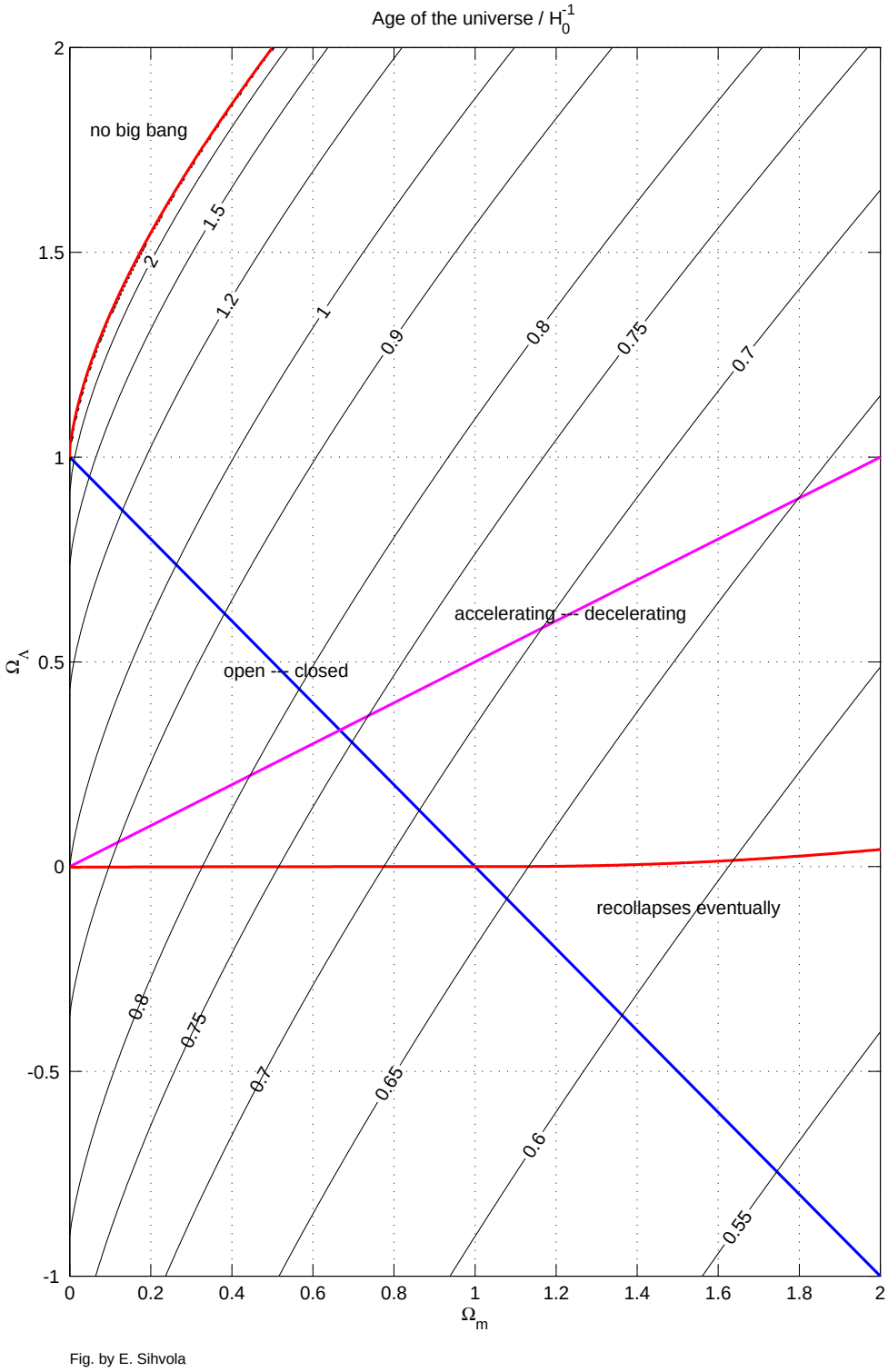

Figure 6: The age of the universe as a function of Ω_m and Ω_Λ.

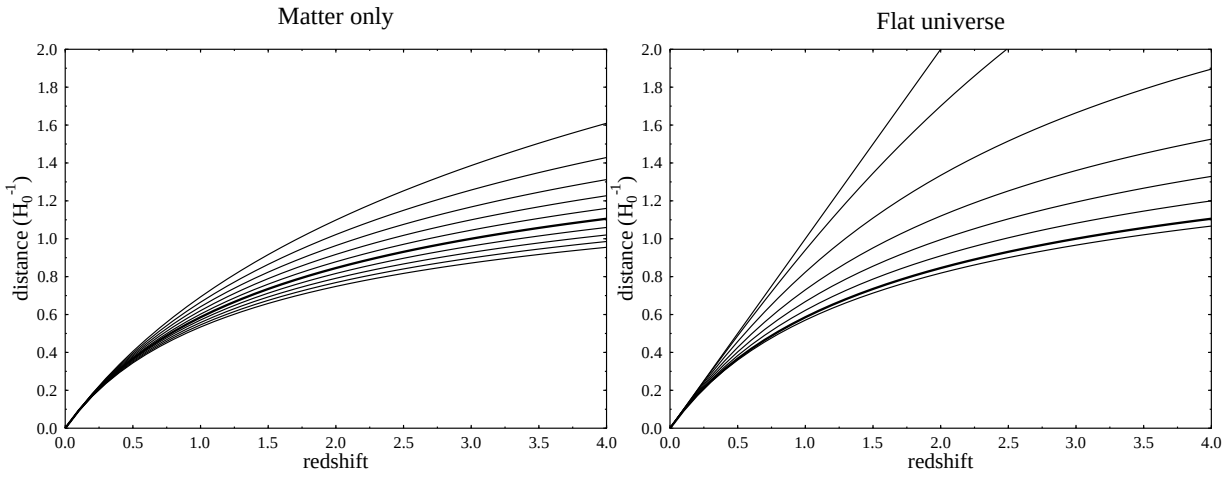

Figure 7: The distance-redshift relation, Eq. (116), for a) the matter-only universe $\Omega_\Lambda = 0$, $\Omega_m = 0$, 0.2,. . . ,1.8 (from top to bottom) b) the flat universe $\Omega_0 = 1$ ($\Omega_\Lambda = 1 - \Omega_m$), $\Omega_m = 0$, 0.05, 0.2, 0.4, 0.6, 0.8, 1.0, 1.05 (from top to bottom).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

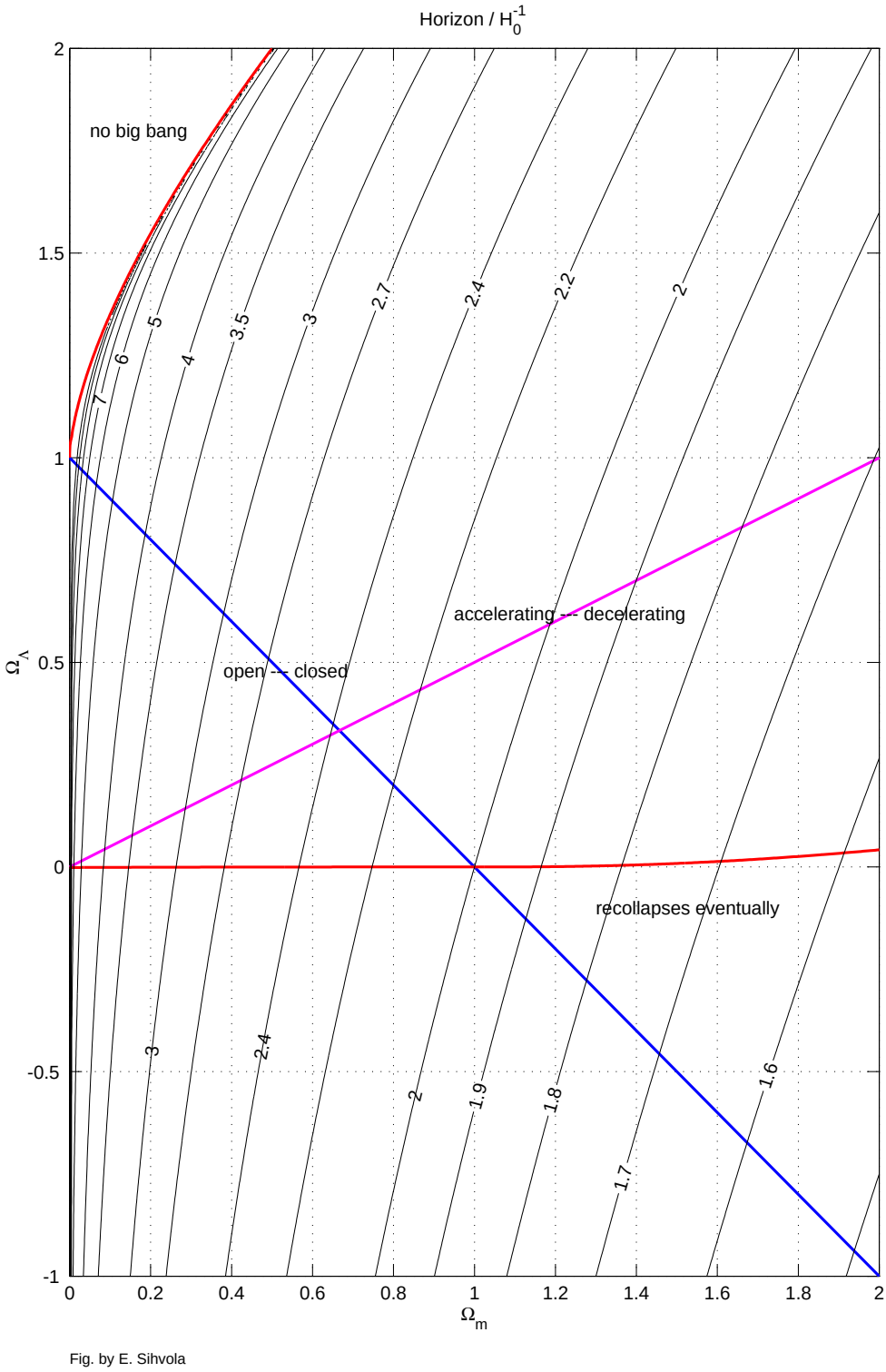

Figure 8: The horizon as a function of Ω_m and Ω_Λ.

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

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

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

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

(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.

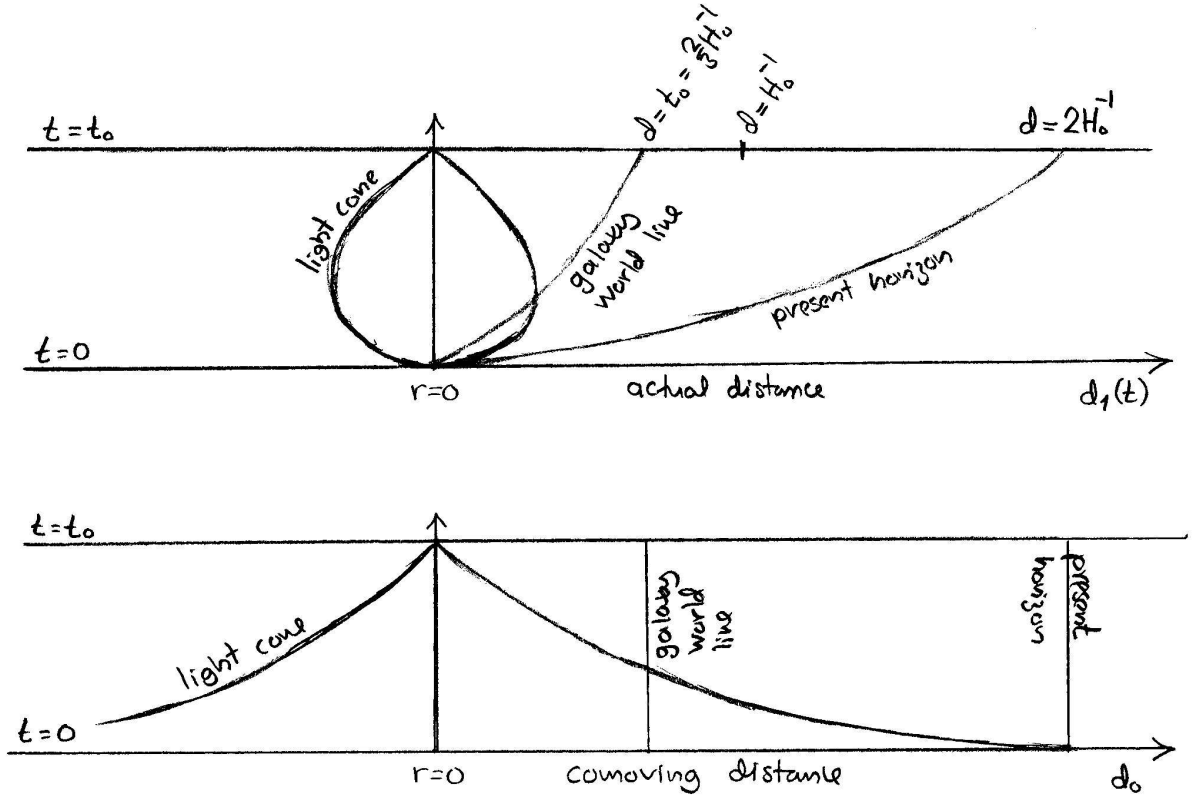

Figure 9: Spacetime diagrams for a flat matter-dominated universe giving a) the proper distance (denoted here as $d_1(t)$ b) the comoving distance (denoted d_0) from origin as a function of cosmic time.

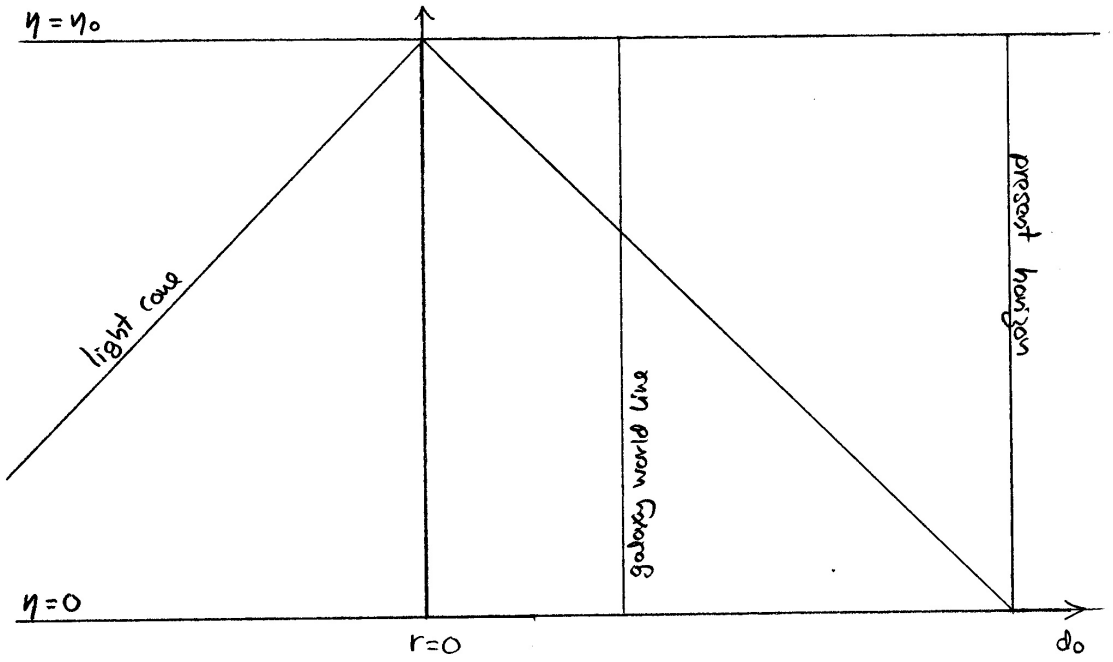

Figure 10: Spacetime diagram for a flat matter-dominated universe in conformal coordinates.

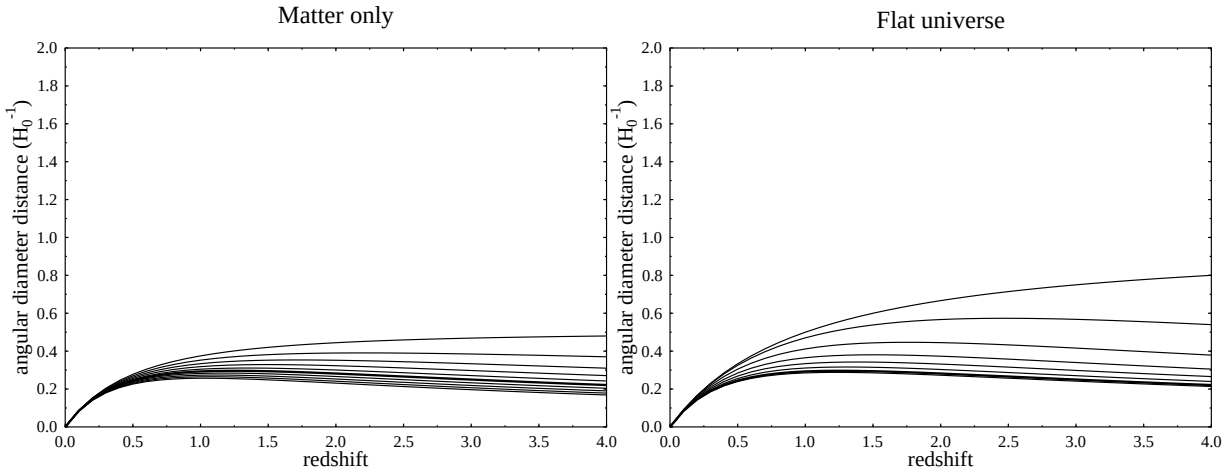

Figure 11: The angular diameter distance -redshift relation, Eq. (129), for a) the matter-only universe $\Omega_\Lambda = 0$, $\Omega_m = 0$, 0.2,...,1.8 (from top to bottom) b) the flat universe $\Omega_0 = 1$ ($\Omega_\Lambda = 1 - \Omega_m$), $\Omega_m = 0$, 0.05, 0.2, 0.4, 0.6, 0.8, 1.0, 1.05 (from 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.

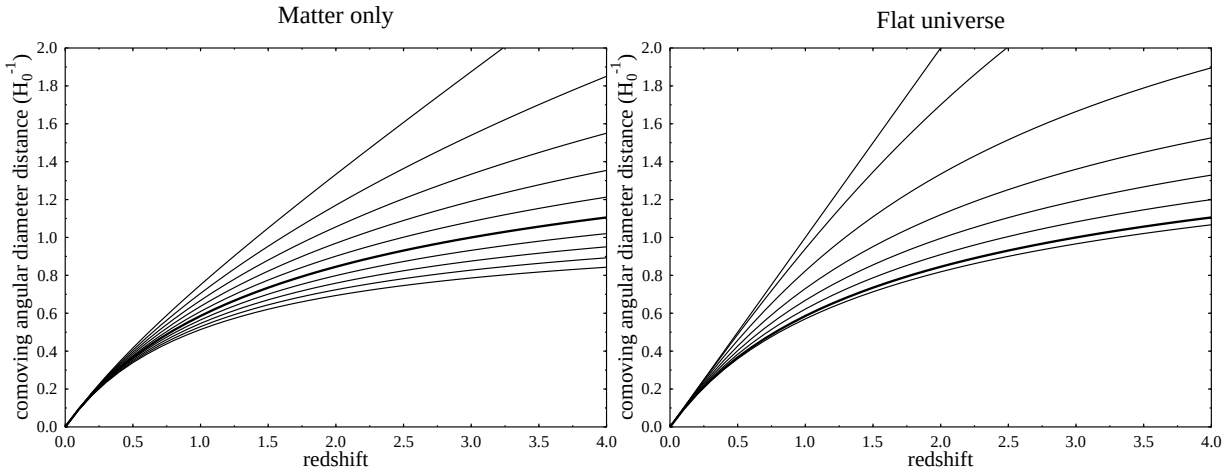

Figure 12: Same as Fig. 11, bur for the *comoving* angular diameter distance. Now the expansion effect is eliminated. For the closed models (for $\Omega_m > 1$ in the case of $\Omega_\Lambda = 0$) even the comoving angular diameter distance may begin to decrease at large enough redshifts. This happens 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

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

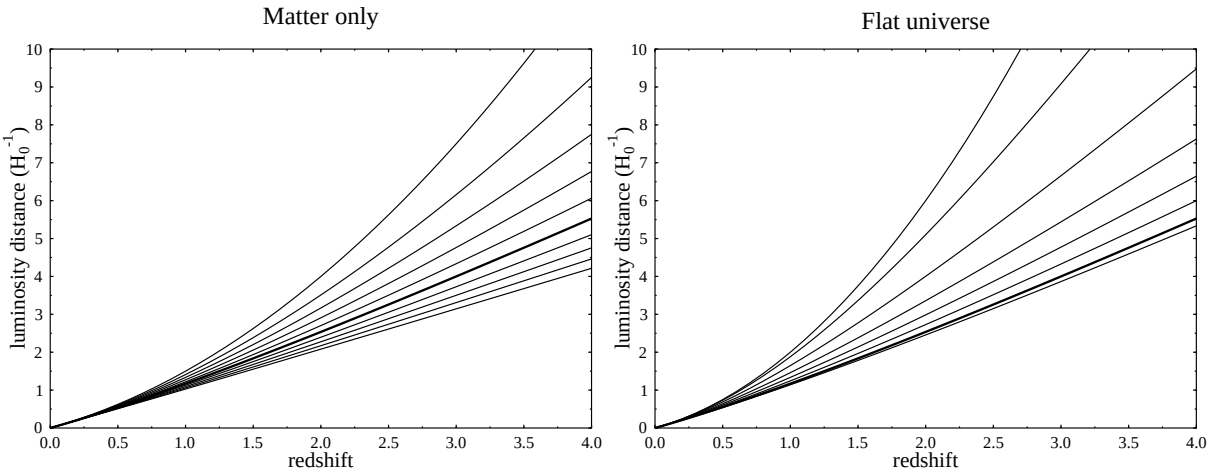

Figure 13: Same as Fig. 11, bur for the *luminosity* distance. Note how the vertical scale now extends to 10 Hubble distances instead of just 2, to have room for the much more rapidly increasing luminosity distance.

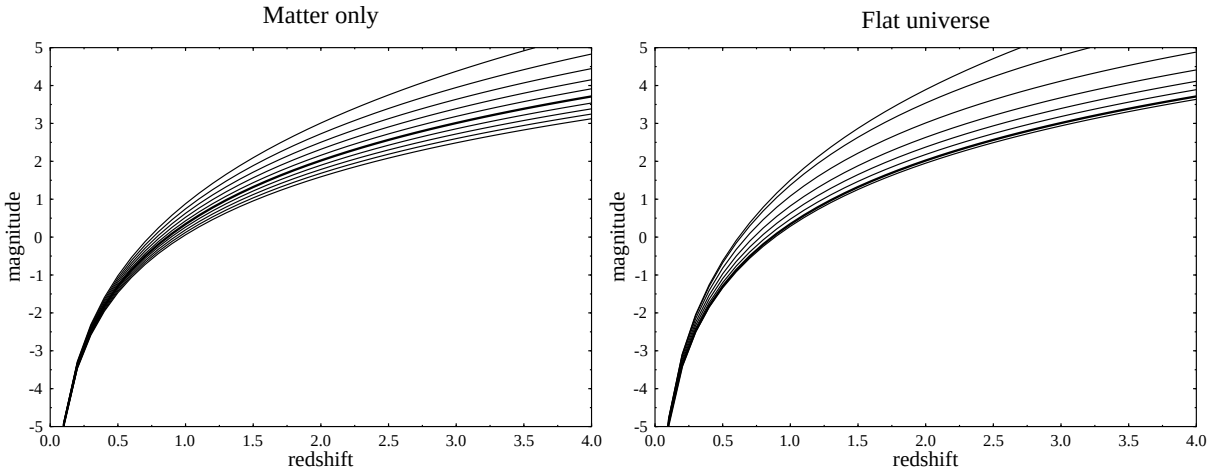

Figure 14: Same as Fig. 11, bur for the *magnitude-redshift* relation. The constant $M - 5 - 5 \lg H_0$ in Eq. (134), which is different for different classes of standard candles, has been arbitrarily set to 0.

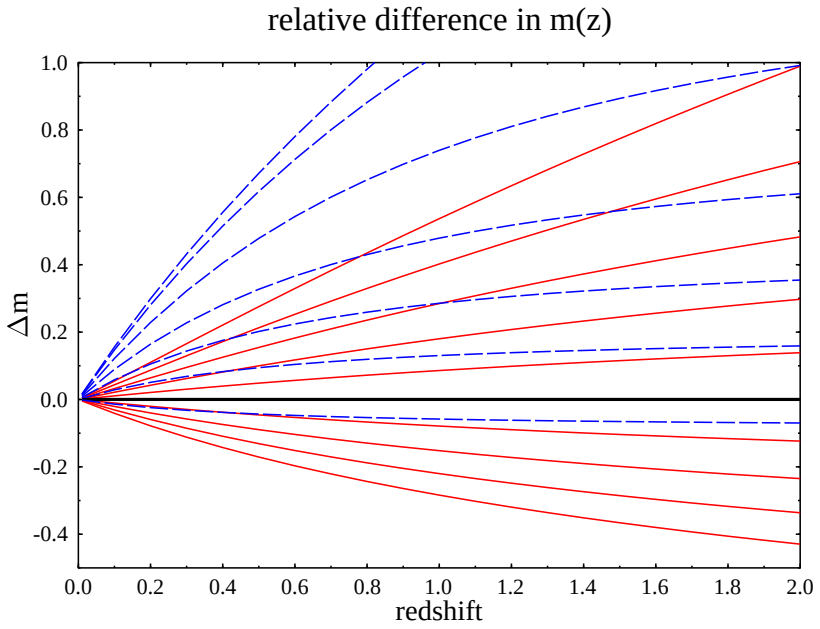

Figure 15: The difference between the magnitude-redshift relation of the different models in Fig. 14 from the reference model $\Omega_m = 1$, $\Omega_\Lambda = 0$ (which appears as the horizontal thick line). The red (solid) lines are for the matter-only ($\Omega_\Lambda = 0$) models and the blue (dashed)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

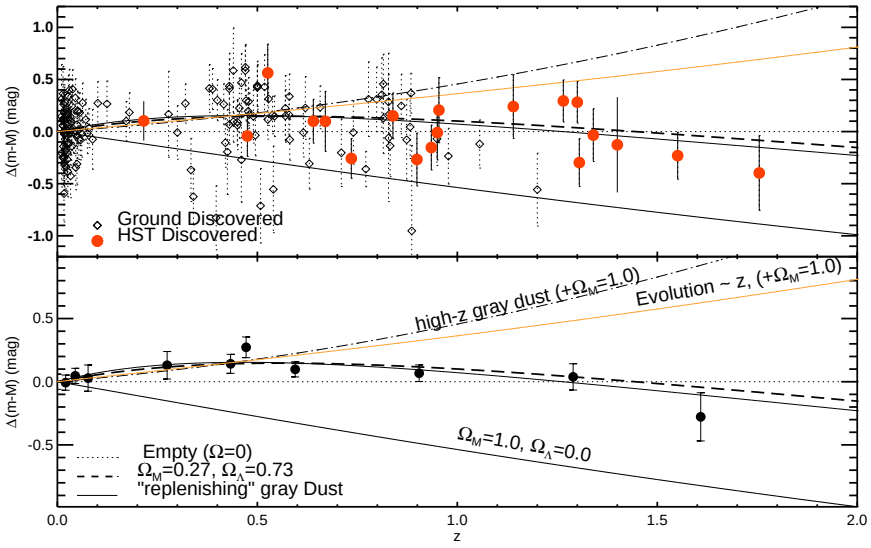

Figure 16: Supernova Ia luminosity-redshift data. The top panel shows all supernova of the data set. The bottom panel show the averages from different redshift bins. The curves corresponds to three different FRW cosmologies, and some alternative explanations: "dust" refers to the possibility that the 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].

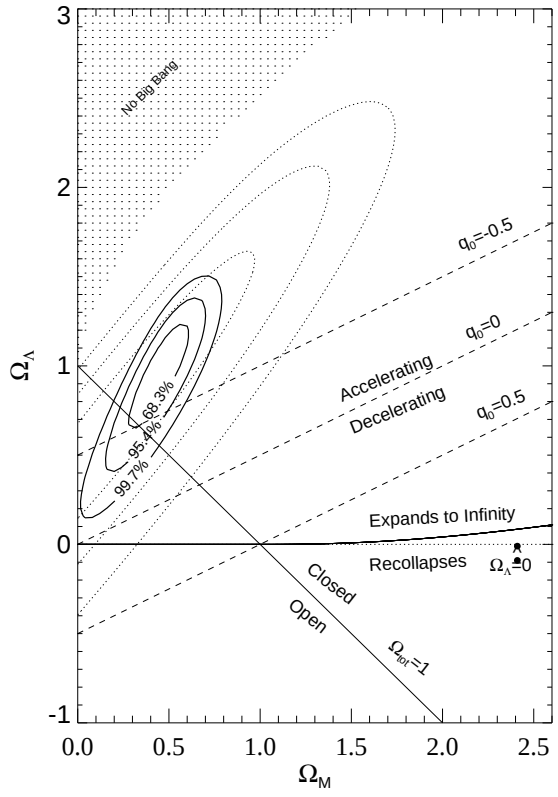

Figure 17: Ω_m and Ω_Λ determined from the Supernova Ia data. The dotted contours are the old 1998 results[3]. From Riess et al., astro-ph/0402512 [6].

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

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

The vacuum and matter energy densities become equal later, when

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

Here

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

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

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

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. The (comoving) distance to the horizon is

$$d^c_{\rm hor} = H_0^{-1}\int_0^1 \frac{da}{\sqrt{\Omega_\Lambda a^4 + \Omega_m a}}\,. \tag{148}$$

which for $\Omega_\Lambda = 0.7$ and $\Omega_m = 0.3$ gives $d^c_{\rm hor} = 3.305H_0^{-1}$.

These results are modified somewhat by inclusion of the effect of radiation ($\Omega_r \approx 10^{-4}$), which requires integrating also $t(a)$ numerically.

Planck 2018 best-fit concordance model. The best-fit ΛCDM model to the final release of Planck satellite data (we discuss this in Cosmology II; see p. 14, Table I, `Plik` best-fit in [9]) has

$$\Omega_m = 0.3158 \qquad \Omega_r = 9.232\times10^{-5} \qquad H_0 = 67.32\,\mathrm{km/s/Mpc} \tag{149}$$

(and $\Omega_\Lambda = 1 - \Omega_m - \Omega_r$).[22] The Hubble distance and time are then

$$H_0^{-1} = 4453.2\,\mathrm{Mpc} = 14.525\times10^9\,\mathrm{yr}\,. \tag{150}$$

[22] We cheat here a little bit, by using the value of Ω_r corresponding to three massless neutrino species; whereas the Planck best-fit model assumed one massive neutrino species with $m_\nu = 0.06\,\mathrm{eV}$, so that this neutrino species counts today as matter (and is included in the $\Omega_m = 0.3158$, where it contributes about 0.0014). A truly accurate treatment would include the change of this species from radiation into matter as the universe expands. It is more accurate to include it also in Ω_r at all times than not to include it there at all, since Ω_r is much more important at early times.

Including the effect of Ω_r, we need to integrate also the age-redshift relation numerically. The age of the universe is

$$t_0 = 0.9499 H_0^{-1} = 13.797\,\mathrm{Gyr}\,. \tag{151}$$

For the redshift when acceleration began, one now has to solve a fourth-order equation (done conveniently with WolframAlpha), to get

$$a_{\rm acc} = \frac{1}{1+z_{\rm acc}} = 0.6136 \qquad \text{and} \qquad t_{\rm acc} = 0.5306 H_0^{-1} = 7.707\,\mathrm{Gyr}\,, \tag{152}$$

i.e., the expansion began to accelarate 6.09 billion years ago. Matter density was equal to vacuum-energy density at

$$a_{\rm eq} = \frac{1}{1+z_{\rm eq}} = 0.7729 \qquad \text{and} \qquad t_{\rm eq} = 0.7100 H_0^{-1} = 10.313\,\mathrm{Gyr}\,, \tag{153}$$

i.e., 3.48 billion years ago. The (comoving) horizon distance (to $z=\infty$) is

$$d_{\rm hor} = 3.1767 H_0^{-1} = 14.146\,\mathrm{Mpc} = 46.14\times 10^9\,\text{light years}\,. \tag{154}$$

The comoving distance to the last scattering sphere ($z=1090$) is

$$d_{\rm ls} = 3.1138 H_0^{-1} = 13.867\,\mathrm{Mpc} = 45.23\times 10^9\,\text{light years}\,. \tag{155}$$

We provide a table of different quantities (age and comoving distance) as a function of redshift for the Planck 2013 (the first data release) best-fit model in Table 1. 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 reason for giving them with so many digits in the table is that one can then take differences between the values for different redshifts.

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

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

$$\begin{aligned} d^c &= \int d\chi = \int_0^\infty \frac{dt}{a(t)} = \left(\frac{\Omega_\Lambda}{\Omega_m}\right)^{1/3} \frac{2}{3\sqrt{\Omega_\Lambda} H_0} \int_0^\infty \sinh^{-2/3} x\, dx \\ &= \Omega_\Lambda^{-1/6}\Omega_m^{-1/3} \frac{\Gamma(\frac{1}{6})\Gamma(\frac{1}{3})}{3\sqrt{\pi}} H_0^{-1} = 2.8044\,\Omega_\Lambda^{-1/6}\Omega_m^{-1/3} H_0^{-1}\,. \end{aligned} \tag{156}$$

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

$$B(p,q) \equiv \int_0^1 t^{p-1}(1-t)^{q-1}\, dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \tag{157}$$

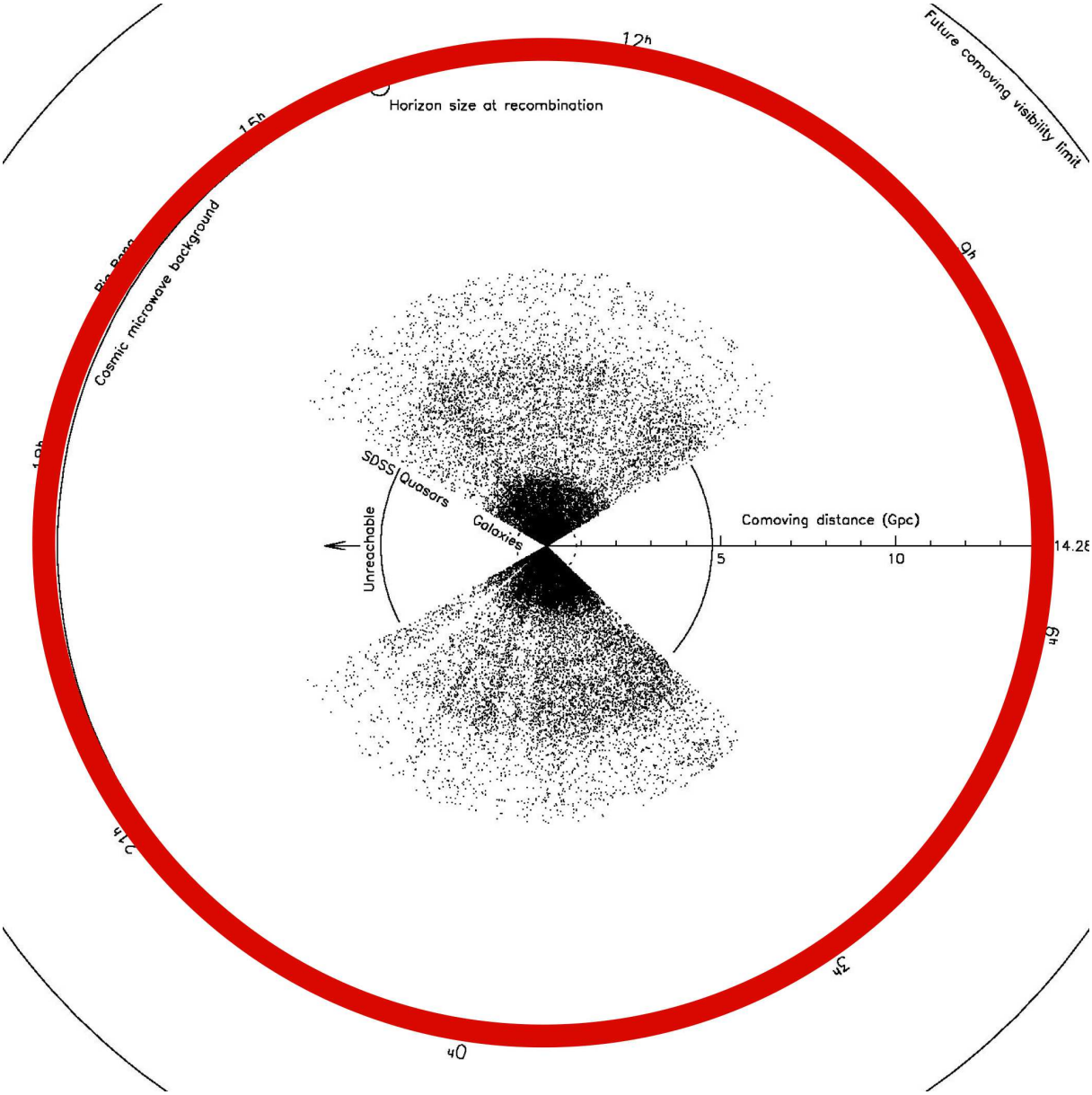

Figure 18: We had this figure already in Chapter 1, but let us look at it again. The figure is in comoving coordinates, so the galaxies do not move in time, except for their peculiar velocity. As time goes on the horizon recedes and we can see further out. The "Future comoving visibility limit" is how 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].

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

$$\int_0^\infty \sinh^\mu x\, dx \;=\; \tfrac{1}{2}\int_0^1 t^{\mu/2-1/2}(1-t)^{-\mu/2-1} \;=\; \tfrac{1}{2}B\left(\tfrac{\mu}{2}+\tfrac{1}{2},-\tfrac{\mu}{2}\right) \;=\; \frac{\Gamma(\frac{\mu}{2}+\frac{1}{2})\Gamma(-\frac{\mu}{2})}{2\Gamma(\frac{1}{2})} \qquad (158)$$

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 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
∞	0	0.963797	3.2446	13895.7	45.322	

Table 2: The cosmological parameters assumed here are $H_0 = 70.0\,\mathrm{km/s/Mpc}$, $\Omega_\Lambda = 0.7$, $\Omega_r = 4.18 \times 10^{-5} h^{-2} = 8.53 \times 10^{-5}$, $\Omega_m = 1 - \Omega_\Lambda - \Omega_r$.

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
∞	0	0.949918	3.1767	14146.7	46.140	

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 $p \equiv |\vec{p}|$ (momentum, 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 sufficiently slow compared to particle interaction rates, so 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). Although the expansion was faster in the early universe than later, the interaction rates were much higher because of higher density and higher particle energies, so that we can have equilibrium in the early universe, but not later.

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

$$\langle E \rangle = \frac{\rho}{n} = \begin{cases} \frac{7\pi^4}{180\zeta(3)} T \approx 3.15137 T & \text{fermions} \\ \frac{\pi^4}{30\zeta(3)} T \approx 2.70118 T & \text{bosons}\,. \end{cases} \tag{15}$$

In the above, ζ is the Riemann zeta function, and $\zeta(3) \equiv \sum_{n=1}^\infty (1/n^3) = 1.202057$.

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.

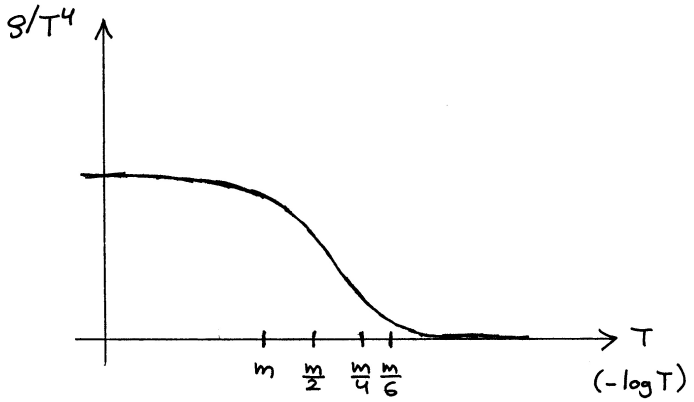

Figure 1: The fall of energy density of a particle species, with mass m, as a function of temperature (decreasing to the right).

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 formulae 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 must be 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	$0 \quad (< 1 \times 10^{-18}\,\mathrm{eV})$			$g = 2$	
						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]

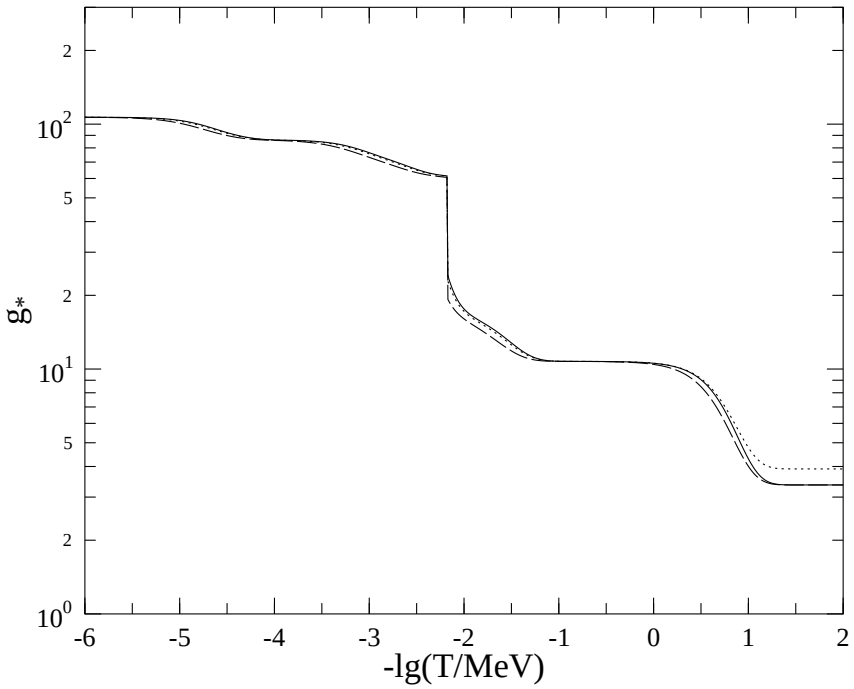

Figure 2: The functions $g_*(T)$ (solid), $g_{*p}(T)$ (dashed), and $g_{*s}(T)$ (dotted) calculated for the standard model particle content.

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$\to \pi^{\pm,0}$, $\quad 47.5 \to 3$)
$T < 100$ 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 transition, 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

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

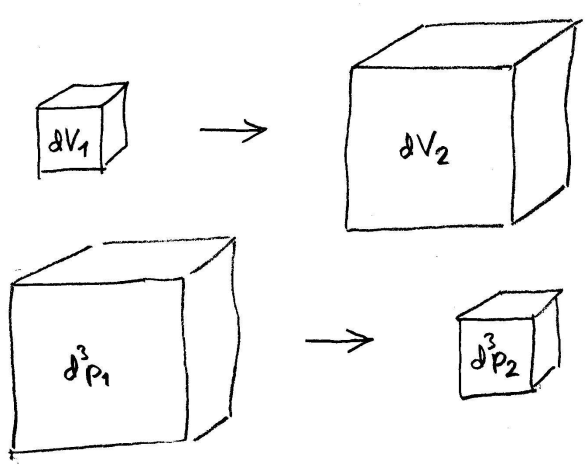

Figure 3: The expansion of the universe increases the volume element dV and decreases the momentum space element d^3p so that the phase space element $d^3p dV$ stays constant.

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

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

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

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

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,

$$\begin{aligned} T_\nu &= \left(\frac{4}{11}\right)^{\frac{1}{3}} T \;=\; 0.71377\, T \\ g_{*s}(T) &= 2 + 5.25 \cdot \frac{4}{11} \;=\; 3.909 \\ g_*(T) &= 2 + 5.25 \left(\frac{4}{11}\right)^{\frac{4}{3}} \;=\; 3.363. \end{aligned} \tag{40}$$

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 1.5% higher (at a given T) than the above calculation gives. The neutrino distribution also deviates slightly from kinetic equilibrium. In the above

$$5.25 = \frac{21}{4} = \frac{7}{4} N_\nu \,, \tag{41}$$

where $N_\nu = 3$ is the number of neutrino species. To correct for the additional energy density we define an *effective number of neutrino species* N_{eff} by

$$\rho_\nu \equiv N_{\text{eff}} \frac{7}{8} \left(\frac{4}{11}\right)^{4/3} \rho_\gamma \tag{42}$$

after e^+e^--annihilation. After many years of hard work by theorists it has been calculated that [3, 4, 5]

$$N_{\text{eff}} \approx 3.046\,. \tag{43}$$

This replaces 5.25 by $\frac{7}{4} N_{\text{eff}} = 5.3305$ in the above, so that

$$\begin{aligned} g_{*s}(T) &= 2 + 5.3305 \cdot \frac{4}{11} \;=\; 3.938 \\ g_*(T) &= 2 + 5.3305 \left(\frac{4}{11}\right)^{\frac{4}{3}} \;=\; 3.384\,. \end{aligned}$$

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

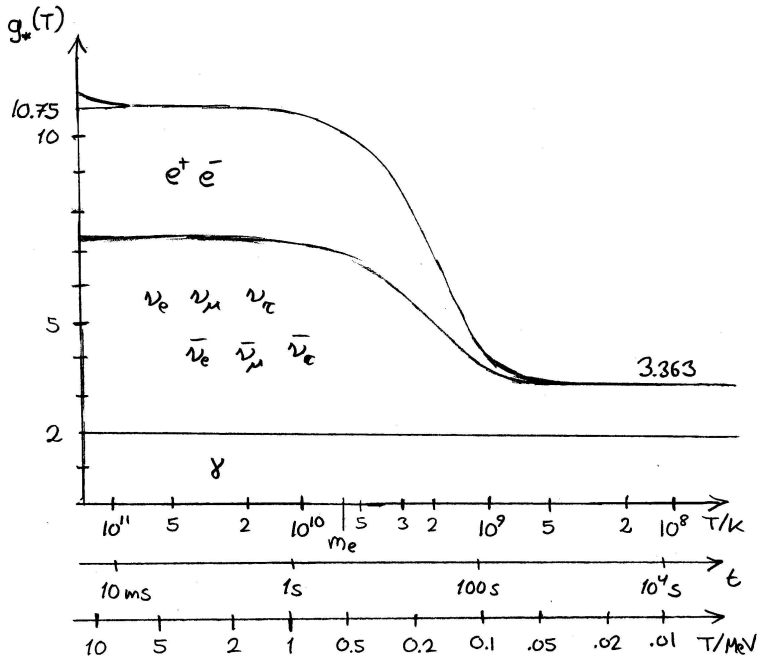

Figure 4: The evolution of the energy density, or rather, $g_*(T)$, and its different components through electron-positron annihilation. Since $g_*(T)$ is defined as $\rho/(\pi^2 T^4/30)$, where T is the photon temperature, the photon contribution appears constant. If we had plotted $\rho/(\pi^2 T_\nu^4/30)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

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 $T = T_0 = 2.7255$ K $= 0.2349$ meV, giving the neutrino background the temperature $T_{\nu 0} = 0.71377 \cdot 2.7255$ 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 $g_{*s}(T) = 3.938$ 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

$$ds^2 = -dt^2 + a^2(t)\left[dr^2 + r^2 d\vartheta^2 + r^2 \sin^2\vartheta \, d\varphi^2\right]. \tag{44}$$

and the Friedmann equation is

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho(T) = \frac{8\pi G}{3}\frac{\pi^2}{30} g_*(T) T^4. \tag{45}$$

To integrate this equation exactly we would need to calculate numerically the function $g_*(T)$ with all the annihilations[6]. For most of the time, however, $g_*(T)$ is changing slowly, so we can

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

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

$$\boxed{t \;=\; \frac{1}{2}H^{-1} \;=\; \sqrt{\frac{45}{16\pi^3 G}}\,\frac{T^{-2}}{\sqrt{g_*}} \;=\; 0.301 g_*^{-1/2}\frac{m_{\rm Pl}}{T^2} \;=\; \frac{2.4}{\sqrt{g_*}}\left(\frac{\rm T}{\rm MeV}\right)^{-2}\,{\rm s}} \tag{46}$$

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

$$m_{\rm Pl} \equiv \frac{1}{\sqrt{G}} = 1.2209 \times 10^{19}\,{\rm GeV}$$

is the Planck mass.[7] Thus

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

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. (46), 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

$$d_{\rm hor} = 2t = H^{-1}. \tag{47}$$

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

[7] Another common definition for Planck mass, which we shall call the *reduced Planck mass*, is

$$M_{\rm Pl} \equiv \frac{1}{\sqrt{8\pi G}} = 2.4353 \times 10^{18}\ \ {\rm GeV}$$

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,

$$n_{\rm B} \;=\; n_{\rm N} - n_{\bar{\rm N}} \;=\; n_{\rm p} + n_{\rm n} - n_{\bar{\rm p}} - n_{\bar{\rm n}}\,. \tag{48}$$

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

$$\eta \equiv \frac{n_{\rm B}(t_0)}{n_\gamma(t_0)}\,. \tag{49}$$

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

$$n_{\rm B} \propto a^{-3}\,. \tag{50}$$

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

$$n_{\rm B}(T) = \eta n_\gamma = \eta\frac{2\zeta(3)}{\pi^2}T^3 \qquad \text{for } T \ll m_{\rm e}\,, \tag{51}$$

and for all times (as long as the universe expands adiabatically and the baryon number is conserved), using Eqs. (36), (50), and (51),

$$n_{\rm B}(T) = \eta\frac{2\zeta(3)}{\pi^2}\frac{g_{*s}(T)}{g_{*s}(T_0)}T^3\,. \tag{52}$$

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").

[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.

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 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:

$$\begin{aligned} \Delta m_{21}^2 &\equiv m_2^2 - m_1^2 \approx 7.5\times10^{-5}\,\mathrm{eV}^2 \qquad \text{(solar)} \\ |\Delta m_{31}^2| &\equiv |m_3^2 - m_1^2| \approx |\Delta m_{32}^2| \approx 2.5\times10^{-3}\,\mathrm{eV}^2 \qquad \text{(atmospheric)}\,. \end{aligned} \tag{53}$$

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

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.

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 $T = T_0 = 2.7255\,\mathrm{K}$. (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_H. Since the universe is electrically neutral, $n_\mathrm{p} = n_\mathrm{e}$. The conservation of baryon number gives $n_\mathrm{B} = n_\mathrm{p} + n_\mathrm{H}$. From Sec. 4.1 we have

$$n_i = g_i \left(\frac{m_i T}{2\pi}\right)^{3/2} e^{\frac{\mu_i - m_i}{T}}. \tag{54}$$

For as long as the reaction

$$p + e^- \;\leftrightarrow\; \mathrm{H} + \gamma \tag{55}$$

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

$$n_\mathrm{H} = \frac{g_\mathrm{H}}{g_\mathrm{p} g_\mathrm{e}} n_\mathrm{p} n_\mathrm{e} \left(\frac{m_\mathrm{e} T}{2\pi}\right)^{-3/2} e^{B/T}, \tag{56}$$

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

$$x \equiv \frac{n_\mathrm{p}}{n_\mathrm{B}} \quad \Rightarrow \quad \frac{n_\mathrm{H}}{n_\mathrm{p} n_\mathrm{e}} = \frac{(1-x)}{x^2 n_\mathrm{B}}. \tag{57}$$

Using (51), Eq. (56) becomes

$$\frac{1-x}{x^2} = \frac{4\sqrt{2}\,\zeta(3)}{\sqrt{\pi}} \eta \left(\frac{T}{m_\mathrm{e}}\right)^{3/2} e^{B/T}, \tag{58}$$

the *Saha equation* for ionization in thermal equilibrium. When $B \ll T \ll m_\mathrm{e}$, 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 $(T/m_\mathrm{e})^{3/2}$ 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. (58) closely, but as this equilibrium fraction begins to fall rapidly, the true ionization fraction begins to lag behind. As

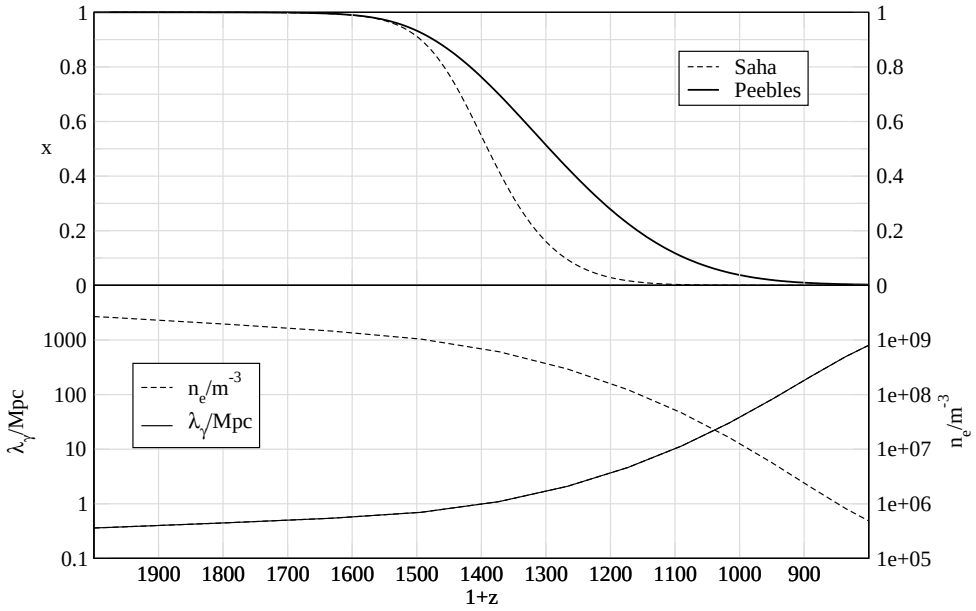

Figure 5: Recombination. In the top panel the dashed curve gives the equilibrium ionization fraction as given by the Saha equation. The solid curve is the true ionization fraction, calculated using the actual reaction rates (original calculation by Peebles). You can see that the equilibrium fraction is followed at first, but then the true fraction lags behind. The bottom panel shows the free electron number density n_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.)

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 (55). 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 $T \propto a^{-1}$.)

We get (for $\eta \sim 10^{-9}$)

$$T_{\rm rec} \sim 0.3 \text{ eV}$$
$$z_{\rm rec} \sim 1300.$$

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

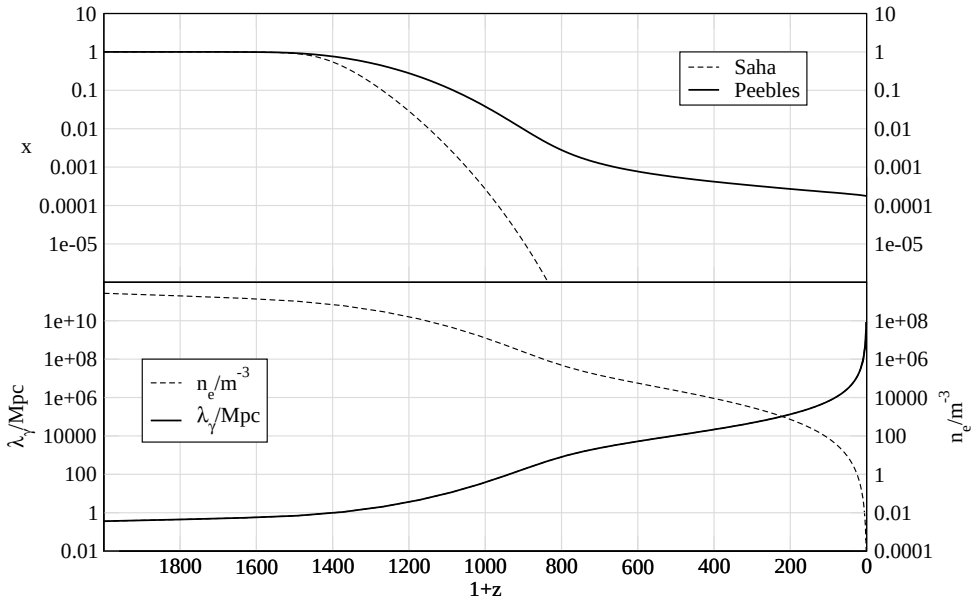

Figure 6: Same as Fig. 5, but with a logarithmic scale for the ionization fraction, and the time (actually redshift) scale extended to present time ($z = 0$ or $1 + z = 1$). You can see how a residual ionization $x \sim 10^{-4}$ remains. This figure does not include the reionization which happened at around $z \sim 10$. (Figure by R. Keskitalo.)

$T \ll B$, 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

$$T_{\rm dec} \approx 2974\,{\rm K} \sim 0.256\,{\rm eV}$$
$$z_{\rm dec} \approx 1090.$$

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

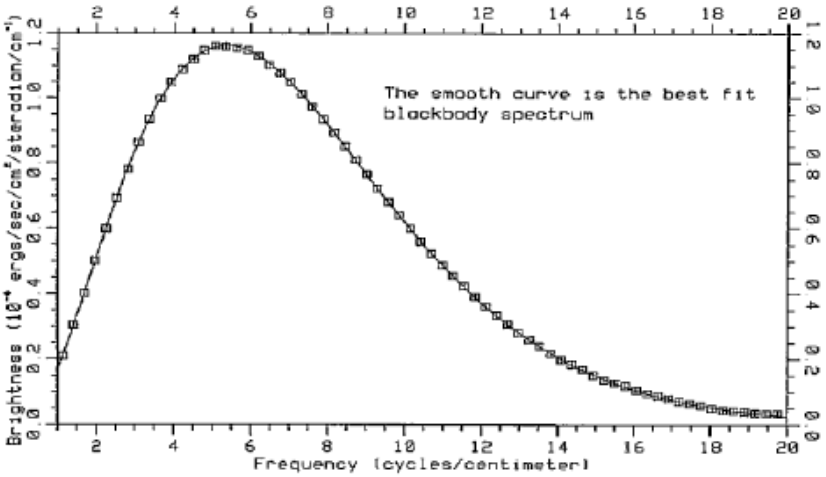

Figure 7: The CMB frequency spectrum as measured by the FIRAS instrument on the COBE satellite [6]. This first spectrum from FIRAS is based on just 9 minutes of measurements. The CMB temperature estimated from it was $T = 2.735 \pm 0.060\,\mathrm{K}$. The current estimate $T = 2.7255 \pm 0.0006\,\mathrm{K}$ (68% confidence) [7].

Electroweak Transition	$T \sim 100$ GeV	$t \sim 20$ ps
QCD Transition	$T \approx 150$ MeV	$t \approx 20\mu$s
Neutrino Decoupling	$T \approx 1$ MeV	$t \approx 1$ 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	$t \sim 370000$ yr

Table 4: Early universe events.

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.

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 photons 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 ($T_0 = 2.7255$ K) is

$$n_{\gamma 0} = \frac{2\zeta(3)}{\pi^2} T_0^3 = 410.727 \text{ photons/cm}^3 \tag{59}$$

and the energy density is

$$\rho_{\gamma 0} = 2\frac{\pi^2}{30} T_0^4 = 2.70118 T_0 n_{\gamma 0} = 4.64509 \times 10^{-31} \text{kg/m}^3 \,. \tag{60}$$

Since the critical density is

$$\rho_{\text{cr}0} = \frac{3H_0^2}{8\pi G} = h^2 \cdot 1.8783 \times 10^{-26} \text{kg/m}^3 \tag{61}$$

we get for the photon density parameter

$$\Omega_\gamma \equiv \frac{\rho_{\gamma 0}}{\rho_{\text{cr}0}} = 2.4730 \times 10^{-5} h^{-2} \,. \tag{62}$$

[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

$$\rho_\nu = \frac{7N_{\rm eff}}{4}\frac{\pi^2}{30}T_\nu^4\,. \tag{63}$$

After e^+e^--annihilation this gives

$$\rho_\nu = \frac{7N_{\rm eff}}{8}\left(\frac{4}{11}\right)^{\frac{4}{3}}\rho_\gamma\,, \tag{64}$$

where $N_{\rm eff} = 3.046$ is the effective 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 $N_{\rm eff} = 2$–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 neutrinos are still relativistic today, the neutrino density parameter is

$$\Omega_\nu = \frac{7N_{\rm eff}}{8}\left(\frac{4}{11}\right)^{\frac{4}{3}}\Omega_\gamma = 0.6918\Omega_\gamma = 1.7107\times10^{-5}h^{-2}\,, \tag{65}$$

so that the total radiation density parameter is

$$\Omega_r = \Omega_\gamma + \Omega_\nu = 4.1837\times10^{-5}h^{-2} \sim 10^{-4}\,. \tag{66}$$

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

$$\omega_\gamma = 2.4730\times10^{-5} \tag{67}$$

$$\omega_\nu = 1.7107\times10^{-5} \tag{68}$$

$$\omega_r = \omega_\gamma + \omega_\nu = 4.1837\times10^{-5}\,. \tag{69}$$

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

$$\rho_\nu = \Omega_\nu\rho_{\rm cr0}a^{-4} \tag{70}$$

using the Ω_ν of Eq. (65), 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 $\rho_{r0} = \rho_{\gamma0} + \rho_{\nu0}$ (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?

---

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

$$B_i \equiv (A-Z)m_{\mathrm{n}} + Z m_{\mathrm{p}} - m_i \tag{5}$$

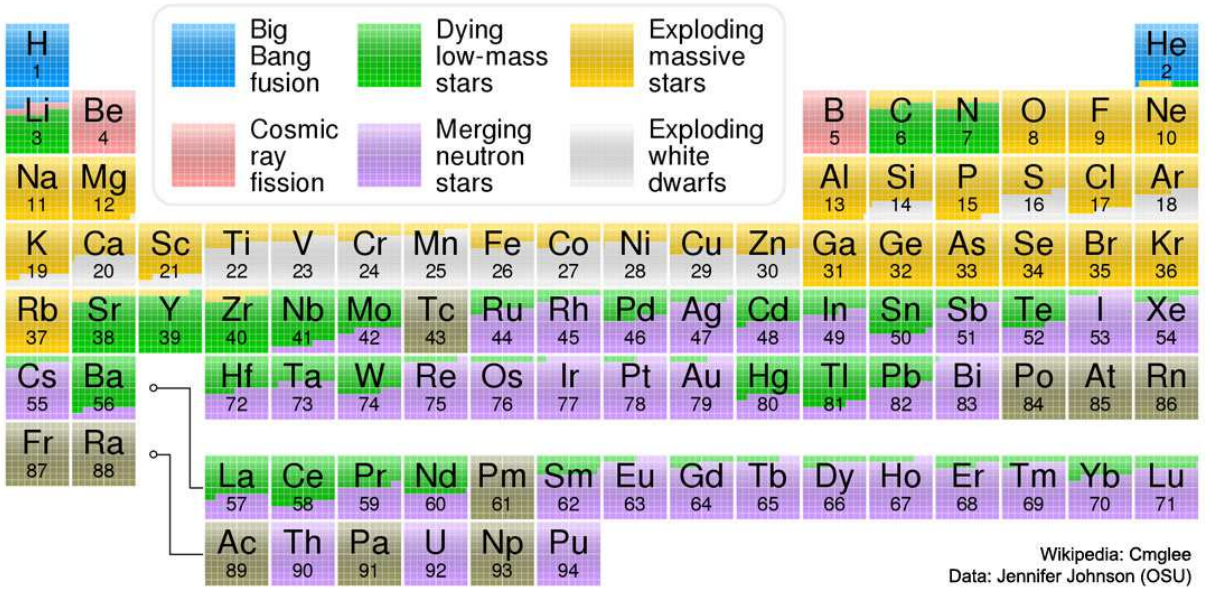

Figure 1: Astronomy Picture of the Day, 2017 October 24 (`https://apod.nasa.gov/apod/ap171024.html`: "Explanation: The hydrogen in your body, present in every molecule of water, came from the Big Bang. There are no other appreciable sources of hydrogen in the universe. The carbon in your body was made 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 ($n_{\rm e^-} - n_{\rm e^+}$) 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

$$X_{\rm n}(t) = X_{\rm n}(t_1)e^{-(t-t_1)/\tau_{\rm n}}, \tag{18}$$

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 relative 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 $g_* = 3.384$. 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

$$t_{\rm ns} - t_1 = 266.0\ {\rm s} \qquad \text{(in reality 264.3 s)}.$$

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 $T = 1\,\mathrm{MeV}$.)

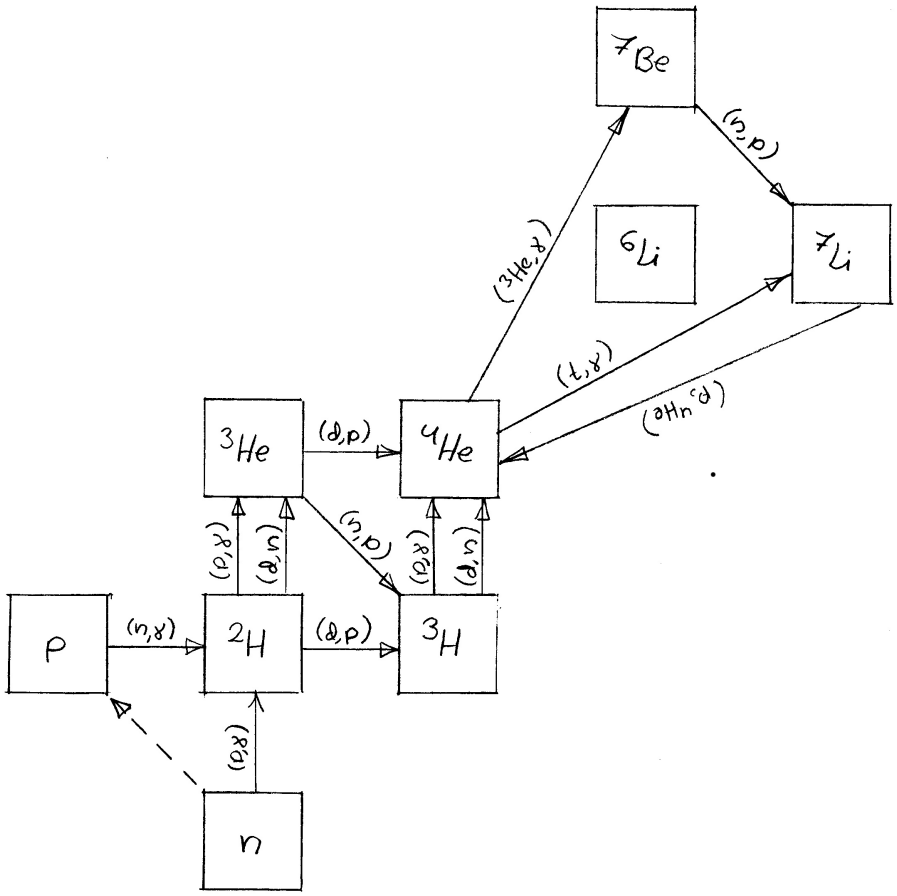

Figure 2: The 12 most important nuclear reactions in big bang nucleosynthesis.

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,

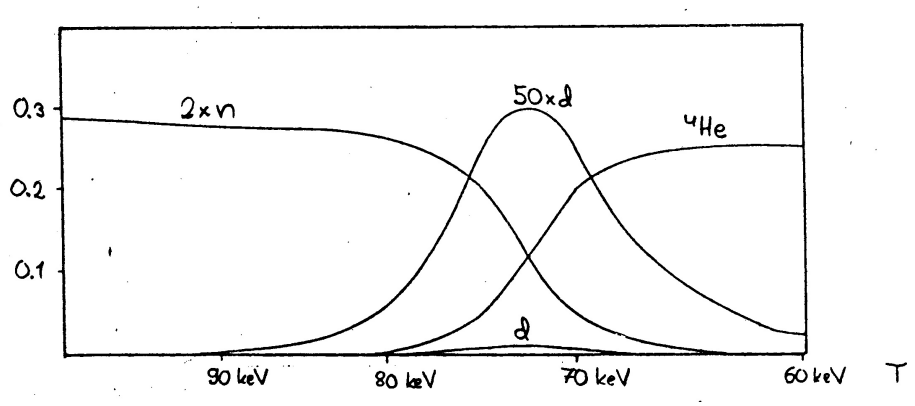

Figure 3: The time evolution of the n, ^{2}H (written as d) and ^{4}He abundances during BBN. Notice how the final ^{4}He abundance is determined by the n abundance before nuclear reactions begin. Only a small part of these neutrons decay or end up in other nuclei. Before becoming ^{4}He, all neutrons 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 $T = (90, 80, 70, 60)\,\mathrm{keV}$ is $(152, 199, 266, 367)\,\mathrm{s}$. Thus the action peaks at about $t = 4\,\mathrm{min}$. 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].

the baryon/photon ratio[5]

$$\begin{aligned} \eta &\equiv \frac{n_\mathrm{B}}{n_\gamma} = \frac{\Omega_b \rho_{\mathrm{cr}0}}{m_N n_{\gamma 0}} = \frac{\Omega_b}{m_N n_{\gamma 0}} \frac{3H_0^2}{8\pi G} \\ &= 273.8 \times 10^{-10} \omega_b = 1.457 \times 10^{18} \left(\frac{\rho_{b0}}{\mathrm{kgm}^{-3}} \right). \end{aligned} \tag{22}$$

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 $\omega_b \equiv \Omega_b h^2$.

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.

[5]To relate η accurately to ω_b we have to decide what value to use for m_N. The most precise determination for ω_b is from the effect of baryon-photon acoustic oscillations on the CMB (Chapter 9) before photon decoupling. At that time the baryonic matter was about 75% protons and 25% helium. Proton mass is $m_\mathrm{p} = 938.272\,\mathrm{MeV}$ and helium nucleus mass divided by four (i.e., mass per nucleon) is 931.995 MeV. Electrons are included in the concept of baryonic matter, so that we have to add $m_\mathrm{e} = 0.511\,\mathrm{MeV}$ for each proton, and half of this to each nucleon in a helium nucleus. Helium atoms form earlier than hydrogen atoms, but atomic binding energies are so small that we can ignore them. This gives an average baryonic mass $m_N \approx 937.068\,\mathrm{MeV}$ per nucleon. Fusion reactions in the later universe will decrease this value somewhat, but the effect is smaller than the effect of helium formation in BBN.

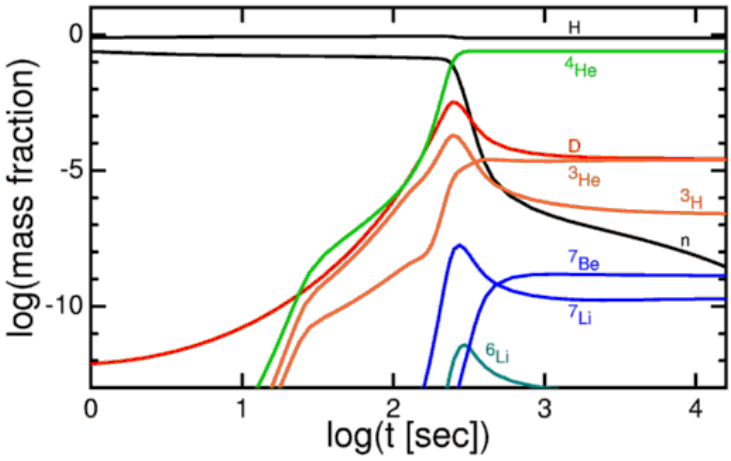

Figure 4: Time evolution of the abundances of the light isotopes during BBN. From `http://www.astro.ucla.edu/~wright/BBNS.html`

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.

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.[6]

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

[6] Many cosmological parameters can be estimated from the CMB anisotropy, as will be discussed in Cosmology II. The Planck estimate[5] is $\omega_b = 0.02242 \pm 0.00014$, or $\eta = (6.14 \pm 0.03) \times 10^{-10}$.

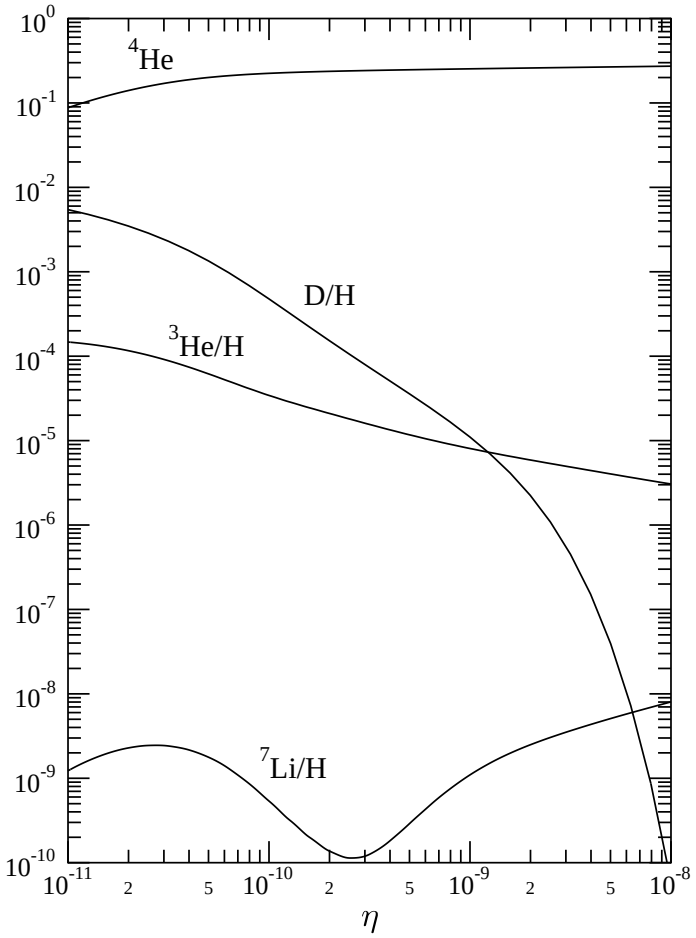

Figure 5: The primordial abundances of the light elements as a function of η. For ^{4}He we give the mass fraction, for D = ^{2}H, ^{3}He, and ^{7}Li the number ratio to H = ^{1}H, i.e., $n_i/n_{\rm H}$.

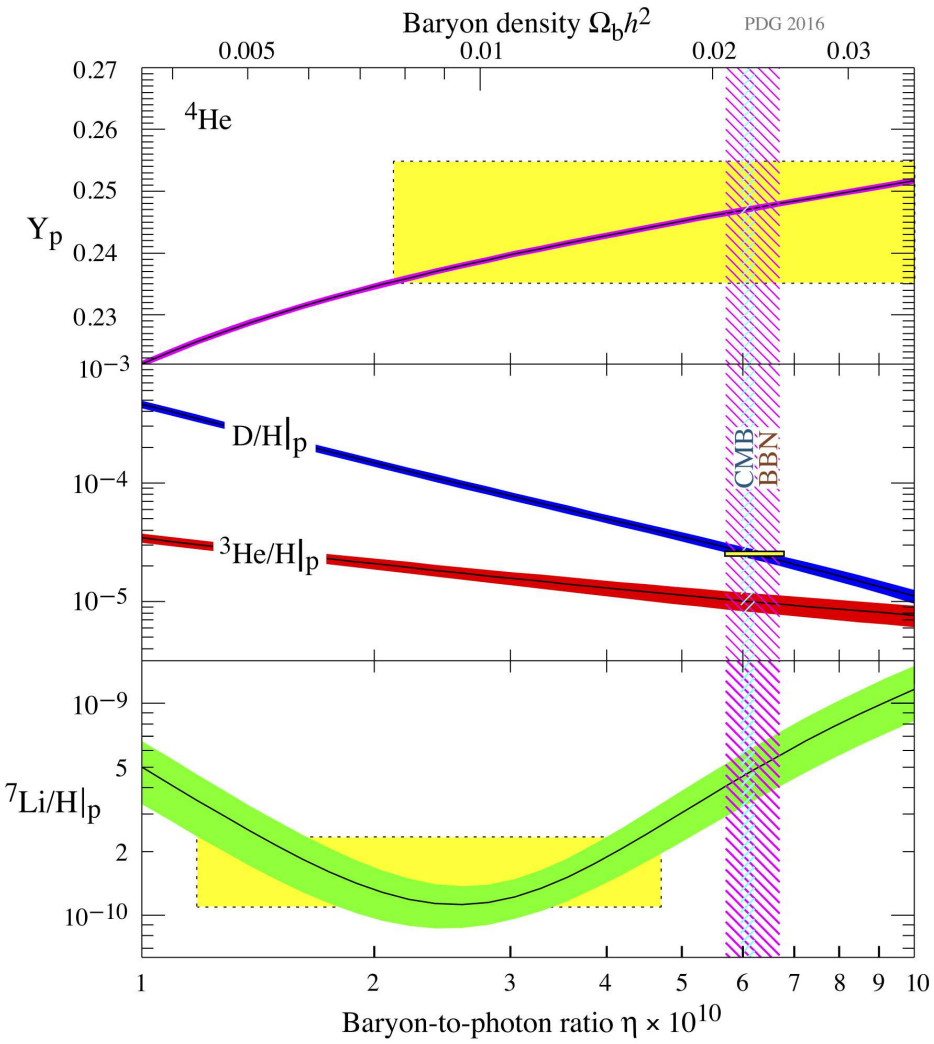

Figure 6: Determining the baryon/photon ration η by comparing BBN predictions to observations. The width of the bands around the the curves represents the uncertainty in BBN prediction due to uncertainty about the reaction rates. The vertical extent of the yellow boxes represent the estimate of the primordial 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].

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.[7] 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 $\eta = 1.5\text{–}7 \times 10^{-10}$. 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 the above range, to $\eta = 5.8\text{–}6.6 \times 10^{-10}$. 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 $\eta = 1.5\text{–}4.5 \times 10^{-10}$. 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 $\eta = 5.8\text{–}6.6 \times 10^{-10}$ (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 $\eta = 1.5\text{–}7 \times 10^{-10}$ corresponds to $\omega_b = \Omega_b h^2 = 3.65 \times 10^7 \eta = 0.0055\text{–}0.026$. With $h = 0.7 \pm 0.07$, this gives $\Omega_b = 0.009\text{–}0.07$ for a conservative range of the baryonic density parameter. With $\eta = 5.8\text{–}6.6 \times 10^{-10}$ 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.

[7] 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.

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 momentum 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),

$$\begin{aligned} n_{\rm x0} \;&\approx\; \frac{g_{*S}(T_0)}{g_{*S}(T_d)} \left(\frac{T_0}{T_d}\right)^3 g_{\rm x} \left(\frac{mT_d}{2\pi}\right)^{3/2} e^{-y} \\ &=\; \frac{g_{\rm x}}{(2\pi)^{3/2}} \frac{g_{*S}(T_0)}{g_{*S}(T_d)} y^{3/2} e^{-y} T_0^3 \\ &=\; \frac{g_{\rm x}}{(2\pi)^{3/2}} \frac{g_{*S}(T_0)}{g_{*S}(T_d)} A^{-1} y^{1+q} T_0^3 \\ &=\; \sqrt{\frac{g_*(T_d)}{90}} \frac{g_{*S}(T_0)}{g_{*S}(T_d)} \frac{\pi}{M_{\rm Pl} m \sigma_0} y^{1+q} T_0^3 , \end{aligned} \tag{38}$$

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

$$\frac{n_{\rm x0}}{n_{\gamma 0}} \;=\; \frac{\pi^3}{\zeta(3)} \sqrt{\frac{g_*(T_d)}{360}} \frac{g_{*S}(T_0)}{g_{*S}(T_d)} \frac{y^{1+q}}{M_{\rm Pl} m \sigma_0} . \tag{40}$$

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 $g_{*S}(T_0) \approx 3.938$, this becomes

$$\frac{n_{\mathrm{x}0}}{n_{\gamma 0}} \;=\; \frac{\pi^3}{\zeta(3)} \frac{1}{\sqrt{360}} \frac{g_{*S}(T_0)}{\sqrt{g_*(T_d)}} \frac{y^{1+q}}{M_{\mathrm{Pl}} m \sigma_0} \;\approx\; 5.35 \frac{y^{1+q}}{\sqrt{g_*(T_d)} M_{\mathrm{Pl}} m \sigma_0} \,. \tag{41}$$

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 $G_F = 1.17 \times 10^{-5}$ 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 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 $y \approx \ln A$)

$$\begin{aligned} \frac{n_{\mathrm{x}0}}{n_{\gamma 0}} &\approx 5.35 \frac{y}{\sqrt{g_*(T_d)} M_{\mathrm{Pl}} m \sigma_0} = 2.1 \times 10^{-9} y \left(\frac{m}{\mathrm{GeV}}\right)^{-3} \\ &\approx 3.5 \times 10^{-8} \left(1 + 0.18 \ln \frac{m}{\mathrm{GeV}}\right) \left(\frac{m}{\mathrm{GeV}}\right)^{-3} . \end{aligned} \tag{48}$$

Taking the baryon-to-photon ratio to be $\eta = 6 \times 10^{-10}$, we get for the ratio of cold thermal relics to baryons

$$\frac{n_{\mathrm{x}0}}{n_{B0}} \approx 58 \left(1 + 0.18 \ln \frac{m}{\mathrm{GeV}}\right) \left(\frac{m}{\mathrm{GeV}}\right)^{-3} , \tag{49}$$

and since $m_N \approx 1$ GeV the mass density ratio

$$\frac{\rho_{\mathrm{x}0}}{\rho_{b0}} \approx 58 \left(1 + 0.18 \ln \frac{m}{\mathrm{GeV}}\right) \left(\frac{m}{\mathrm{GeV}}\right)^{-2} . \tag{50}$$

Since $\rho_{b0} \approx 0.05 \rho_{\mathrm{cr}0}$, the requirement that such relic particles (with energy density $\rho_{\mathrm{x}0} + \rho_{\bar{\mathrm{x}}0} = 2\rho_{\mathrm{x}0}$) 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 ρ_{cr} the critical density.

[7] The electroweak scale is more like 100 GeV, but this is considered close enough.

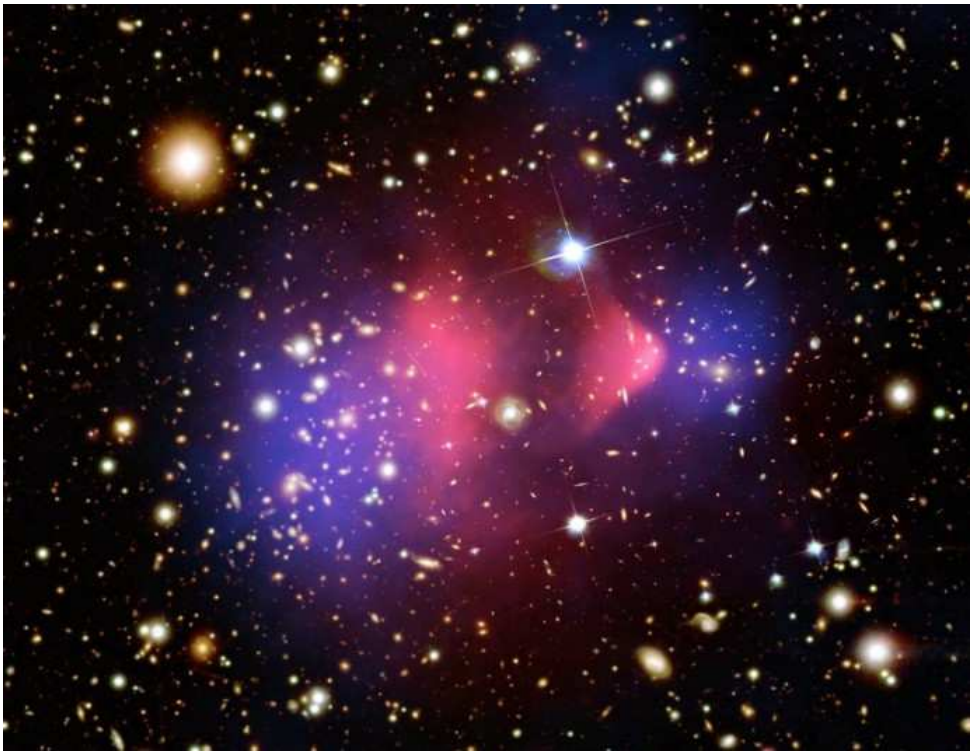

Figure 1: This is a composite image of galaxy cluster 1E 0657-56, also called the Bullet Cluster. It consists of two subclusters, a larger one on the left, and a smaller one on the right. They have recently collided and traveled through each other. One component of the image is an optical image which 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.

C Numerical Constants

C.1 Defining constants

The following constants of nature are used to define SI units and therefore they have exact numerical values in them [1]. The second equalities define the natural units we use in this course.

$$\begin{aligned} c &= 299\,792\,458\,\mathrm{m/s} = 1 && \text{defines meter} \\ h &= 6.626\,070\,15 \times 10^{-34}\,\mathrm{Js} = 2\pi && \text{defines kilogram} \\ k_B &= 1.380\,649 \times 10^{-23}\,\mathrm{J/K} = 1 && \text{defines kelvin} \\ e &= 1.602\,176\,634 \times 10^{-19}\,\mathrm{C} && \text{defines ampere} \end{aligned} \tag{1}$$

C.2 Other constants of nature

$$\begin{aligned} G &= 6.674\,30 \pm 15 \times 10^{-11}\,\mathrm{m^3/kgs^2} \\ m_{\mathrm{Pl}} &= 1.220\,890 \pm 14 \times 10^{19}\,\mathrm{GeV} \end{aligned} \tag{2}$$

C.3 Mathematical constants

Approximate values:

$$\begin{aligned} \pi &\approx 3.14159\,26535\,8979 \\ e &\approx 2.71828\,18284 \\ \zeta(3) &\approx 1.20205\,69032 \end{aligned} \tag{3}$$

C.4 Astronomical units

$$\begin{aligned} \text{Julian year} &\equiv 365.25\,\text{days} = 31\,557\,600\,\mathrm{s} \\ \text{light year} &= 9\,460\,730\,472\,580\,800\,\mathrm{m} \quad \text{(exact)} \\ \mathrm{AU} &= 149\,597\,870\,700\,\mathrm{m} \quad \text{(2012 definition)} \\ \mathrm{pc} &\equiv 3600\,(180/\pi)\,\mathrm{AU} \approx 206\,264.806\,247\,\mathrm{AU} \\ &\approx 3.085\,677\,581\,49 \times 10^{16}\,\mathrm{m} \approx 3.261\,564\,\text{light years} \end{aligned} \tag{4}$$

C.5 Cosmological quantities

Present CMB temperature [1]

$$\begin{aligned} T_0 &= 2.7255 \pm 0.0006\,K \\ \Rightarrow\ n_{\gamma 0} &= \frac{2\zeta(3)}{\pi^2} T_0^3 \approx 410.73\ \text{photons/cm}^3 \\ \rho_{\gamma 0} &= 2\,\frac{\pi^2}{30} T_0^4 \approx 4.6451 \times 10^{-31}\,\mathrm{kg/m^3} \end{aligned} \tag{5}$$

[1] physics.nist.gov/cuu/Units/current.html