Cosmology

Useful numbers for reference

$\hbar = 1.05457148 \times 10^{-34} \mathrm{m}^2\,\mathrm{kg\,s}^{-1}$ $\quad$ $c \equiv 299792458\ \mathrm{m\,s}^{-1}$

$k_B = 1.3806504 \times 10^{-23} \mathrm{m}^2\,\mathrm{kg\,s}^{-2}\,\mathrm{K}^{-1}$ $\quad$ $G = 6.67428 \times 10^{-11} \mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2}$

$\sigma_T = 6.6524616 \times 10^{-29} \mathrm{m}^2$ $\quad$ $\sigma_{SB} = 5.6704 \times 10^{-8} \mathrm{J\,s}^{-1}\,\mathrm{m}^{-2}\,\mathrm{K}^{-4}$

$m_p = 1.672621637 \times 10^{-27} \mathrm{kg} \approx 0.938 \mathrm{GeV}$ $\quad$ $m_e = 9.10938215 \times 10^{-31} \mathrm{kg} \approx 0.511 \mathrm{MeV}$

$m_n = 1.6749272 \times 10^{-27} \mathrm{kg}$ $\quad$ $m_n - m_p \approx 1.293 \mathrm{MeV}$

$1\mathrm{eV} = 1.60217646 \times 10^{-19} J = 1.78266173 \times 10^{-36} \mathrm{kg}$ $\quad$ $1\mathrm{Mpc} = 3.08568025 \times 10^{22} \mathrm{m}$

$1\mathrm{GYr} = 3.1556926 \times 10^{16} \mathrm{s}$ $\quad$ $H_0^{-1} \equiv [100h \mathrm{km\,s}^{-1}\,\mathrm{Mpc}^{-1}]^{-1} \approx h^{-1} 9.78 \mathrm{Gyr}.$

$\zeta(3) \approx 1.202056903$ $\quad$ $\zeta(4) = \pi^4/90 \approx 1.082323234$

$$\text{Planck mass } m_P \equiv \sqrt{\hbar c/G} = 1.2209 \times 10^{19} \mathrm{GeV} = 2.17644 \times 10^{-8} \mathrm{kg}$$

$$\text{Reduced Planck mass } M_P \equiv \sqrt{\hbar c/(8\pi G)} \approx 2.43 \times 10^{18} \mathrm{GeV} \approx 4.34 \times 10^{-9} \mathrm{kg}$$

We will always (except where explicitly stated) use "natural" units where $k_B = c = \hbar = 1$. As an example, a temperature of 1 K corresponds to about 8.6×10^{-5} eV and the Hubble constant $H_0 = 100h$km/s/Mpc is also $h(9.78\mathrm{Gyr})^{-1}$.

I. OBSERVATIONAL OVERVIEW

A. The dark sky

Perhaps the most obvious cosmological observation that we can make is to look at the sky at night. Why is most of the sky dark rather than light? This is called *Olbers' paradox*. This single observation tells us something important about the universe: it cannot be infinite and static, with any constant density of stars, otherwise in every direction we look we would see a star (so light, not dark). In the Big Bang model the dark sky is explained both by the expansion of space and the finite age of the universe: most lines of sights do not intersect a star, and the radiation that we would see from the early universe has been redshifted (wavelength becomes much longer, see later) so that it is no longer visible to the naked eye.

B. Light - photons

1. *Visible Light*

The sun is about 8 light minutes away from Earth. In astronomy a more common distance unit is the parsec = 3.261 light years, corresponding to the typical distance of other stars.

Out of our solar system we can see

- **stars** Objects like our sun (mass $M_\odot = 2 \times 10^{30}$kg). Nearest stars are about 4 light years away (Alpha Centauri).

- **galaxies** Can contain billions of stars, the Milky Way has $\sim$ 300 billion stars (total mass $\sim 10^{12} M_\odot$). The nearest large galaxy is Andromeda, at 770pc, but smaller galaxies are part of our local group (the Large Magellanic Cloud at 50kpc). A megaparsec (Mpc $= 3.086 \times 10^{22} m$) is a good unit in cosmology, corresponding to about the typical separation of galaxies. On the largest scales the distribution of galaxies is thought to trace the total matter of the universe (more matter $\implies$ more galaxies).
- **clusters** (of galaxies). e.g. the Coma cluster is about 100 Mpc away, with about 10,000 galaxies. These are the largest gravitationally collapsed objects. Most nearby galaxies are parts of groups or clusters, but on larger scales most of the less bright galaxies are not (*field galaxies*). Clusters are sometimes grouped into superclusters, often joined by filaments or walls of galaxies (with large $\sim$ 50Mpc voids in between).

2. Microwaves

Penzias and Wilson detected the radiation in 1964 and received the Nobel prize in 1978. The first precise observation had to wait for the COBE satellite in 1990, which proved that the spectrum is very close to a perfect blackbody form, with a temperature of $T_0 = 2.725 \pm 0.001$ K.

COBE also showed that the CMB is near-perfectly isotropic, with the dominant departure being a dipole pattern due to the doppler shift from the relative motion of the earth and the CMB rest frame (i.e. the dipole is not cosmologically interesting and depends on how the earth happens to be moving). The small 10^{-5} anisotropies are measured by WMAP and Planck, and provide a view of fluctuations in the very early universe (see later). The fact that the anisotropies are so small indicates that the very early universe was very smooth (with the structures we see today forming much later by gravitational collapse).

3. Radio waves

Very distant galaxies can be seen in the radio. We can also pick up the hyperfine (21cm) transition in hydrogen, which allows us to observe neutral gas in the universe (not necessarily in galaxies) over a huge size of the observable universe.

4. X-rays

These are emitted by very hot gas and is a good way to see clusters, where gas can have temperatures of tens of millions of degrees ($\sim 90\%$ of the gas in clusters is not in galaxies).

5. Infra red and gamma rays

Not so useful for cosmology, but are useful for looking close to the galactic plane.

C. Other radiation

Neutrinos are very weakly interacting, but high-energy ones from eg. supernovae can be detected on earth in large detectors. Neutrinos from the big bang are unfortunately too low energy to be detected.

Gravitational waves from e.g. black-hole in-spiral may also be detectable. (Cosmological gravitational waves from the big bang may also be detectable indirectly via their imprint in the polarization of the microwave background)

Cosmic rays are detectable when the collide with the atmosphere or directly with detectors, but are not of that much direct use for cosmology since they don't come from that far away.

There may of course also be other kinds of matter we don't know about, and dark matter detection experiments are looking for these.

II. COPERNICAN PRINCIPLE, ISOTROPY AND HOMOGENEITY

A. Copernican principle

Also known as the cosmological principle. This states that we are at a fairly typical place in the universe, in that observers in other galaxies would see roughly the same things as us on large scales. The assumption is that universe as a whole is similar to what we see locally, at least on scales $\gtrsim 100$Mpc.

Of course we aren't actually at a typical place in the universe - most of the universe is empty space not the surface of a comfy planet. However all observers have to be somewhere where they could have evolved, so we can expect to be typical of observers, if not necessarily typical of locations in space or time (this is just an observational selection effect, sometimes loosely called the *anthropic principle*).

B. Homogeneity and isotropy

The Copernican principle implies that we should expect statistical homogeneity: on average observers at any fixed time from the big bang should see the same thing at any different location in the universe. This is consistent with the large-scale distribution of galaxies which looks fairly uniform if you smooth the number over a scale of ~ 100Mpc. The CMB also tells us that the early universe was very smooth in all directions: the CMB is nearly isotropic. We don't know that the CMB looks nearly isotropic to other observers, but from the Copernican principle we would expect that to be case, in which case we know the early universe was nearly homogeneous and isotropic at early times.

The fundamental assumption that we shall make is

- The large-scale universe is accurately modelled as spatially homogeneous and isotropic

This assumption is self-consistent in that at early times the universe is very smooth indeed (from the CMB), with gravitational growth gradually forming the structures (galaxies etc) that we see today. What is much less obvious is that this assumption is valid in the late universe, e.g. the last billions of years till today, when the universe is actually very lumpy on small scales (galaxies, clusters, and voids); in fact it remains an open research question to what level of precision the assumption is valid. Here we shall simply assume that it is valid, and we shall see that this is sufficient to describe a wide range of cosmological phenomena.

In practice cosmologist usually use *perturbation theory*: the background model is that the universe is *exactly* homogeneous and isotropic (the *FRW* - Friedmann-Robertson-Walker model), and then this is made more realistic by separately modelling the effect of small perturbations. In this course we are only going to focus on the background large-scale properties of the universe, which is sufficient to understand many key results (the Early Universe course will then go into some detail about the origin and evolution of the perturbations).

III. COSMOLOGICAL OBSERVABLES

Cosmology is a science, where the source of data is observations. Unlike other sciences we cannot make controlled experiments, and we also cannot easily get off Earth to explore other places. The data is therefore limits to experiments on earth, and short-timescale Earth-based observations. Ideally we can observe in all directions, and measure things as a function of angular coordinate. So what can we actually measure? All that we can easily currently detect is light; so directly, only photons (light) as a function of angle on the sky, frequency, polarization and time. Ultimately anything else we wish to know about the universe at large has to be inferred from the observations of these kinds.

A. Redshift

Distances and cosmological times are not directly observable; we are stuck on earth, and the timescale of life is far too short compared to the age of the cosmos for anything to change signif-

icantly over a lifespan. Instead, we have to use what we can actually measure, which is the properties of objects that we can see; i.e. objects on our past light cone. When we measure light from distant objects we can also measure a spectrum; when we do this we see spectral lines. We also know from measurements on earth what frequency and pattern of spectral lines to expect from many different elements and ions. So we can compare the observed spectral lines with known spectra, and infer something about the composition of the distant objects. However the spectra we observe are not actually identical to those on earth: they are *redshifted*, meaning all the wavelengths (or equivalently frequencies) are scaled by some factor compared to what we would measure in the lab (and hence also by assumption would be measured by a lab at the source object). This ratio of the measured to the source frequencies is used to define *redshift* by

$$z \equiv \frac{\lambda_{\rm obs} - \lambda_{\rm source}}{\lambda_{\rm source}} = \frac{\nu_{\rm source}}{\nu_{\rm obs}} - 1, \tag{1}$$

where $\nu_{\rm source} = c/\lambda_{\rm source}$ is the source (lab) frequency. The redshift is a directly related to observables, and with spectroscopy measurable accurately because we know very accurately the frequency of the spectral lines (for example the Lyman-α transition in Hydrogen). However spectroscopy becomes difficult if the source is too dim, so measuring *spectroscopic* redshifts becomes expensive for very distant objects (i.e. require a large collecting area and long observation time to get enough photons). For this reason sometimes we have to make do with *photometric* redshifts, where the redshift is inferred approximately by looking at the colour of observed objects (intensity measured through some different coloured filters).

In cosmological observations angular coordinate and redshift are often the key observable quantities used to describe light cone location of specific objects. We cannot directly see anything not on our light cone, and we also cannot directly convert redshift into a distance or time, though as we shall see higher redshift usually means further away and older. In the case of the smooth CMB spectral lines cannot be observed, but instead the spectrum is measured to have a nearly thermal blackbody spectrum; in this case observations instead can conveniently be described by an angular coordinate and temperature.

IV. EXPANSION AND COSMOLOGICAL REDSHIFT

A. Doppler effect, expansion and Hubble's law

The Doppler effect describes how the frequency of light changes with the velocity of the source or observer: a source moving away from us will be observed to radiation at longer wavelengths than one that is static, i.e. the Doppler effect causes redshift. In general the redshift is a combination of the velocity of the source and the observer, but since the velocity of the sun is only changing very slowly (e.g. rotation of the sun about the galaxy), the observer velocity is basically fixed and known, so the relative line-of-sight velocities of different sources can be determined from observation. For non-relativistic recession velocity $v \ll c$, the redshift is simply given by

$$z \equiv \frac{\lambda_{\rm obs} - \lambda_{\rm source}}{\lambda_{\rm source}} = \frac{v}{c} = \frac{\hat{\mathbf{n}} \cdot \mathbf{v}}{c}, \tag{2}$$

where the last equality is to emphasize that we are only sensitive to the velocity along our direction of observation $\hat{\mathbf{n}}$. Of course this does not work for velocities that are not much less than the speed of light, so the relation is only valid for $z \ll 1$.

What is found is that almost all objects are redshifted, very few are blue shifted. Taking this to arise from a Doppler shift implies that objects are moving away from us. Furthermore the further away an object is (e.g. on average dimmer galaxies are further away), the more redshifted it appears. The observed average nearly-linear relationship between distance and recession velocity is called *Hubble's law*, and can be written

$$\mathbf{v} = H_0 \mathbf{r}, \tag{3}$$

where H_0 is the *Hubble constant*. So objects in the universe appear to be expanding away from us. The verification of this law and the value of H_0 was one of the key aims of early modern cosmology. Figure

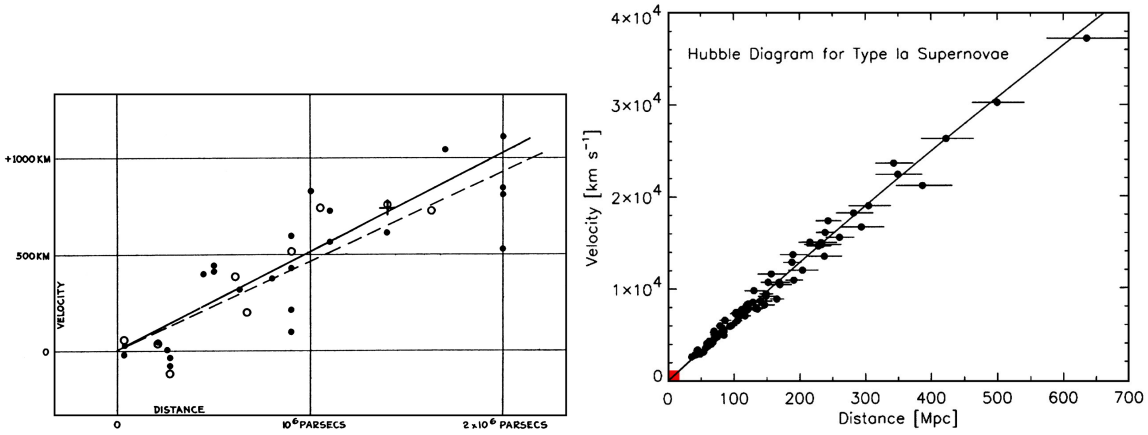

FIG. 1: Hubble diagrams, showing the relationship between recessional velocities of distant galaxies (inferred directly from the observed redshifts) and their distances (indirectly inferred). The left plot shows the original data of Hubble (1929), the right plot more recent data by Riess et al (1996). Notice the difference in scale.

1 shows the original data of Hubble replotted, together with a more recent compilation of data. Any given object will not have a redshift exactly satisfying Hubble's law because it has some dynamical local motion, called a *peculiar velocity*, but averaging over lots of objects (so the local motions average out), the law is found to be a good fit to nearby objects. For very distant objects z becomes larger, and the $v/c \sim z$ approximation breaks down; also high redshift objects are seen when the universe was significantly younger, and this evolution can change the fit from just a straight line. Hubble's law applies to nearby objects, with $z \ll 1$, where H_0 is the expansion rate today.

Note that in reality it is rather difficult to measure H_0. Firstly we need know a redshift; in principle this is easy because it is directly observable, but peculiar velocities give a scatter. At large distances the peculiar velocities become relatively less important, but it then becomes very difficult to infer the distance. The distance is not directly observable, so it has to be inferred, traditionally using a sequence of involved steps called the *distance ladder*. The first step uses parallax measurements, which allow for a fairly direct determination of distances to nearby objects using the change in angular position of the object as the earth rotates about the sun (using the accurately-measured radius of the orbit of the Earth). Angles can be measured to milliarcseconds, which limits the distance away for which parallax will work to a few hundred parsecs. Distances to objects further away are often obtained by using the dimming with distance for *standard candles*. For example it is possible to determine a relationship between the pulsation rate and absolute luminosity of the variable Cepheid-type stars, so that the observed luminosity can be used to infer a distance. Observations of these stars in other galaxies can then be extended with secondary distance determinations, especially using type-Ia supernova (SN-Ia).

The Hubble Key Project (using the Hubble Space Telescope, HST) came up with a distance ladder measurement of $H_0 = (72 \pm 8)\text{kms}^{-1}\text{Mpc}^{-1}$. More recent very indirect measurement from Planck, using some cosmological assumptions, gives $H_0 \sim (68 \pm 1)\text{kms}^{-1}\text{Mpc}^{-1}$; Currently H_0 is not known to much better than the percent level, and historically much less accurately, so people often parameterize it as $H_0 = 100h\,\text{kms}^{-1}\text{Mpc}^{-1}$, and express other results in terms of h (observations give $h \sim 0.7$).

B. Scale factor and comoving distances

How is this consistent with the Cosmological Principle, surely we aren't at the centre of the expansion? The Cosmological Principle implies that the expansion should be the same everywhere at any given time, or integrating up the expansion we can define a *scale factor* $a(t)$ that describes how relatively large the universe is at time t (there is *no* dependence on spatial coordinate $\mathbf{x}$ by homogeneity of the background model). Often $a(t)$ is normalized so that $a(t_0) = 1$ today, so that e.g. $a = 1/2$ means the universe was half as expanded as it is today.

Consider two observers only moving apart because of the expansion (no *peculiar velocity*). Their

physical separation is therefore given by $r(t) = a(t)\chi$ where χ is a fixed number equal to their separation today (where $a = 1$). The distance χ is called the *comoving distance*, which is just the separation of two objects with the effect of the expansion taken out. If the comoving separation χ is fixed, we can define a relative velocity given by the rate of change of proper distance $r(t) = a(t)\chi$, i.e.

$$v_r \equiv dr/dt = \dot{a}\chi = (\dot{a}/a)r \equiv Hr, \tag{4}$$

where dots denote time derivatives d/dt. Thus the relative velocity is proportional to the distance (if H is constant) — Hubble's law. This is independent of the position of the observer, and hence is consistent with the Copernican principle: all galaxies are moving apart from all other galaxies due to the expansion of space. If we lay down a uniform grid in space, then an observer will see all grid points around him recede or move towards him with a velocity proportional to the distance of the grid points. Although an observer might from this conclude that she is in the center of the universe, this is not true; homogeneity is fully respected since any other position in the grid would see the same thing. The expansion rate $H(t) \equiv \dot{a}/a$ is in general a function of time and called the *Hubble parameter*; its value today is the Hubble constant H_0. Since H varies in time, the Hubble law is really only useful for relatively nearby objects where it is roughly constant, $H \sim H_0$.

The universe is therefore thought to have been expanding fairly uniformly in the past; this implies that long ago it was much smaller, leading to the idea of an initial Big Bang 'explosion'.

C. The cosmological redshift

Recall that we can measure redshift using the known frequencies of atoms and molecules in the source rest frame. Consider a static source at a fixed radial comoving distance χ. For incoming light $a(t)d\chi = dr = -cdt$, and hence

$$\chi = \int_0^{\chi} d\chi' = -\int_{t_2}^{t_1} \frac{dt}{a(t)} = \int_{t_1}^{t_2} \frac{dt}{a(t)}. \tag{5}$$

We can think of t_1 as being the time a wave crest is emitted, and t_2 the time when it is received. We can also consider the next wave crest emitted a time interval $\delta t_1 = 1/\nu_1$ later in the source rest frame, which will reach $\chi = 0$ at a time $t_2 + \delta t_2$. Since the comoving distance of the source is assumed to be fixed we also have

$$\chi = \int_{t_1+\delta t_1}^{t_2+\delta t_2} \frac{dt}{a(t)} = \int_{t_1}^{t_2} \frac{dt}{a(t)}. \tag{6}$$

The only what that this can be true is if

$$\frac{\delta t_2}{a(t_2)} - \frac{\delta t_1}{a(t_1)} = 0 \quad \Longrightarrow \frac{\nu_1}{\nu_2} = \quad \frac{\delta t_2}{\delta t_1} = \frac{a(t_2)}{a(t_1)}. \tag{7}$$

Since we can observe the frequency of the radiation ν_2 when observed, and we know the source frequency ν_1, we can define a cosmological redshift between emission and observation, given by

$$1 + z \equiv \frac{\lambda_2}{\lambda_1} = \frac{\nu_1}{\nu_2} = \frac{a(t_2)}{a(t_1)}. \tag{8}$$

If $t_2 \geq t_1$ and the universe is expanding, as is the case for observation today of objects in the past, then $1 + z \geq 1$ and so the redshift is positive with $0 \leq z < \infty$: frequencies are observed to be redder than when they were emitted, and the ratio simply tells us how much smaller the universe was when the light was emitted. This makes sense, as one can think of the wavelength of the light stretching with the universe as it expands, so the total amount of stretching just gives the overall expansion ratio.

V. COSMOLOGICAL DISTANCES

We have already discussed redshift, which can be used as a kind of distance measure (if we know the source is static), though to relate redshifts to radial distance we need to know $a(t)$, so the radial distance is not directly observable. In order to test different cosmological models (which determine possible different $a(t)$ expansion histories), other handles on distance are very useful. There are two basic other ways in which distances can be measured: we can either consider the apparent angular size of an object and compare it to its known diameter, or we can measure the apparent observed luminosity and compare to a known (or modelled) source luminosity of an object. First we will recap on the comoving distance, then discuss the more directly observable distance measures.

1. Comoving distance

For purely radial distances, at any fixed time $a(t)\chi$ is the proper radial distance. The coordinate χ is the comoving distance, which has the overall expansion scale factored out, as previously explained.

For observations of objects in the universe, the relevant line-of-sight comoving distance is the distance travelled by the light since it was emitted. In time dt light travels physical distance cdt ($= dt$ in our units), hence comoving distance $d\chi = -dt/a$. Hence for radial rays the comoving distance travelled by light is determined by

$$\chi_s = \int_0^{\chi_s} d\chi = \int_{t_s}^{t_0} \frac{dt}{a(t)}, \tag{9}$$

where t_0 is the time today, and t_s is the time of emission from the source. This is easy to calculate in any given model that defines $a(t)$, but is not directly measurable. If the source has no peculiar velocity, χ_s is also the distance the object would have today, but of course we can only observe objects in the past.

2. Transverse comoving distance (comoving angular diameter distance)

This distance is useful for considering objects of small angular size transverse to the line of sight, and is defined so that the comoving size of an object (perpendicular to the line of sight) is given by $\delta\theta \times d_m$ where $\delta\theta$ is the angular size we observe on the sky and d_m is the comoving angular diameter distance.

Consider two light rays coming from either side of a small object, observed with angular separation $\delta\theta$. The two directions on the sky locally define a plane, and by symmetry (in the unperturbed background universe) we can expect the light rays to remain on that plane as they propagate through space. In Euclidean space the rays would just define a triangle, but of course space is expanding, so the rays are physically closer in the past than they would be in an un-expanding spacetime. But let's remove the expansion and thing about the comoving separation of the rays (the projection of the arrays into today's expanded space). Now again the simplest thing to expect is that they form a triangle in comoving distances, with comoving separation $\chi\delta\theta$.

However the "plane" defined by the rays can be curved in General Relativity, perhaps it is not a Euclidean plane after all? To preserve homogeneity all points in space should see the same behaviour, so the most general possibility is that the "plane" is in fact a constant-curvature surface. One example easily springs to mind: the surface of a sphere. The curvature at all points on a sphere is the same, i.e. it satisfies homogeneity. However on a sphere light rays form spherical triangles, i.e. they tend to converge. The rate of convergence depends on the radius of the sphere, so there is an extra curvature parameter $K > 0$ that determines the constant curvature (related to the radius of curvature). A less obvious possibility is a negative $K < 0$ which defines a *hyperbolic* space, sometimes visualized as the surface of a saddle. In this case the light rays diverge faster than in Euclidean space.

Imagine the case of the surface of a sphere. The length of an arc from the north pole is given by $\chi = r_c\theta$ where r_c is the radius. The separation of the end of two arcs with angle $d\phi$ at the north pole is $r_c \sin(\theta) d\phi = r_c \sin(\chi/r_c) d\phi$. Instead of r_c we can use the curvature parameter defined by $K = 1/r_c^2$, and hence the comoving angular diameter distance is given by $d_m = \sin(\chi\sqrt{K})/\sqrt{K}$.

The hyperbolic case is similar, but K is now negative and in general we can define

$$S_K(\chi) \equiv \begin{cases} \frac{1}{\sqrt{K}}\sin(\sqrt{K}\chi), & K>1 \quad \text{closed universe} \\ \chi & K=0 \quad \text{flat universe} \\ \frac{1}{\sqrt{|K|}}\sinh(\sqrt{|K|}\chi), & K<0 \quad \text{open universe} \end{cases} \tag{10}$$

so that the comoving angular diameter distance is $d_m = S_K(\chi)$.

If we imagine sending out two light beams separated by small angle $\delta\theta$, $d_m\delta\theta = S_K(\chi)\delta\theta$ would be the comoving separation of the beams once they have reached a comoving distance χ. In a flat geometry this is just the standard Euclidean result, $\chi\delta\theta$. However in a closed universe $S_K(\chi) < \chi$, and hence the beams are closer than they would be in a flat universe: the light rays are converging. So looking at objects in a closed universe is a bit like looking through a magnifying lens. Indeed at $\chi = \pi/\sqrt{K}$ the rays focus as the separation goes to zero, and at $\chi = 2\pi/\sqrt{K}$ they converge again at the point of emission having gone all the way round the universe[1]! Conversely in an open universe $S_K(\chi) > \chi$, and the light rays are diverging and become for ever (exponentially) further apart than they would be in a flat universe.

3. The (physical) angular diameter distance

We can now go back to the physical (non-comoving) angular-diameter distance. This is given simply by the result in the previous section converted back to physical units by multiplying by $a(t)$, hence the angular diameter distance is $d_A = a(t)d_m = a(t)S_K(\chi)$, where t is the time at which the object is being observed. See Figure 3.

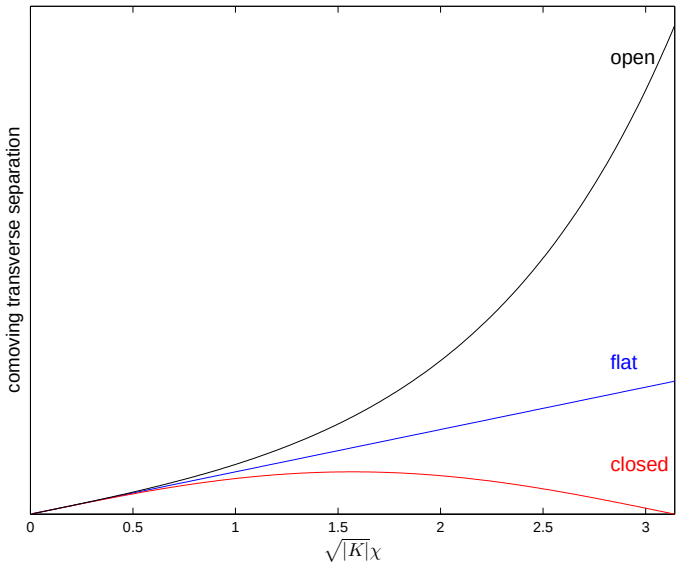

FIG. 2: Imagine sending out radial light rays with some small separation. This figure shows the comoving separation of the rays as a function of the radial comoving distance (in curvature radius units). In the closed universe the rays re-focus at the antipodal distance $\chi = \pi/\sqrt{K}$. In an open (and flat) universe the rays get forever further apart.

[1] Assuming of course the universe actually lasts long enough for light to have the time to go all the way round.

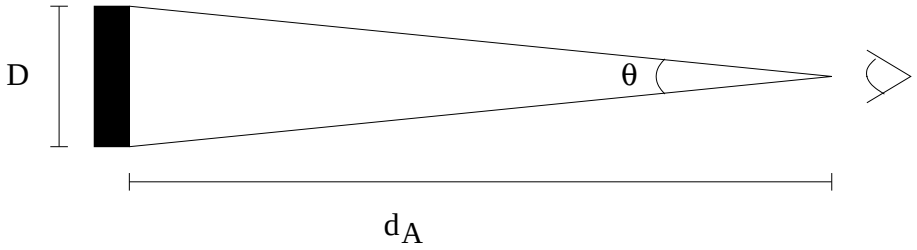

FIG. 3: Diagram of the definition of the angular diameter distance so that standard triangle relations apply: an object with known physical size D is observed to have angular size θ.

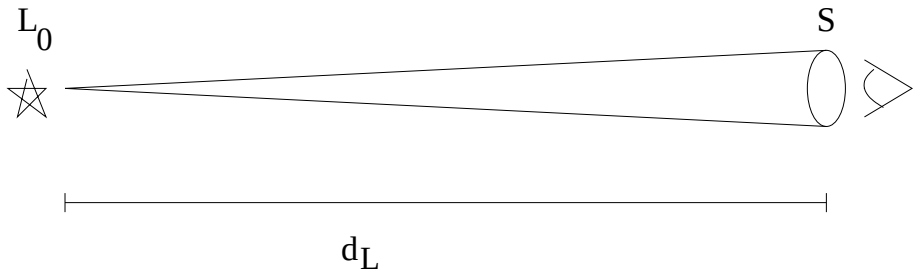

FIG. 4: Definition of the luminosity distance as an analogy to $1/r^2$ dimming in flat Euclidean space; a source with absolute luminosity L is observed on earth with Flux $S \propto 1/d_L^2$.

4. The luminosity distance

We assume that we know the intrinsic, absolute luminosity L of a certain object, called a *standard candle*, that radiates isotropically. Type Ia supernovae are believed to be approximate standard candles, and they are used to map out cosmological distances. Let us assume that we observe a flux S from the standard candle with absolute luminosity L at a fixed comoving distance. The definition of the luminosity distance d_L is then

$$S \equiv \frac{L}{4\pi d_L^2} \tag{11}$$

This is shown in Figure 4.

Imagine a pulse of a number of photons reaching us from the source. In the source rest frame the pulse lasts time δt_1 and contains energy $L\delta t_1$. The total number of photons reaching us per comoving area δA will be diluted as the photons move out over a spherical wavefront, so we only receive a fraction $\delta A/(4\pi d_m^2)$ of the photons. The energy of each photon ($h\nu$) is redshifted by a factor $1/(1+z)$ as the frequency is redshifted, so the total energy we will receive is $\delta E = \delta A \delta t_1 L/[(1+z)4\pi d_m^2]$. The flux (energy per unit time per unit area) at observation time t_2 is then

$$S = \frac{\delta E}{\delta A \delta t_2} = \frac{L}{(1+z)4\pi d_m^2}\frac{\delta t_1}{\delta t_2} = \frac{L}{(1+z)^2 4\pi d_m^2} \tag{12}$$

from Eq. (7). Hence the luminosity distance is given by

$$d_L = (1+z)d_m. \tag{13}$$

We see that there is a relation between the luminosity distance and the angular diameter distance, $d_L = (1+z)^2 d_A$. This actually also holds very generally (e.g. in general metrics) using a theorem called the *reciprocity relation* as long as no photons are lost along the line of sight (e.g. by scattering).

VI. FRIEDMANN EQUATION AND ENERGY CONSERVATION

A. Friedmann Equation

What determines the expansion rate $H(t)$ (and hence the scale factor $a(t)$)? This will depend on how much and what kind of stuff there is in the universe, an in general is given by solving Einstein's equations in general relativity. Here we just give a simple hand-waving argument based on Newtonian gravity for non-relativistic matter.

Consider a box filled with mass density ρ. From Gauss' theorem, the potential at distance r from any point is given by the mass enclosed, so

$$V(r) = -GM/r = -4\pi G\rho r^2/3 \tag{14}$$

The points should be moving apart, with kinetic energy in a spherical shell of mass m (on any particle in it) is given by

$$T = mv^2/2 = m\dot{r}^2/2. \tag{15}$$

The spherical shell has potential energy $V(r)m$ and so the total energy is

$$U = T + V = (\dot{r}^2/2 - 4\pi G\rho r^2/3)m \tag{16}$$

This energy should be conserved as the particles move apart. Let's take out the expansion using $r = a\chi$ giving

$$\frac{U}{ma^2\chi^2} = \frac{1}{2}\left(\frac{\dot{a}}{a}\right)^2 - \frac{4\pi G\rho}{3}. \tag{17}$$

The RHS is not a function of χ (homogeneity assumption), so the LHS cannot be either, and we can define a constant $K = -2U/(m\chi^2)$ so that

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{\kappa\rho}{3} - \frac{K}{a^2}. \tag{18}$$

where $\kappa \equiv 8\pi G$. This dodgy argument has in fact lead to the correct Friedmann equation, where K is precisely the curvature constant that we used before! The Friedmann equation tells us that the expansion rate is directly related to the curvature and density. It may seem unintuitive that universes with *higher* densities expand *faster* (H is bigger), for the same curvature; this is basically because if it didn't the curvature would instead have to be very different, and observations constrain the curvature to be small.

Note that this equation is for the smooth homogeneous background which applies to the universe on large scales. It does not mean small objects are expanding; for example the solar system is totally dominated by the gravitational force of the sun and has stable physical size (it is disconnected from the *Hubble flow* which expands dynamically uninteracting points).

B. Energy conservation

We assume an adiabatic expansion, so that the conservation of energy equation (first law of thermodynamics) is

$$dE = -PdV. \tag{19}$$

In terms of energy density ρ we can rewrite this using $H = (1/a)da/dt$ as

$$d(\rho a^3) = -Pd(a^3) \qquad \Longrightarrow \qquad \dot{\rho} + 3H(\rho + P) = 0, \tag{20}$$

since ρa^3 is proportional to the total energy, and a^3 to the volume. This is the energy conservation equation.

The relation between the pressure and the energy density (if one exists), $P = P(\rho)$ is called the *equation of state*. For (ideal) matter, radiation and vacuum energy it takes a very simple form,

$$P = w\rho, \quad \begin{cases} w = 0 & \text{"pressureless" matter('dust')} \\ w = 1/3 & \text{radiation} \\ w = -1 & \text{vacuum energy or cosmological constant} \end{cases} \tag{21}$$

C. The second Friedmann equation: acceleration equation

The second Friedmann equation can be obtained by differentiating the first one and using the energy conservation equation, giving

$$\left(\frac{\ddot{a}}{a}\right) = -\frac{4\pi G}{3}\left(\rho + 3P\right). \tag{22}$$

D. Equation summary

Any one of the three equation s can be derived from the other two. Written in terms of H and $\kappa = 8\pi G$ we can summarize these important results as

$$H^2 + \frac{K}{a^2} = \frac{\kappa}{3}\rho, \tag{23}$$

$$\dot{H} + H^2 = -\frac{\kappa}{6}(\rho + 3P), \tag{24}$$

$$\dot{\rho} + 3H(\rho + P) = 0. \tag{25}$$

E. Equation of state and fluid energy conservation

For the simple equation of state $P = w\rho$ with w constant we can also easily solve for $a(t)$ by writing

$$\frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{a}}{a} \quad \Longrightarrow \quad d\ln\rho = -3(1+w)d\ln a. \tag{26}$$

Integrating we find immediately that

$$\rho \propto a(t)^{-3(1+w)} \propto \begin{cases} a(t)^{-3} & \text{for } w = 0 \quad \text{(pressureless matter)} \\ a(t)^{-4} & \text{for } w = 1/3 \quad \text{(radiation)} \\ \text{const.} & \text{for } w = -1 \quad \text{(vacuum energy)} \end{cases} \tag{27}$$

For pressureless matter all the energy is in the mass, and the mass density simply dilutes with the volume scale $a(t)^3$. The energy density in radiation decreases more rapidly because as the universe expands the wavelength and hence energy is also redshifted $\propto 1/a(t)$. In an expanding universe, radiation will dominate at early times, while matter will become more important later on. If there is a non-vanishing contribution from vacuum energy, it will always start to dominate eventually (as shown in figure 5).

In general of course the universe will contain a mixture of things (e.g. a photons, dark matter, etc.). For all components that are not interacting, so they cannot exchange energy or momentum, the energy conservation equation will also apply to each separately

$$\dot{\rho}_i + 3H(\rho_i + P_i) = 0. \tag{28}$$

Note that the Friedmann equation involves the *total* energy densities and pressure, with $\rho = \sum_i \rho_i$ if there are multiple fluids. The simple result of Eq. 27 will hold for each uninteracting fluid, but the solution of the Friedmann equation will be more complicated since it involves the sum of energy densities that are evolving in different ways.

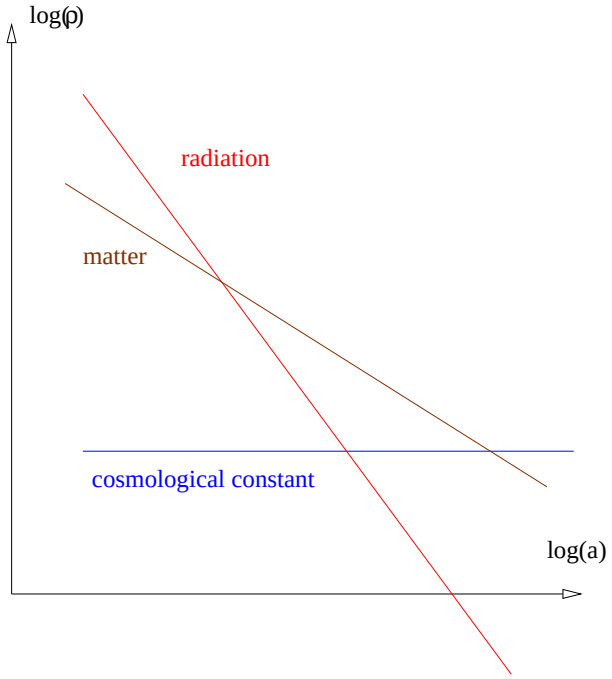

FIG. 5: The densities of the different ingredients of the universe (radiation, matter and a cosmological constant) as a function of scale factor.

VII. THE CRITICAL DENSITY AND THE GEOMETRY OF THE UNIVERSE

Looking at the first Friedmann equation, (23), we can re-write it as

$$\frac{K}{a^2H^2} = \frac{\kappa\rho}{3H^2} - 1 \equiv \frac{\rho}{\rho_c} - 1 \equiv \Omega - 1. \tag{29}$$

If the total energy density in the universe, ρ, is equal to the *critical energy density* $\rho_c \equiv 3H^2/\kappa$ then the curvature vanishes and the spatial sections of the universe is flat. In SI units $\rho_c = 1.88h^2 \times 10^{-26}\mathrm{kg\,m^{-3}}$, where $H_0 = 100h\,\mathrm{km\,s^{-1}Mpc^{-1}}$ and observations suggest $h \sim 0.7$ (so the universe is pretty empty on average!). In more cosmological units

$$\rho_c = 2.78h^{-1} \times 10^{11}\frac{M_\odot}{(h^{-1}\mathrm{Mpc})^3}, \tag{30}$$

so since $10^{11}M_\odot$ is around the typical mass of a galaxy, and galaxy separations are around 1Mpc, we can expect the density to be around this value (galaxies in fact turn out not to have enough mass on their own, but dark energy and non-galactic matter appear to make up the missing density so that actual the universe is close to flat).

Expressed with the *density parameter* $\Omega(t) \equiv \rho/\rho_c$:

$$\begin{aligned}
\Omega(t) > 1 &\;\Rightarrow\; K > 0 \Rightarrow \textbf{closed} \text{ universe} \\
\Omega(t) = 1 &\;\Rightarrow\; K = 0 \Rightarrow \textbf{flat} \text{ universe} \\
\Omega(t) < 1 &\;\Rightarrow\; K < 0 \Rightarrow \textbf{open} \text{ universe}
\end{aligned}$$

We will often split the density parameter into its constituents and write Ω_r for the contribution from radiation, Ω_m for the contribution by matter and Ω_Λ for the vacuum energy contribution, assuming these are the only relevant constituents of the universe. We can also define a "curvature" density parameter, $\Omega_K \equiv -K/(a^2H^2)$. The Friedmann equation becomes then simply

$$1 = \Omega(t) + \Omega_K(t) = \Omega_r(t) + \Omega_m(t) + \Omega_\Lambda(t) + \Omega_K(t). \tag{31}$$

Of course if the universe is flat, i.e. if $\Omega_K = 0$, then $\Omega = 1$ at all times.

There is an additional, useful way to rewrite the Friedmann equation (23), by writing out the dependence of the energy density on the scale factor explicitly. Normalizing so that $a(t_0) = 1$ today as usual, this leads to

$$H^2 = \frac{\kappa}{3}\left[\frac{\rho_r^{(0)}}{a^4} + \frac{\rho_m^{(0)}}{a^4} + \rho_\Lambda^{(0)}\right] - \frac{K}{a^2}. \tag{32}$$

With the critical energy density today given by $H_0^2 = \kappa\rho_c^{(0)}/3$ and by noting that $K = -\Omega_K^{(0)}H_0^2$ we can write

$$H(a)^2 = H_0^2\left[\frac{\Omega_r^{(0)}}{a^4} + \frac{\Omega_m^{(0)}}{a^3} + \frac{\Omega_K^{(0)}}{a^2} + \Omega_\Lambda^{(0)}\right]. \tag{33}$$

This provides an easy way to compute the Hubble constant as a function of the scale factor or the redshift. For the latter, we only need to use $a = (1+z)^{-1}$.

It should be noted that in the literature, the superscript (0) is normally dropped, so that e.g. Ω_m usually denotes the matter density *today* in terms of the critical density *today*. From now on we will often follow this notation and drop the (0) superscripts on the density parameters.

A. Acceleration

For nearby objects one can approximate H as a constant, however for objects at higher redshifts the universe can expand non-negligibly during the time that the photons need to reach us. In this case, the Hubble parameter does not remain constant. Observers often use an additional quantity, the acceleration parameter denoted as q_0 that determines the rate of change of the expansion today; it can be defined by a Taylor expansion about the time today t_0:

$$(1+z)^{-1} = a(t) = 1 + H_0(t-t_0) - \frac{1}{2}q_0H_0^2(t-t_0)^2 + \dots \tag{34}$$

where $q_0 = -\ddot{a}/(aH_0^2)|_{t=0}$. If $q_0 < 0$ then $\ddot{a}(t=0) > 0$ and the expansion is accelerating, otherwise it is decelerating. Generally one might expect the gravitational pull of matter to slow down the expansion, but in fact observations suggest that at low redshifts the expansion is actually accelerating.

Using the second Friedmann equation, (22), we find that

$$\frac{\ddot{a}}{a} = \frac{-\kappa}{6}(\rho + 3P). \tag{35}$$

So if $(\rho + 3P) > 0$, or $w > -1/3$ which is the "normal" case of matter and radiation, then $\ddot{a} < 0$ and the universe is decelerating. In the opposite case, it is accelerating. Specifically, $w = P/\rho < -1/3$ is required for acceleration.

Since $\rho = \Omega\rho_c$ where Ω is the total density parameter, defining w as effective equation of state parameter (or it can be written as the sum over all contributions) we have

$$\frac{\ddot{a}}{a} = \frac{-\kappa}{6}\rho(1+3w) = \frac{-\kappa}{6}\Omega(t)\rho_c(1+3w) = -\frac{1}{2}\Omega(t)H^2(1+3w). \tag{36}$$

The acceleration parameter $q_0 = -\ddot{a}_0/(a_0H_0^2)$ can then be expressed in terms of the equation of state today as,

$$q_0 = \frac{1}{2}\Omega_0(1+3w_0). \tag{37}$$

So if we measure $q_0 < 0$ as indicated by current data, then we need some exotic component with a negative equation of state, $P_{de}/\rho_{de} = w_{de} < -1/3$. Note that w_0 in the above equation is $P_{\rm tot}/\rho_{\rm tot}$, so since the current matter density is non-zero, neglecting radiation we need $w_0 = P_{de}/(\rho_m + \rho_{de}) < -1/3$ in order that $q_0 < 0$. A cosmological constant with $w_{de} = -1$ is a good fit to most current data, with $\Omega_m \sim 0.3$, $\Omega_\Lambda \sim 0.7$.

B. Age of the universe

It is straightforward to compute the age of the universe using the Hubble parameter:

$$t_0 = \int_0^{t_0} dt = \int_0^1 \frac{da}{\dot{a}} = \int_0^1 \frac{da}{aH(a)} = \int_0^\infty \frac{dz}{H(z)(1+z)}. \tag{38}$$

In general, we need to integrate the equation numerically. In the specific (non-realistic) case where the universe contains only matter, and $\Omega_m = 1$, $H^2 = H_0^2/a^3$, so that $H = H_0(1+z)^{3/2}$ in which case the equation integrates easily to give

$$t_0 = \frac{2}{3H_0}. \tag{39}$$

You can also calculate analytic results for matter and cosmological constant, or matter and radiation. On general dimensional grounds (H_0 has dimensions of inverse time) the age is usually given roughly by $t_0 = \mathcal{O}(1/H_0)$. The time H^{-1} is called the *Hubble time* and gives the characteristic time scale for the expansion.

Observationally we can easily set bounds on the age of the universe. We know the Earth formed about five billion years ago. Dating of isotopes in the galactic disk suggests around 10 billion years. The chemical evolution of old starts in globular clusters is also useful, as these turn out to be rather old: about 13 billion years. This is older than the estimate from $\frac{2}{3H_0} \sim 9$ billion years, which is evidence that in fact the universe does not contain only matter (a cosmological constant/dark energy component makes it older).

C. Special cases

In the following sections we use the Friedmann equations to study some special cases and to get a feeling how the universe evolves dynamically.

1. *The radiation dominated universe*

In the early universe, where the scale factor $a(t)$ is much smaller, radiation will dominate over matter, curvature and the cosmological constant, as its energy density scales with the highest negative power of the scale factor, $\rho_r \propto a^{-4}$. For $t \to 0$ we can thus approximate the first Friedmann equation by

$$H^2 = H_0^2 \frac{\Omega_r}{a^4}. \tag{40}$$

We can rewrite this as $a\dot{a} = H_0\Omega_r^{1/2} = \text{const.}$, or $ada \propto dt$, with the solution

$$a(t) \propto t^{1/2}. \tag{41}$$

2. *The matter dominated universe*

Currently the energy density in the radiation is negligible. We only have to take into account the matter, curvature and possibly a cosmological constant. We start by assuming that $\Omega_\Lambda = 0$ and that the universe is initially expanding. The equation to be solved is

$$H^2 = H_0^2 \left(\frac{\Omega_m}{a^3} + \frac{\Omega_K}{a^2} \right). \tag{42}$$

We will study the three cases of qualitatively distinct curvature:

1. Euclidian space, $K = 0$: The Friedmann equation is now $H = \dot{a}/a \propto a^{-3/2}$, which can be solved immediately to give

$$a(t) \propto t^{2/3}. \tag{43}$$

We can also notice that the expansion rate H is always positive, $\dot{a} \to 0$ for $t \to \infty$, so the universe expands forever, but ever more slowly. The age of the universe is easy to compute in this special case, as we computed earlier: $t_0 = 2/(3H_0)$. If we write $H_0 = 100h$km/s/Mpc it is $1/H_0 = 9.8 \times 10^9 h^{-1}$ years and so

$$t_0 = \frac{2}{3}\frac{1}{H_0} = 6.5/h\text{Gyr}. \tag{44}$$

A cosmological model which is flat and only contains matter is sometimes called the *Einstein-de Sitter* model.

2. Closed model, $K > 0$ ($\Omega_K < 0$): The Friedmann equation is now

$$\dot{a}^2 = \frac{\Omega_m H_0^2}{a} - K. \tag{45}$$

We see that there is a critical value of the expansion factor where $\dot{a} = 0$, at $a_{\max} = \Omega_m H_0^2/K$. This is when the universe changes from expansion to contraction. The universe ends in a "Big Crunch" when $a = 0$ is reached again in finite time.

[See question sheet for similar results] We can analytically solve the equation by setting $a = a_{\max}\sin^2\theta = a_{\max}(1 - \cos 2\theta)/2$, in which case the equation has solution

$$t = K^{-1/2}\int \frac{\sqrt{a}da}{\sqrt{a_{\max} - a}} = 2\frac{a_{\max}}{\sqrt{K}}\int d\theta \sin^2\theta = \frac{a_{\max}}{\sqrt{K}}(\theta - \cos\theta\sin\theta). \tag{46}$$

The "Big Crunch" is reached when $a = 0$, i.e. for $\theta = \pi$ at a time $t = \pi a_{\max}/\sqrt{K}$. For $\theta \ll 1$ we have $\theta - \cos\theta\sin\theta \approx 2\theta^3/3 \propto a^{3/2}$, recovering the previous result that $a \propto t^{2/3}$. As expected, the curvature contribution is subdominant at early times.

Note that we have only shown there is a big crunch for a universe made of matter; an additional dark energy contribution can avoid re-collapse.

3. Hyperbolic model, $K < 0$: In this case we also have to solve the equation

$$\dot{a}^2 = \frac{\Omega_m H_0^2}{a} - K, \tag{47}$$

but since $-K$ is positive, in this case $\dot{a} \geq 0$ at all times, so the universe will expand forever. Defining $\bar{a} = \Omega_m H_0^2/|K|$ this time we use the equivalent hyperbolic substitution $a = \bar{a}\sinh^2\theta$

$$t = |K|^{-1/2}\int \frac{\sqrt{a}da}{\sqrt{\bar{a} + a}} = 2\frac{\bar{a}}{\sqrt{|K|}}\int d\theta \sinh^2\theta = \frac{\bar{a}}{\sqrt{|K|}}(\cosh\theta\sinh\theta - \theta). \tag{48}$$

Again for $\theta \ll 1$ we recover once more $a \propto t^{2/3}$.

3. Non-zero cosmological constant

If $\Omega_\Lambda = 1$ and $\Omega_r = \Omega_m = \Omega_K = 0$, then the first Friedmann equation becomes

$$\dot{a}^2 = \frac{1}{3}\Lambda a^2 \tag{49}$$

which has the solution

$$a(t) = \exp(Ht), \quad H = \sqrt{\frac{\Lambda}{3}}. \tag{50}$$

In this model, also known as the *de Sitter* model, the universe is completely dominated by a cosmological constant, and the scale factor grows exponentially. Both the acceleration parameter $q = -1$ and the Hubble constant do not change over time.

Since the energy density due to the cosmological constant is constant, but other densities dilute with the expansion, models with a cosmological constant tend to have a cosmological constant dominated state as their end point. The model is also relevant for the early universe in the model of exponential expansion during *inflation*.

Let's now discuss qualitatively the realistic cases with matter, curvature and Λ contributions:

$$\dot{a}^2 = \frac{\Omega_m H_0^2}{a} - K + \frac{\Lambda}{3}a^2. \tag{51}$$

1. $K < 0$ **:** If $\Lambda > 0$ then $\dot{a}^2 > 0$ at all times, with a minimum expansion rate $\dot{a}$ at $a_{\min} = (3\Omega_m H_0^2/2\Lambda)^{1/3}$. For small values of a (at early times) Eq. (51) reduces to the hyperbolic case with vanishing cosmological constant, and $a \propto t^{2/3}$: matter still dominates at early times. At late times the cosmological constant dominates and $a \propto \exp(Ht)$. If $\Lambda < 0$ then there exists an a_c where $\dot{a} = 0$, so the universe will expand until a_c and then collapse again.

2. $K = 0$ **:** This case behaves similar to the previous one, but Eq. (51) can be solved explicitly. We find

$$a(t)^3 = \frac{3\Omega_m H_0^2}{\Lambda} \sinh\left(\frac{\sqrt{3\Lambda}}{2}t\right)^2 \qquad (\Lambda > 0) \tag{52}$$

$$a(t)^3 = \frac{3\Omega_m H_0^2}{|\Lambda|} \sin\left(\frac{\sqrt{3|\Lambda|}}{2}t\right)^2 \qquad (\Lambda < 0). \tag{53}$$

3. $K > 0$ **:** This is a rather complicated case. There can now be a solution for $\dot{a} = 0$, as determined by the relevant cubic equation. The number of real solutions depends on the relative sizes of the different terms. For $\Lambda < 0$ there is always a solution, so the universe will re-collapse eventually. For $\Lambda > 0$ there can be a solution depending on the relative size of the terms, but usually the universe expands forever unless the matter density is high enough to re-collapse the universe before the cosmological constant term comes to dominate.

VIII. CONTENT AND DENSITIES IN THE UNIVERSE

A. Baryons

In cosmology, 'baryons' refers to the mass in atoms and ionized plasmas (including electrons, though they are only a small fraction of the mass). This is the stuff of normal matter, and protons and electrons are sufficiently heavy that the baryons almost always move non-relativistically (for cosmological purposes on large scales they can be treated as 'cold' or 'dust' - pressureless matter, though the pressure becomes important when thinking about gravitational collapse into galaxies.).

Note that when atoms are ionized there is a very strong Coulomb force, so electrons and protons do not get far separated; for this reason one often talks about ionized baryons as a single fluid, which electrons and protons moving around together.

Adding up the mass in stars gives $\Omega_{stars} \sim 0.01$, a small fraction of the critical density. Gas that is not in stars is though to be about 5 times larger with in total $\Omega_b \sim 0.05$, or $\Omega_b h^2 \sim 0.022$.

The fact that baryons massively outnumber anti-baryons requires some kind of asymmetry between baryons and anti-baryons in the very early universe. The process by which this asymmetry arose in unknown, and known as *baryogenesis*; possibilities include GUT-scale physics, though the required baryons number violation can also be non-perturbatively generated in the electroweak standard model.

B. Photons

We see the radiation in the microwave background. In quantum terms ,we can think of it as made up of photons, symbol γ. Light interacts only weakly with atoms, but if there are free electrons

(in an ionized plasma), there can be a much stronger interaction (Compton scattering, or Thomson scattering in the common non-relativistic case); in the early ionized universe the photons and baryons are tightly coupled together.

Photons are spin-1 (gauge) *bosons*, and are their own antiparticle. They can have two spins, up and down (or left and right polarization, as you prefer).

C. Neutrinos

Some neutrinos eigenstates are known to have a small mass (from neutrino oscillation experiments), but all the masses are rather small, probably $\lesssim 1\,\mathrm{eV}$. They interact only via the weak interaction (which is weak!), and come in three flavours: electron, muon and tau neutrinos. They are important for cosmology largely because as we shall see they make a significant contribution to the energy density in the universe at early times.

We know neutrinos must be present in large numbers because as we show later the early universe is very hot, and due to the weak interaction at some point photons and neutrinos would have been in equilibrium, leading to a relic neutrino background today (much like the CMB is the relic photon background, but unfortunately the neutrinos are essentially unobservable).

Neutrinos are spin-1/2 *fermions*, and hence obey the Pauli Exclusion Principle — this will be important later when we study their distribution in more detail. There are also anti-neutrinos. However neutrinos are always left-handed, so there are only two distinct types for each flavour: one left-handed neutrino and one right-handed anti-neutrino.

D. Dark matter

Other non-relativistic stuff that does not interact with baryons or light (or at least only very weakly), is generically called dark matter: it's stuff that is important gravitationally but does not otherwise interact.

Does dark matter exist? On theoretical grounds there's no particular reason why not - why should all the stuff in the universe be easy to see? There are plenty of ways to make weakly-interacting particles, e.g. in supersymmetric theories. If the dark matter particles are very heavy, so that their velocities are negligible and the matter has no pressure, it is called *cold dark matter*. If the pressure is non-negligible it is called *warm* dark matter (or *hot* dark matter if relativistic). Current data points strongly towards the dark matter being rather cold.

In fact there is a lot of evidence for dark matter.

1. Rotation curves

Perhaps the most direct is galaxy rotation curves. If a galaxy is rotating, we can measure the rotation speed (projected along our line of sight) using the relative Doppler shift. This can be compared to what you'd expect from Newtonian gravity: equating centrifugal and gravitational forces gives

$$\frac{v^2}{r} = \frac{GM(r)}{r^2} \qquad \Longrightarrow \qquad v = \sqrt{\frac{GM(r)}{r}}. \tag{54}$$

If we look at the edge of a rotating galaxy most of the light is near the centre (it typically falls nearly exponentially at large radii), so if the light traced the matter we would have $M(r) \sim$ const and hence expect $v \propto r^{-1/2}$. However observations actually give $v \sim$ constant. This implies that actually $M(r) \propto r$, even though the light that we can see is falling off rapidly with radius. Assuming that standard gravity is correct, this implies that there is actually a lot more matter at large radii that we can't see. If it is non-interacting, it would not have formed a disk (like the visible galaxy), and hence is expected to form a nearly spherical halo.

The rotation curve argument does not rule out baryonic matter that happens to be dark (brown dwarves or similar), but other lines of evidence suggest the missing matter cannot be mostly baryonic.

2. Galaxy clusters

Clusters are gravitationally bound systems, about 90% of the baryons in gas and the rest in galaxies. The gas temperature can be measured by X-ray observations, and from this the pressure and gas mass to be inferred (using some assumptions). Do gravitational attraction and pressure balance to make a stable cluster? The answer is no: the gravitational mass must be larger than inferred from the gas, with about 10 times the mass of dark matter as baryons. Since clusters are very large, when they form you'd expect the ratio of baryons and dark matter to be roughly the same as the universe as a whole, so this provides evidence that most of the mass is not in baryons.

3. Lensing

Matter perturbations cause gravitational bending of light: gravitational lensing. This causes arcs, as seen in famous Hubble images, and also small distortions to all observed galaxy shapes. Measuring the amount of distortion, it is possible to estimate the mass. Since lensing is due to the *total* mass, this can be compared to mass inferred e.g. from X-ray gas. In fact in the case of the 'Bullet' Cluster the gas and the mass seem to be in completely different places, which can be taken as 'direct' evidence for dark matter.

4. Structure formation

Details calculation of how large-scale structures in the universe grow, and the expected form of anisotropies in the CMB, both also suggest that most of the matter is dark, with $\Omega_c \sim 0.2$—0.3.

5. What is the dark matter?

It can't be standard neutrinos, they are too light (would stream out of galaxies quickly).
Options include:

- *WIMPS* (weakly interacting massive particles): The lightest supersymmetry particle (LSP) is typically stable is supersymmetric models, and hence could make a good dark matter candidate since they tend to be massive and interact only very weakly with normal matter and light: things like the photino, gravitino and neutralino.
- *MACHOs* (massive compact halo objects): typically stellar mass objects that are dark (baryonic or non-baryonic), e.g. brown dwarfs. These could be detected by micro-lensing, and some definitely exist - but not at the densities required to explain most of the dark matter.
- *modified gravity* Some of the arguments for dark matter are based on gravitational calculations: perhaps gravity is different on the relevant scales? Suggestions include 'MOND', but these days have difficulty fitting all the data, esp. the bullet cluster where there is a clear separation of the lensing mass from the gas.
- *new ideas?*

Because by definition dark matter is weakly interacting, it is hard to detect directly. However if they interact via the weak force (e.g. WIMPs), they may just be detectable in the lab. The idea is to observe a really large chunk of mass and look for the rare interactions due to the dark matter particles streaming through the earth (being careful to exclude other events, like radioactive decays). Annihilations or decays could also potentially be observed e.g. via the production of gamma rays in our galaxy. Currently there is no convincing signal.

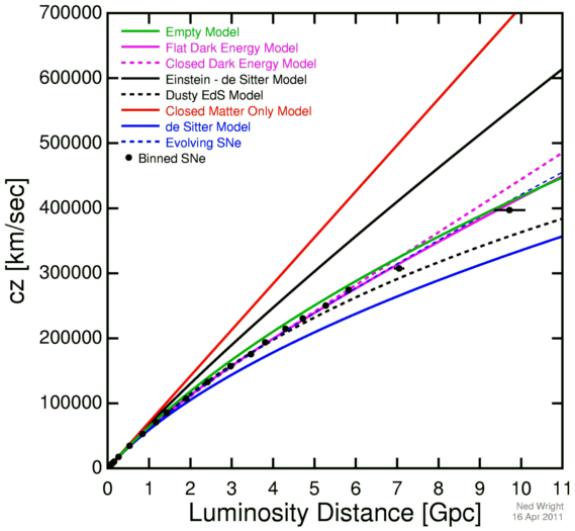

FIG. 6: Fits to the observed inferred luminosity distance of Type Ia supernovae against of redshift. The luminosity distance depends on $H(z)$, and the data strongly prefer a model with dark energy (pink). The luminosity distances are inferred from observed supernovae luminosities by assuming that the intrinsic luminosities of the supernovae are *standard candles* (or, more accurately, standardizable candles) - i.e. we know the intrinsic luminosity so the observed flux directly depends on the luminosity distance. *From* `http://www.astro.ucla.edu/~wright/sne_cosmology.html`.

E. Dark energy

This is the mysterious stuff that seems to provide about 70% of the energy density today making the universe nearly flat (rather than open if it only had $\Omega_m \sim 0.3$ in the baryons and dark matter), and making the expansion accelerate. Observations of supernovae luminosities indicate accelerated expansion predicted by dark energy, and and the age of globular clusters indicate that the universe must be older than predicted if it only contains matter. There is also significant other evidence from the abundance of clusters, the distribution of anisotropies in the microwave background, the growth structure at low redshift, and other observations.

Supernovae Type Ia are able to probe the luminosity distance out to $z > 1$, because the luminosities are thought to be 'standardizable', in that from the observations you can calculate the luminosity distance d_L. People often discuss the effective magnitude, $m = \log d_L + \text{const}$, where the constant is conventional, so that since the SN-1a redshifts are also measured, there is effectively a measurement of $\log d_L(z)$ to within a constant. This can be compared with the expected scaling with redshift in different cosmologies. We showed before that for a universe filled with different perfect fluids, the luminosity distance for a source a redshift z is given by

$$d_L(z) = (1+z)S_K(\chi_z) = (1+z)S_K\left(\int_0^z \frac{dz'}{H(z')}\right) \tag{55}$$

(where in $S_K(\chi)$ we used the comoving χ that light has travelled, $\chi = \int_{t_z}^{t_0} dt'/a = \int_{a_z}^{1} da/Ha^2 = -\int_z^0 dz/H$), and the Hubble parameter is given (in the case of a cosmological constant) by

$$H(z)^2 = H_0^2\left[\Omega_r(1+z)^4 + \Omega_m(1+z)^3 + \Omega_\Lambda + \Omega_K(1+z)^2\right]. \tag{56}$$

The Ω_r term is negligible for $z \sim \mathcal{O}(1)$. Figure 6 shows the luminosity distance as measured by the SN-Ia compared to various possible models, which strongly favours a dark energy model.

Some possibilities include:

- *cosmological constant*: Perhaps the simplest explanation: a uniform vacuum energy that fills space, with negative pressure ($w = -1$). The vacuum energy is very small, much smaller (many

orders of magnitude) than would be expected in the standard model vacuum, which constitutes the *cosmological constant problem*. The default assumption would be that some as-yet unknown high-energy physics regularizes or cancels the main contributions, leaving the small residual we see.

- *quintessense*: This can behave similar to a cosmological constant, but now $w \geq -1$ and in general w varies in time. This behaviour can be produced by a cosmological scalar field, but it doesn't help with the cosmological constant problem (except than it could in principle allow for a dynamical mechanism to drive the energy density to very low values; in practice this is difficult).
- *modified gravity*: We only detect the effects of dark energy on really large scales (larger than clusters), so we cannot easily rule out some modification to gravity (i.e. not General Relativity) at least on large scales (explaining dark matter this way was much harder). However there are no very compelling models, and again the cosmological constant problem remains.

Observations currently mostly give $w \approx -1 \pm 0.1$, so quite consistent with a cosmological constant.

F. Summary

Data suggests that today the energy in the universe is about 1% in luminous baryonic matter (e.g. stars), and 5% ($\Omega_b \sim 0.05$) in total including cool gas that we can't see directly. There's much more (nearly cold) dark matter [$\Omega_c \sim 0.2$] and most of the energy density today is in dark energy ($\Omega_\Lambda \sim 0.75$). Today the photons and neutrinos are cool (CMB temperature is only 2.726K), and hence have low energy and only contribute $< 10^{-4}$ to the total energy density.

Of course the energy densities scale very differently with redshift, so at earlier times the balance will be quite different (initially radiation dominated, then matter dominated, and only recent dark energy dominated).

IX. THE THERMAL HISTORY OF THE UNIVERSE

Although the treatment of the contents of the universe as perfect fluids allowed us to derive the basic equations governing the expansion of the universe, it is not sufficient to study in detail the behaviour and interactions of realistic particles at high temperature. For this we need to use statistical mechanics, which tells us the equilibrium probability distribution of the particles and hence how the bulk properties of all the particles behave.

A. Review of equilibrium distributions

If particles interact often and exchange energy, they will rapidly reach an equilibrium state where at any given time each possible micro-configuration with the same total energy is equally likely. For example, for a gas in a box any give atom is just as likely to be in one position in the box as any another. Many of the micro-configurations are however macroscopically indistinguishable, i.e. can be described by a macrostate with a particular temperature T, total number of particles N, etc. The most likely macroscopic state is the one in which has the largest fraction of the possible microstate configurations for the system (subject to any constraints, like conservation of energy or particle number), and this is the state the system will almost always be in when it has reached equilibrium. For example, the macroscopic state corresponding to having 50 coins heads and 50 coins tails is vastly more likely in a distribution of 100 coin tosses than a macroscopic state corresponding to 100 heads, because there are vastly more sequences of tosses that give a 50/50 result than an unlikely sequence of 100 heads. Likewise gas particles in a box could all be in one corner, but there are vastly more ways of arranging them dispersed throughout the box, and hence the latter is what we expect to see if the particles are free to move around and randomly interact.

In the case of particles, the form of the most likely distribution depends on whether the particles are bosons (integer spin, like the photon), or fermions (half-integer spin, like neutrinos, which obey

the Pauli exclusion principle and hence cannot have more than one particle in each distinct quantum state).

Consider a set of energy levels, each having energy ϵ_i and occupied by n_i particles. The levels may be degenerate, in that there are g_i distinct quantum states (sub-levels) all of the same energy. For example in the case of a gas of particles the energy is determined by the particle momentum, and there are many states of the same energy because the particle could be in many possible different locations, and the momentum could be in many different directions. Statistical mechanics applies to large systems, so that the occupation probability of any quantum is independent of the number of states that exist. We can therefore consider large g_i and calculate what fraction of these states are occupied in order to calculate the average occupation number.

Consider having n_i identical fermions, so that each distinct quantum state can only have either 0 or 1 fermion in it. We now want to find the most likely distribution of the fermions amongst the available energy levels, which depends on the number of different ways that you can arrange the particles in the levels[2]. The number of ways of putting n_i identical particles into g_i distinct states is given by the binomial coefficient

$$w_f(n_i, g_i) = \frac{g_i!}{n_i!(g_i - n_i)!}. \tag{57}$$

For bosons each state can have more than one particle in, and calculating the number of ways the states can be more populated is more tricky. Consider writing down a list of the particles and grouping them into the different states, with a boundary separating particles in different states (and consecutive boundary lines if there are no particles in a given state). There are $g_i - 1$ boundaries between each group of particles in this list (because there are g_i sub-levels), and the total number of particles is n_i. The number of ways this could happen is the number of ways of choosing $g_i - 1$ boundaries and n_i particles from a set of $n_i + g_i - 1$ boundaries and particles; this gives[3]

$$w_b(n_i, g_i) = \frac{(n_i + g_i - 1)!}{n_i!(g_i - 1)!}. \tag{58}$$

Given a set of occupation numbers $\{n_i\}$ for each level, the total number of ways the levels and sub-levels can be populated is

$$W = \prod_i w(n_i, g_i). \tag{59}$$

We now want to find the most likely distribution (that with the largest W), subject to the constraint of fixed energy $E = \sum_i n_i \epsilon_i$ and number of particles $N = \sum_i n_i$ in a fixed large volume. We can do this with Lagrange multipliers[4], i.e. maximize

$$f \equiv \ln(W) + \alpha(N - \sum_i n_i) + \beta(E - \sum_i n_i \epsilon_i). \tag{60}$$

In general maximizing using $\frac{\partial f}{\partial n_i} = 0$ gives

$$\frac{\partial \ln W}{\partial n_i} = \alpha + \beta \epsilon_i. \tag{61}$$

For the cases in hand, in the large g limit we can use Stirling's approximation $\ln n! \approx n \ln n - n$. For Fermions this gives

$$\ln W \approx \sum_i \left[-n_i \ln n_i - (g_i - n_i) \ln(g_i - n_i) + g_i \ln g_i \right] \tag{62}$$

---

[2] This corresponds to maximizing the entropy. In statistical mechanics the entropy is $S \equiv -k_B \sum_i P_i \ln P_i$, where P_i is the probability of being in the ith micro-configuration of the system (microstate). We assume each microstate is equally likely to be occupied, so that $P_i = 1/W$ where W is the number of microstates (W is the number of distinct ways of arranging things). In this case $S = k_B \ln W$.

[3] If you are confused, see `http://en.wikipedia.org/wiki/Bose-Einstein_distribution` first notes section.

[4] For an introduction see `http://en.wikipedia.org/wiki/Lagrange_multiplier`

and hence the maximum is for $\hat{n}_i$ where

$$\frac{\partial \ln W}{\partial n_i} = -\ln \hat{n}_i + \ln(g_i - \hat{n}_i) = \alpha + \beta\epsilon_i \tag{63}$$

which rearranges to give

$$\hat{n}_i = \frac{g_i}{e^{\alpha+\beta\epsilon_i} + 1} \qquad \text{(fermions)}. \tag{64}$$

For bosons with $g_i \gg 1$ so $g - 1 \approx g$ we have

$$\ln W \approx \sum_i \left[(n_i + g_i)\ln(n_i + g_i) - (n_i + g_i) - n_i \ln n_i + n_i - g_i \ln g_i + g_i\right] \tag{65}$$

which gives the maximum at

$$\hat{n}_i = \frac{g_i}{e^{\alpha+\beta\epsilon_i} - 1} \qquad \text{(bosons)}. \tag{66}$$

For large g_i we expect the distribution to be symmetric, so the maximum is also the mean, and the probability that any sub-level will be occupied is given the large-g_i fraction $\hat{n}_i/g_i$:

$$\mathcal{N}_i = \frac{1}{e^{\alpha+\beta\epsilon_i} \pm 1}, \tag{67}$$

with + for fermions and − for bosons.

The distribution is determined by two constant α and β, which can be used as the thermodynamic variables labelling the distinct macrostates. As such they must be related to the usual thermodynamic state variables (temperature and chemical potential). Also, the number of ways of arranging things is conventionally measured by the *entropy*, defined with a convenient constant so that $S = k_B \ln W$. The mostly likely state, the equilibrium state, is therefore the state of maximum entropy (subject to the constraints).

Note that since the constraints must be satisfied $S = k_B \ln(W) = k_B f$. From Eq. 60 we then see that

$$\frac{\partial S}{\partial E} = k_B\beta, \qquad \frac{\partial S}{\partial N} = k_B\alpha. \tag{68}$$

So α describes how the entropy responds to changes in the total number of particles, and β describes the response to changes in total energy. Eq. (68) can then be used give a statistical mechanical *definition* of the temperature and chemical potential in terms of β and α, and hence the response of the entropy.

We can also relate α and β to familiar classical thermodynamic quantities by comparing with the relation

$$\mathrm{d}E = T\mathrm{d}S - P\mathrm{d}V + \mu\mathrm{d}N \quad \Longrightarrow \quad \mathrm{d}S = \frac{1}{T}\left(\mathrm{d}E + P\mathrm{d}V - \mu\mathrm{d}N\right), \tag{69}$$

so that for a fixed volume using Eqs. (68) we see that

$$\left.\frac{\partial S}{\partial E}\right|_{V,N} = k_B\beta = \frac{1}{T}, \qquad \left.\frac{\partial S}{\partial N}\right|_{V,E} = k_B\alpha = -\frac{\mu}{T}. \tag{70}$$

Hence β is related to the temperature, and α to the chemical potential (and temperature):

$$\beta = \frac{1}{k_B T}, \qquad \alpha = \frac{-\mu}{k_B T}. \tag{71}$$

Another way to see the relation to the classical quantities is using Eq. (61): assuming the energy levels do not change we then have[5]

$$\mathrm{d}S = k_B \mathrm{d}\ln W = k_B \sum_i \frac{\partial \ln W}{\partial n_i} \mathrm{d}n_i = k_B \sum_i (\alpha + \beta\epsilon_i)\mathrm{d}n_i = k_B(\alpha dN + \beta dE). \tag{72}$$

Hence comparing coefficients with Eq. (69) at fixed volume gives Eq. (71).

Finally we can now go back and write the equilibrium occupation number of Eq. (67) directly in terms of the temperature and chemical potential as

$$\mathcal{N}_i = \frac{1}{e^{(\epsilon_i - \mu)/k_B T} \pm 1}. \tag{73}$$

For brevity we will often use units with $k_B = 1$; remember to put k_B back in where required to make the dimensions correct when calculating numerical answers.

For a particle in a large volume we can calculate the relevant density of states using the fact that in a box of side L the wavelengths in each direction are quantized so that $\lambda_i = 2L/n_i$, where n_i is integer. In terms of momentum $|\mathbf{p}| = h\nu = \hbar 2\pi c/\lambda = 2\pi/\lambda = 2\pi\sqrt{n_x^2 + n_y^2 + n_z^2}/2L$ in natural units. In n-space, the states are spaced in a cubic grid with unit grid point separation. However the n-space points are only in the range where $n_x, n_y, n_z \geq 0$, but $\mathbf{p}$ can point in any direction, so there are eight points in momentum space for every point in the positive octant of n-space. So the number of states per unit momentum volume is $\frac{1}{8} \times \left(\frac{2L}{2\pi}\right)^3$. Over a particular range of momenta and positions the number of states is therefore

$$dg_{\mathbf{p}} = \frac{1}{8}\frac{(2L)^3 d^3\mathbf{p}}{(2\pi)^3} = \frac{d^3\mathbf{p}d^3\mathbf{x}}{(2\pi)^3}. \tag{74}$$

For a particle A in addition there may be a spin-degeneracy factor g_A.

B. Distribution function

For a particle species A (with mass m) in statistical equilibrium, the number density n, energy density ρ and pressure P are given as integrals over the distribution function $f_A(\mathbf{x}, \mathbf{p}, t)$. This is defined so that in a 3-momentum element $d^3\mathbf{p}$ and spatial volume element $d^3\mathbf{x}$ there are $f_A(\mathbf{x}, \mathbf{p}, t)d^3\mathbf{p}d^3\mathbf{x}$ particles, where $\mathbf{p}$ is the 3-momentum. We only consider the homogeneous case here, so $f_A(\mathbf{x}, \mathbf{p}, t) = f_A(\mathbf{p}, t)$, and statistical isotropy implies $f_A(\mathbf{p}, t) = f_A(|\mathbf{p}|, t) \equiv f_A(p, t)$. We will mainly leave the time dependence implicit, which will be manifest via the temperature dependence of the equilibrium distribution function. Particles of different species will be interacting constantly, exchanging energy and momentum. If the rate of these reactions $\Gamma(t) = n\langle\sigma v\rangle$ (where σ is the cross-section and v is the rms velocity) is much higher than the rate of expansion $H(t)$, then these interactions can produce and maintain thermodynamic equilibrium with some temperature T. Therefore, particles may be treated as an ideal (Bose or Fermi) gas, with the equilibrium distribution function $f_A d^3\mathbf{p}d^3\mathbf{x} = \mathcal{N}_{\mathbf{p}} g_A dg_{\mathbf{p}}$ so that

$$f_A(p) = \frac{g_A}{(2\pi)^3}\frac{1}{e^{(E_A - \mu_A)/T_A} \pm 1} \tag{75}$$

where g_A is the spin degeneracy factor, μ_A is the chemical potential, T_A is the temperature of this species and $E(p) = \sqrt{p^2 + m^2}$, where $p = |\mathbf{p}|$. The "+" sign corresponds to fermions, and the "-" sign to bosons, and we are using units with Boltzmann constant $k_B = 1$.

---

[5] If the energy levels can change $\mathrm{d}E = \sum_i (n_i \mathrm{d}\epsilon_i + \epsilon_i \mathrm{d}n_i)$ so that $\sum_i \beta\epsilon_i \mathrm{d}n_i = \beta \mathrm{d}E - \beta \sum_i n_i \mathrm{d}\epsilon_i$, where $\sum_i n_i \mathrm{d}\epsilon_i = \mathrm{dWork} = -P\mathrm{d}V$ is the work done on the system, consistent with the usual thermodynamical result when the volume is not fixed.

C. Chemical potential

The chemical potential may not be very familiar, and for a given system in general is unknown; however we know some things about it.

If the number of particles is not constrained (so that chemical as well as kinetic equilibrium is obtained), we do not need the $\sum_i n_i = N$ Lagrange multiplier, i.e. $\alpha = \mu = 0$ and the chemical potential is zero. For example in the very early universe photons are not conserved (double Compton scattering $e^- + \gamma \leftrightarrow e^- + \gamma + \gamma$ happens in equilibrium at high temperatures), so the number of photons can change to maximize the entropy. The maximum is where $\partial S/\partial N = 0$, and hence from Eq. (70) $\mu_\gamma = 0$.

For any interaction between particles that takes place frequently in the equilibrium (where $dS = 0$), we must also have

$$dS = \sum_i \frac{\partial S}{\partial N_i} \mathrm{d}N_i = -\sum_i \frac{\mu_i}{T} \mathrm{d}N_i = 0, \tag{76}$$

and hence $\sum_i \mu_i dN_i = 0$ (since the temperatures must also be the same in equilibrium). If this were not the case the particles could convert into each other to increase the entropy further.

If different species are in chemical equilibrium through the reactions $A + B \rightleftharpoons C + D$, then the chemical potentials satisfy $\sum_i \mu_i \mathrm{d}N_i = -\mu_A - \mu_B + \mu_C + \mu_D = 0 \implies \mu_A + \mu_B = \mu_C + \mu_D$[6]. This can be used to relate unknown chemical potentials to each other. For example as we mentioned photons are not conserved at high temperature, so we know[7] $\mu_\gamma = 0$, and if pair production and annihilation takes place, eg. $\mathrm{e}^- + \mathrm{e}^+ \leftrightarrow \gamma + \gamma$, then the particle and antiparticle have equal and opposite chemical potentials, e.g. $\mu_{\mathrm{e}^-} = -\mu_{\mathrm{e}^+}$.

D. Blackbody spectrum

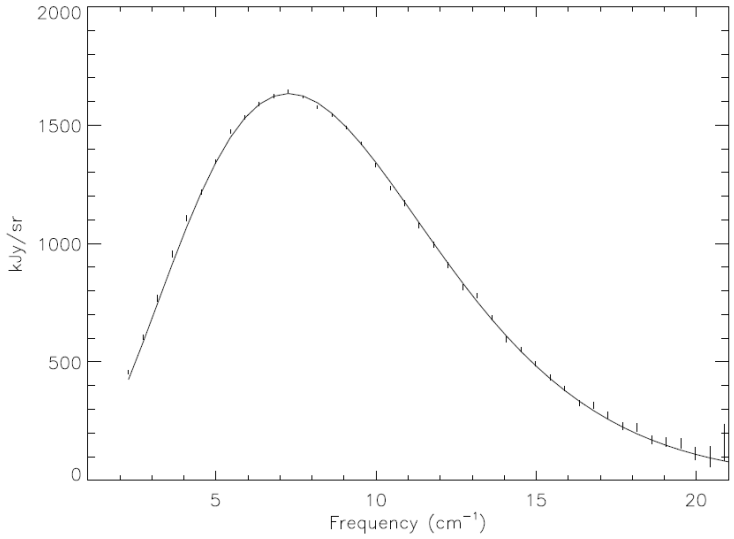

FIG. 7: Measurements of the CMB intensity as a function of frequency by COBE's Firas experiment. The data are in good agreement with a blackbody spectrum (solid line) with $T = 2.726K$. *From* arXiv:astro-ph/9605054

---

[6] You can also derive this by considering the constraints (and hence Lagrange multipliers) $A + B \rightleftharpoons C + D$ imposes on the individual numbers and energy of the joint system; see e.g. the Mukhanov book, Sec 3.3.

[7] When the universe cools enough ($z \lesssim 2 \times 10^6$) double Compton scattering is inefficient, and it's possible for a "μ-distortion" to develop where μ_γ becomes non-zero if new energy is injected; see e.g. arXiv:1201.5375 and refs therein.

Of particular importance are the photons in the early universe and observed today in the CMB. The Firas experiment on the COBE satellite measured the energy spectrum of the CMB with high accuracy. If the photons originated from a thermal equilibrium distribution in the early universe what would we expect to see? Since photons are bosons with $\mu = 0$, and including the two photon spins (polarizations) we have the equilibrium distribution function

$$f_\gamma(E) = \frac{2}{(2\pi)^3}\frac{1}{e^{E/k_BT}-1} \tag{77}$$

where in terms of the photon frequency ν we know $E = p = h\nu$. Each photon has an energy E so the total energy received over an area dA with photon direction (momentum) in solid angle $d\Omega$ is $Ef_\gamma d^3\mathbf{p}d^3\mathbf{x} = Ef_\gamma p^2 dp d\Omega_\mathbf{p} d^3\mathbf{x} = E^3 f_\gamma dE d\Omega_\mathbf{p} dAcdt$. Hence the intensity (power per unit area per unit solid angle per unit frequency) observed is (with constants put back in)

$$B(\nu) = \frac{2h\nu^3}{c^2}\frac{1}{e^{\frac{h\nu}{k_BT}}-1}. \tag{78}$$

This is the blackbody spectrum, and observations by Firas fit this spectrum to high accuracy with a temperature of $T = 2.726K$. So we know that photons were very close to equilibrium at some point in the early universe.

For small frequencies (large wavelengths) a series expansion in ν gives the *Rayleigh-Jeans* approximation

$$B(\nu) \approx \frac{2\nu^2 k_B T}{c^2} \qquad (h\nu \ll k_BT). \tag{79}$$

In the limit of high frequencies (the *Wien tail*) there is instead the Wien approximation

$$B(\nu) \approx \frac{2h\nu^3}{c^2}e^{-\frac{h\nu}{k_BT}} \qquad (h\nu \gg k_BT). \tag{80}$$

Sometimes people use $B(\lambda)$ rather than $B(\nu)$, remember to account for the Jacobian if you are switching between the two.

E. Number densities, energy densities and pressure

Given the distribution function, we can then calculate the number density n, energy density ρ and pressure P:

$$n = \int f(\mathbf{p})d^3\mathbf{p} = \frac{g}{2\pi^2}\int_0^\infty \frac{p^2dp}{e^{(E-\mu)/T}\pm 1} = \frac{g}{2\pi^2}\int_m^\infty \frac{(E^2-m^2)^{1/2}}{e^{(E-\mu)/T}\pm 1}EdE\,, \tag{81}$$

$$\rho = \int E(p)f(\mathbf{p})d^3\mathbf{p} = \frac{g}{2\pi^2}\int_0^\infty \frac{E}{e^{(E-\mu)/T}\pm 1}p^2dp = \frac{g}{2\pi^2}\int_m^\infty \frac{(E^2-m^2)^{1/2}}{e^{(E-\mu)/T}\pm 1}E^2dE\,, \tag{82}$$

$$P = \int \frac{p^2}{3E(p)}f(\mathbf{p})d^3\mathbf{p} = \frac{g}{2\pi^2}\frac{1}{3}\int_m^\infty \frac{(E^2-m^2)^{3/2}}{e^{(E-\mu)/T}\pm 1}dE\,. \tag{83}$$

Don't confuse P for pressure with p for momentum[8]. We now give the number density, pressure, temperature, etc, for some useful limits:

---

[8] Why $p^2/3E$ factor? Pressure is the force per unit area, so momentum change per unit time per unit area; momentum in the x direction is p_x per particle, hence a change in momentum of $\Delta p_x = 2|p_x|$ if it hits the perpendicular area dA. The volume swept out in time dt is $d^3\mathbf{x} = |v_x|dtdA = |p_x|dtdA/E$, so the number that hit is $dN = fd^3\mathbf{p}d^3\mathbf{x} = (|p_x|dtdA/E)fd^3\mathbf{p}$. Hence the contribution to the pressure (force per unit area), is $dP = \Delta p_x/(dtdA) = (2p_x^2/E)fd^3\mathbf{p}$ for particles moving in the right direction ($p_x > 0$). For given p_x^2 half are moving in the wrong direction, so the total pressure is $P = \int (p_x^2/E)fd^3\mathbf{p}$. For an isotropic distribution $\int d^3\mathbf{p}p_x^2 = \int d^3\mathbf{p}p_y^2 = \int d^3\mathbf{p}p_z^2$, hence $P = \int \frac{p^2}{3E}fd^3\mathbf{p}$. If the distribution is *not* isotropic, the pressure is defined to be the angle-averaged quantity given by the first part of Eq. (83).

- **Relativistic species**: $m \ll T$, $\mu \ll T$:

$$n = T^3 \frac{g}{2\pi^2} \int_0^\infty \frac{x^2 dx}{e^x \pm 1} \propto T^3 \,, \tag{84}$$

where $x = E/T$. Using the Riemann zeta function

$$\zeta(n) = \frac{1}{\Gamma(n)} \int_0^\infty du \frac{u^{n-1}}{e^u - 1}$$

for bosons we have

$$n_B = T^3 \frac{g\zeta(3)}{\pi^2} \,, \tag{85}$$

where $\zeta(3) \approx 1.202056903$, and for fermions, we can use a cunning trick of writing:

$$\frac{1}{e^x + 1} = \frac{1}{e^x - 1} - \frac{2}{e^{2x} - 1} \,, \tag{86}$$

and then:

$$n_F = T^3 \frac{g\zeta(3)}{\pi^2} (1 - \frac{1}{4}) = \frac{3}{4} n_B \,. \tag{87}$$

For the energy density, using $\zeta(4) = \pi^4/90$ and $\Gamma(n) = (n-1)!$ we get:

$$\rho_B = T^4 \frac{g}{2\pi^2} \int_0^\infty \frac{x^3 dx}{e^x - 1} = \frac{g}{30} \pi^2 T^4 \,, \tag{88}$$

$$\rho_F = T^4 \frac{g}{30} \pi^2 (1 - \frac{1}{8}) = \frac{7}{8} \rho_B \,, \tag{89}$$

and pressure

$$P = \frac{\rho}{3} \,. \tag{90}$$

There are several points here worthy of note: Firstly, we have derived now from statistical mechanics the equation of state for radiation, and indeed we find that $w = P/\rho = 1/3$. Secondly, in eq. (88) we derived the Stefan-Boltzmann relation for photons ($g = 2$), $\rho_\gamma = a_R T^4 = 4\sigma_{SB} T^4/c$, where the Stefan-Boltzmann constant is given by

$$\sigma_{SB} \equiv \frac{k_B^4 \pi^2}{60 \hbar^3 c^2} = 5.6704 \times 10^{-8} \mathrm{Js^{-1} m^{-2} K^{-4}}. \tag{91}$$

We also showed in the last chapter that the energy density in radiation scales like $\rho_\gamma \propto a^{-4}$. Combined with the above, it follows immediately that the temperature of radiation (and of any relativistic species) scales like

$$T_\gamma \propto 1/a. \tag{92}$$

The important consequence is that the universe was much *hotter* when it was smaller. Here we use the temperature of the radiation as the "temperature of the universe", both because it is well defined as the radiation has a thermal spectrum (and thus a unique, well-defined temperature) and because at early times the other particle species interact with the radiation and so share its temperature. We will later discuss what happens when this is no longer true.

- **Relativistic species, small chemical potential**: $m \ll T$, $|\mu| \ll T$ Consider the number density of Fermions, and doing a series expansion in μ:

$$\begin{aligned} n &= \frac{g}{2\pi^2}\int_0^\infty \frac{p^2 dp}{e^{(p-\mu)/T}+1} \\ &\approx n(\mu=0) + \frac{\mu}{T}\frac{g}{2\pi^2}\int_0^\infty \frac{e^{p/T}p^2 dp}{[e^{p/T}+1]^2} + \dots \\ &= n(\mu=0) + \frac{g\mu T^2}{12} + \mathcal{O}(\mu^2/T^2). \end{aligned} \tag{93}$$

  In particular for a particle X and anti-particle $\bar{X}$ with $\mu_X = -\mu_{\bar{X}}$ (as expected if e.g. $X + \bar{X} \leftrightarrow \gamma + \gamma$), then

$$n_X - n_{\bar{X}} = \frac{g\mu_X T^2}{6} + \mathcal{O}(\mu_X^3/T^3). \tag{94}$$

  In particular, if $n_X \approx n_{\bar{X}}$ (e.g. because the universe is charge neutral), then $\mu_X/T \approx 0$. The chemical potential is closely related to conserved charges and particle-antiparticle asymmetries.

- **Non-relativistic (massive) species**, $m \gg T$:

$$\begin{aligned} n &\approx \frac{g}{2\pi^2}e^{\mu/T}\int_0^\infty e^{-E/T}p^2 dp \approx \frac{g}{2\pi^2}e^{\mu/T}\int_0^\infty e^{-(m+p^2/2m)/T}p^2 dp \\ &= g\left(\frac{mT}{2\pi}\right)^{3/2} e^{-(m-\mu)/T}\,, \end{aligned} \tag{95}$$

$$\rho = mn\,, \tag{96}$$

$$P \approx \frac{g}{2\pi^2}e^{\mu/T}\int_0^\infty e^{-(m+p^2/2m)/T}\frac{p^2}{3m}p^2 dp \simeq nT \ll \rho \quad (P \simeq 0)\,. \tag{97}$$

  This recovers the ideal gas law, $P = nT$. Again we can look at the relative numbers of particles and anti-particles in the case that they are massive, and now get

$$n_X - n_{\bar{X}} = 2g\left(\frac{mT}{2\pi}\right)^{3/2} e^{-m/T}\sinh(\mu_X/T). \tag{98}$$

  Again this is zero if $\mu_X = 0$, and for small chemical potentials the numbers are all very small because of the exponential $e^{-m/T}$ suppression.

Therefore, in general for relativistic species the number densities go as T^3 and the energy density behaves as T^4, while for massive species they are suppressed by the Boltzmann factor $\exp(-m/T)$. This exponential suppression of the number density means that non-relativistic particles soon drop below the limit where they interact sufficiently often to stay in equilibrium.

1. Several relativistic species: number of degrees of freedom

If we have a collection of relativistic species, each of them in equilibrium at different temperatures T_i, we can write the total energy density ρ_R, summing over all the contributions and neglecting the chemical potentials ($\mu_i \ll T$), as:

$$\rho_R = \sum_i \rho_i = \frac{T_\gamma^4}{30}\pi^2 g_* \tag{99}$$

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

$$g_* = \sum_{bosons} g_i \left(\frac{T_i}{T_\gamma}\right)^4 + \frac{7}{8}\sum_{fermions} g_i \left(\frac{T_i}{T_\gamma}\right)^4 . \tag{100}$$

As the temperature decreases, the effective number of degrees of freedom in radiation will decrease, as massive particles start behaving non-relativistically when their mass become larger then T:

$T \ll 1$ MeV: the only relativistic particles would be the 3 neutrino species (fermions with 2 degrees of freedom each) and the photon (boson, 2 polarisation states). Neutrinos at this temperature are decoupled from the thermal bath and they are slightly colder than the photons, with a temperature $T_\nu = (4/11)^{1/3} T_\gamma$ [we'll derive this later on]. Therefore:

$$g_* = 2 + \frac{7}{8} \times 3 \times 2 \times \left(\frac{4}{11}\right)^{4/3} \simeq 3.36\,. \tag{101}$$

1 MeV $< T <$ 100 MeV: Electrons and positrons have a mass of about 0.5 MeV, and so they are now also relativistic. As the difference between neutrino and photon temperature is due to the electron-positron annihilation (see discussion in the section IX H), we have $T_\nu = T_\gamma$ and so

$$g_* = 2 + \frac{7}{8}\,(3 \times 2 + 2 \times 2) = 10.75 \tag{102}$$

$T <$ 300 GeV: this is above the electroweak unification scale, and for the particles in the Standard Model we have $g_* \simeq 106.75$

$\vdots$

F. Entropy

The universe has far more photons than baryons, so the entropy of a uniform universe is dominated by that of the relativistic particles. The fundamental relation of thermodynamics for a system in equilibrium with negligible chemical potential (or no change in particle number) is

$$dE = TdS - PdV. \tag{103}$$

The change in energy is the work done changing the volume plus the temperature times the change in entropy (the temperature is effectively the energy per internal state, each with energy $\sim k_B T$). In a cosmological volume V we have $E = \rho V$ so

$$Vd\rho + \rho dV = TdS - PdV. \tag{104}$$

We know about $d\rho/dt$ from the energy conservation equation; using $V \propto a^3$ this gives

$$\frac{d\rho}{dt} = -3H(\rho + P) = -\frac{1}{V}\frac{dV}{dt}(\rho + P), \tag{105}$$

so substituting in Eq. (104) we have[9]

$$-\frac{dV}{dt}(\rho + P) + \rho\frac{dV}{dt} = T\frac{dS}{dt} - P\frac{dV}{dt} \tag{106}$$

$$\implies \frac{dS}{dt} = 0. \tag{107}$$

So the total entropy is in a comoving volume is conserved, which is what we might expect for a closed system (there is nowhere for heat to flow from or to). It is also useful to consider the entropy density $s = S/V$, where substituting for $dS = d(sV)$ in Eq. (104)

$$T(sdV + Vds) = Vd\rho + \rho dV + PdV. \tag{108}$$

---

[9] This argument as presented is circular since we derived the energy conservation equation from $dS = 0$. However the energy conservation equation can also be derived directly from conservation of stress-energy in general relativity.

$$\implies d\rho - Tds = (Ts - \rho - P)\frac{dV}{V}. \tag{109}$$

For a system at equilibrium the entropy density, energy density and pressure are intensive quantities that can be written as functions only of the temperature, $\rho = \rho(T)$, $s = s(T)$, $P = P(T)$, so that $d\rho - Tds \propto dT$. The coefficients of the dT and dV terms must separately be zero because one is intensive (independent of volume) and the other is extensive (depends on the size of the system). For example you could consider a volume change at constant temperature $dT = 0$, which implies then the dV term must be zero. Hence for the dV coefficient to be zero we get an expression for the entropy density

$$s = \frac{1}{T}(\rho + P). \tag{110}$$

The dT coefficient is then zero by the $\dot{\rho}$ energy conservation equation.

We therefore expect $S \propto a^3 s$ to be conserved between different times when a homogeneous universe is in thermodynamic equilibrium. This also applies to separate decoupled (uninteracting) components if they are each separately in a thermal distribution with their own temperature.

For a relativistic species A in thermal equilibrium at a temperature T_A, we have seen that:

$$\rho_A = g_A^{\rm eff}\frac{\pi^2}{30}T_A^4 = 3P_A\,, \tag{111}$$

where $g_A^{\rm eff} = g_A$ for bosons, and $g_A^{\rm eff} = 7g_A/8$ for fermions. The total entropy density s, summing over all possible contributions is given by:

$$\begin{aligned} s &= \sum_i s_i = \frac{\pi^2}{30}(1+\frac{1}{3})T_\gamma^3\left[\sum_{\rm bosons} g_i\left(\frac{T_i}{T_\gamma}\right)^3 + \frac{7}{8}\sum_{\rm fermions} g_i\left(\frac{T_i}{T_\gamma}\right)^3\right] \\ &= \frac{2\pi^2}{45}g_{*S}T_\gamma^3\,, \end{aligned} \tag{112}$$

where T_γ is the temperature of the photons, and we have defined the effective number of degrees of freedom in entropy:

$$g_{*S} = \sum_{\rm bosons} g_i\left(\frac{T_i}{T_\gamma}\right)^3 + \frac{7}{8}\sum_{\rm fermions} g_i\left(\frac{T_i}{T_\gamma}\right)^3\,. \tag{113}$$

Notice that g_{*S} is equal to g_* only when *all* the relativistic species are in equilibrium at the *same* temperature.

We can also consider independent thermal components that are at different temperatures separately. Using the entropy conservation law for a thermal system i at temperature T_i we have then:

$$d(g_{*S,i}a^3T_i^3) = 0 \;\;\Rightarrow\;\; T_i \propto g_{*S,i}^{-1/3}a^{-1}\,. \tag{114}$$

If $g_{*S,i}$ is a constant, then the T_i of the system decreases as the inverse of the scale factor, $T_i \propto a^{-1}$, and $s_i \propto T_i^3 \propto a^{-3}$. If there is one or more independent thermal systems, as we shall see is the case after decoupling, then this law will apply separately to each one, where the temperatures of each system can in general be different.

G. Decoupling

At any time, the Universe will contain a blackbody distribution of photons with some temperature T_γ. If a species interacts (directly or indirectly) with the photons with a rate $\Gamma_{A\gamma}$ high enough ($\Gamma_{A\gamma} \gg H$), then these particles will have the same temperature as the photons: $T_A = T_\gamma$. Any set of particles species which are interacting among themselves at a high enough rate will share the same temperature. But in general this may not be the case and the Universe could be populated by different species each with its own temperature (uninteracting species behave like independent thermal

systems). As a rule of thumb, a species A will maintain an equilibrium distribution when there are many photon interactions in the time it takes the universe to expand significantly, i.e. $\Gamma_A \gg H$, and it will decouple from the thermal bath when the interaction rate drops below the rate of expansion, $\Gamma_A \ll H$.

The rate of interaction can be expressed as:

$$\Gamma_A \equiv n_T \langle \sigma_X v \rangle \,, \tag{115}$$

where n_T is the number density of the target particles, v is the relative velocity, and σ_X is the interaction cross section. $\langle \sigma_X v \rangle$ denotes an average value of this combination (σ_X usually depends on the energy). We should note that the interaction rates between particles A and B are not symmetric, e.g. if $n_A \gg n_B$ then also $\Gamma_B \gg \Gamma_A$. It is then possible that particle B is still in thermal equilibrium with A, while A has already decoupled from B.

As long as $\Gamma_A \gg H$ and the interactions can maintain equilibrium, the distribution function f_A maintain the form of the equilibrium distribution. But once the species A is completely decoupled ($\Gamma_A \ll H$) the particles will be just travelling freely. The distribution function is then frozen in to the form it had at decoupling, though particle momenta will gradually redshift as the universe expands.

Notice that if a massive particle decouples when it is relativistic $T_D \gg m$, then the distribution function is "frozen" in the form of the distribution function $f_{(\mathrm{eq})}$ of massless particles. These particles will become non-relativistic when the temperature of the thermal bath drops below their mass, such that their energy will now be $E \simeq m$. The distribution function and number density of the particles will still be given by the frozen-in form corresponding to relativistic particles, but the energy density will be that of non-relativistic particles $\rho \simeq nm$. This is exactly what happens for massive neutrinos.

H. The thermal history from neutrino decoupling onwards

1. *Neutrino decoupling*

At temperature below $T \simeq 10^{12}\mathrm{K} \simeq O(100)$ MeV, the energy density of the universe is essentially given by that of the relativistic particles $e^{\pm}$, ν, $\bar{\nu}$ and photons. Since they are in equilibrium with the same temperature, the effective number of degrees of freedom is $g_* = 10.75$, and the rate of expansion in this radiation-dominated epoch is given by:

$$H(T) = \frac{\rho_R^{1/2}}{\sqrt{3} M_P} \,. \tag{116}$$

Neutrinos are kept in equilibrium via weak interaction processes (for example $\bar{\nu}\nu \leftrightarrow e^+ e^-$ via Z, elastic scattering of ν and e^- via Z exchange, or $e^- \bar{\nu} \leftrightarrow e^- \bar{\nu}$ via W^-, etc.), with a cross section given by:

$$\sigma_F \simeq G_F^2 E^2 \simeq G_F^2 T^2 \,, \tag{117}$$

where G_F is the Fermi constant ($G_F = \pi \alpha_W / (\sqrt{2} m_W^2) = 1.1664 \times 10^{-5}\,\mathrm{GeV}^{-2}$). The interaction rate per (massless) neutrino is:

$$\Gamma_F = n \langle \sigma_F v \rangle \simeq 1.3 G_F^2 T^5 \,, \tag{118}$$

and

$$\frac{\Gamma_F}{H(T)} \simeq 0.24 T^3 G_F^2 m_P \simeq \left(\frac{T}{1\ \mathrm{MeV}} \right)^3 \,. \tag{119}$$

Therefore neutrinos decouple from the rest of the matter at a temperature around $T_D \simeq 1$ MeV. Below 1 MeV, the neutrino temperature scales as a^{-1}.

2. Electron-positron annihilation

Shortly after neutrino decoupling, the temperature drops below the mass of the electron ($T < 0.5$ MeV), and the electron-positron pairs all annihilate into photons[10]. Due to the strong electromagnetic forces the photon-electron-positron gas remains in thermal equilibrium, but the neutrinos are decoupled. We can therefore treat the photon-electron-positron gas as a separate thermal system from the neutrinos. Let's call them thermal system one and thermal system two, where being decoupled means there are no interactions so they behave independently. For the photon-electron-positron gas before and after $e^- e^+$ annihilation we have respectively:

$$g_{*S,1}(T_D > T \gg m_e) = 2 + \frac{7}{8}4 = \frac{11}{2}\,, \qquad g_{*S,1}(T \ll m_e) = 2\,. \tag{120}$$

These results are of course only approximate near $T \sim m_e$, but due to the $e^{-m/T}$ factor in the equilibrium abundance for massive particles, the number of electrons and positrons is rapidly exponentially suppressed for $T \ll m_e$.

The conservation of the entropy $S_1 = g_{*S,1}(aT_\gamma)^3$ for the particles which are in equilibrium with radiation shows that $g_{*S,1}(T_\gamma a)^3$ remains constant during expansion. Because $g_{*S,1}$ decreases after $T < m_e$, the value of $(aT_\gamma)^3$ will be larger after $e^- e^+$ annihilation than its value before:

$$\frac{(aT_\gamma)^3_{\rm after}}{(aT_\gamma)^3_{\rm before}} = \frac{(g_{*S,1})_{\rm before}}{(g_{*S,1})_{\rm after}} = \frac{11}{4}\,. \tag{121}$$

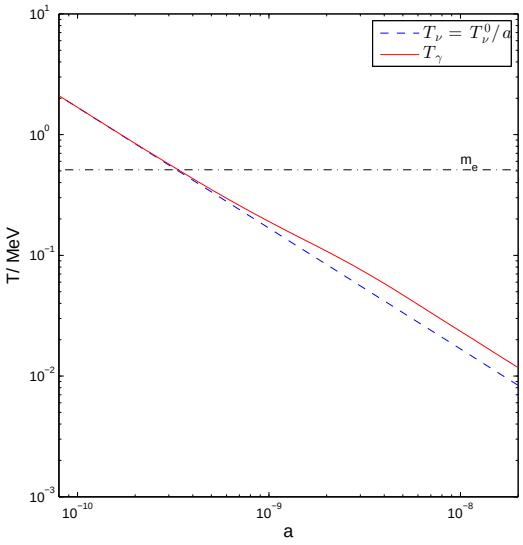

FIG. 8: Thermal history through electron-positron annihilation. Neutrinos are decoupled and their temperature redshifts $\propto 1/a$. When $T \lesssim m_e$ the electrons and positrons (which are in equilibrium with the photons) annihilate, injecting energy and reducing the cooling due to redshifting. Note that the main effect is after $T \sim m_e$, because the contributions to the energy density are skewed towards $E > T$ because of the large phase space and higher weighting for larger momenta (integrand is $\propto f(E)p^2 E dp$).
Ex: See if you can write a Matlab or python script to reproduce this plot by numerically integrating the Fermi-Dirac distribution.

---

[10] Not *all* the electrons annihilate since there is an excess of electrons to have charge neutrality with the protons, $n_e = n_p$. However we will show shortly that this excess is tiny compared to the number of photons, and hence the number of electrons and positrons produced in equilibrium above 0.5 MeV.

Neutrinos do not participate in this process and their entropy is separately conserved, with $(aT_\nu)_{\rm before} = (aT_\nu)_{\rm after}$ since the neutrino degrees of freedom do not change. But before $e^- e^+$ annihilation began, photons and neutrinos had the same temperature, $(aT_\gamma)_{\rm before} = (aT_\nu)_{\rm before}$ since they were in equilibrium prior to neutrino decoupling. Therefore:

$$(aT_\gamma)_{\rm after} = \left(\frac{11}{4}\right)^{1/3} (aT_\nu)_{\rm after}\,. \tag{122}$$

The temperature of the photons is larger than that of the neutrinos today by a factor $(11/4)^{1/3} \sim 1.4$ (so $T_{\nu 0} \sim 1.92K$ for $T_{\gamma 0} = 2.726K$). And because $T_\gamma \neq T_\nu$, today $g_* \neq g_{*S}$, with $g_* \simeq 3.36$ and $g_{*S} \simeq 3.91$.

3. Matter-radiation equality

The total matter density and radiation today is (for CMB temperature today $T_{\gamma 0} = 2.726K$, which is observed to be very close to blackbody)

$$\rho_{m0} = \frac{3H_0^2 \Omega_m}{8\pi G} = 1.88 \times 10^{-32} \Omega_m h^2 \,\mathrm{kg\ cm^{-3}} \tag{123}$$

$$\rho_{r0} = g_* \frac{\pi^2}{30} \frac{(k_B T_{\gamma 0})^4}{c^5 \hbar^3} = 7.8 \times 10^{-37}\,\mathrm{kg\ cm^{-3}} \tag{124}$$

Here we used the definition that the Hubble parameter today is $H_0 = h 100 \mathrm{km\,s^{-1} Mpc^{-1}}$. Using the fact that $\rho_r/\rho_m \propto a_0/a = 1 + z$, it follows that the redshift of equal matter and radiation energy densities (*matter-radiation equality*) is given by:

$$1 + z_{(\rm eq)} \;=\; 2.4 \times 10^4 \Omega_m h^2. \tag{125}$$

Evidence suggests $\Omega_m h^2 \sim 0.133$, so $z_{(\rm eq)} \sim 3200$ ($T_{eq} \sim 0.75$eV). Prior to this redshift the universe was *radiation dominated*, afterwards it was *matter dominated* until dark energy became important at low redshift ($z \sim 1$).

4. Photon decoupling and recombination

We start by estimating the number densities of photons and baryons, since this ratio is critical to how recombination (and nucleosynthesis) proceed. Only a fraction of the matter is baryons, with $\Omega_b h^2 \sim 0.022$ observationally. Most of the baryons are in the form of hydrogen, so the number density of baryons is roughly $3H_0^2 \Omega_b / 8\pi G m_H \sim 2 \times 10^{-7} \mathrm{cm^{-3}}$. The number density of photons is given by the observed blackbody temperature as $2\frac{(k_B T_0)^3 \zeta(3)}{\pi^2 c^3 \hbar^3} \approx 200 \mathrm{cm^{-3}}$. So there are vastly more photons than baryons, with $\equiv n_b/n_\gamma \sim 10^{-9}$ (Planck parameters give $\approx 6.03 \times 10^{-9}$).

Since there are $\sim 10^9$ more photons than baryons (which are mostly protons), decoupling of the photons from the baryons happened first when $\Gamma_\gamma \simeq H$. The protons were very tightly coupled to the free electrons by electromagnetic interactions, and the relevant rate is given by electron scattering:

$$\Gamma_\gamma = n_e \sigma_T\,, \tag{126}$$

where n_e is the number density of free electrons, and $\sigma_T = 8\pi\alpha_{em}^2/(3m_e^2) = 6.65 \times 10^{-25}\,\mathrm{cm^2}$ is the Thomson cross section (scattering of classical electromagnetic radiation by a free e^-; at the times of interest $m_e \gg T$ so scattering was elastic). This happened when the number density of electrons n_e fell rapidly due to recombination[11] of free electrons and protons to form a neutral hydrogen.

---

[11] This odd terminology is conventional; they were never previously combined.

Note that scattering from protons can be neglected because of the $1/m^2$ factor in the cross-section. Also electrons bound into hydrogen atoms have their charge effectively shielded, and hence do not contribute to Thomson scattering; a hydrogen gas is transparent.

The equilibrium abundance of free electrons is determined by the Saha equation. Let n_H, n_p and n_e denote the number density of hydrogen, free protons and free electrons respectively. We are going to assume that all the baryons in the Universe are in the form of protons (neglecting the smaller but significant helium fraction). Because of the charge neutrality of the Universe $n_p = n_e$, and baryon conservation gives $n_B = n_p + n_H = n_e + n_H$. In thermal equilibrium, with $T < m_i$, for $i = \mathrm{e}^-, \mathrm{p}, H$, we have

$$n_i = g_i \left(\frac{m_i T}{2\pi}\right)^{3/2} \exp\left(\frac{\mu_i - m_i}{T}\right) . \tag{127}$$

Recall that $\mu_\gamma = 0$, and due to the reaction $\mathrm{p} + \mathrm{e}^- \leftrightarrow H + \gamma$, we have $\mu_p + \mu_e = \mu_H$. Energy conservation also gives $m_H = m_p + m_e - B$, where B is the binding energy of hydrogen $B = 13.6\, eV$. Then, in order to remove the dependency on the chemical potentials we can take the ratio

$$\frac{n_H}{n_e n_p} = \frac{n_H}{n_e^2} = \frac{g_H}{g_p g_e}\left(\frac{m_e T}{2\pi}\right)^{-3/2} \exp\left(\frac{\mu_H - \mu_p - \mu_e - m_H + m_p + m_e}{T}\right) \tag{128}$$

$$= \left(\frac{m_e T}{2\pi}\right)^{-3/2} \exp\left(\frac{B}{T}\right) , \tag{129}$$

with $g_p = g_e = 2$ and $g_H = 4$ (because the spins of the electron and proton in a hydrogen atom can be aligned or anti-aligned, giving one singlet state and one triplet state, so $g_H = 1 + 3 = 4$). Here we neglected the small mass difference between m_p and m_H in the prefactor (but not in the exponential). The ionization fraction is defined as

$$X_e \equiv \frac{n_p}{n_B} = \frac{n_e}{n_e + n_H} . \tag{130}$$

We find X_e indirectly by calculating $(1 - X_e)/X_e^2 = n_H n_B / n_e^2$, as this depends only on $n_B = \eta n_\gamma$,

$$\frac{1 - X_e^{(\mathrm{eq})}}{(X_e^{(\mathrm{eq})})^2} = \frac{n_H}{n_e^2}\frac{n_B}{n_\gamma} n_\gamma = \frac{2\zeta(3)}{\pi^2}\frac{n_B}{n_\gamma}\left(\frac{2\pi T}{m_e}\right)^{3/2} \exp\left(\frac{B}{T}\right) . \tag{131}$$

This is the Saha equation for the equilibrium ionization fraction, which determines X_e as a function of temperature.

In fact it turns out to be hard to keep the distributions in equilibrium, so the Saha equation is only really a good approximation near the start of recombination. Recall that there are far more photons than baryons, so recombination happens only when there are few enough photons in the high-energy tail of the blackbody distribution. The main problem then is that if you do $e + p \to H + \gamma$ directly, the photon released can immediately ionize a neutral hydrogen atom, giving no net recombination. If the electron is captured to an excited state and cascades down to 2p, the final Lyman alpha transition to the 1s ground state gives a resonance Lyman-alpha photon, which can immediately excite the 1s state of another atom. Thermalizing these energetic resonance photons is difficult (there is only some small thermal broadening of the line); there is a 'bottleneck' at the 2p state, and the process goes out of equilibrium. Recombination actually proceeds more slowly, mostly by capture to an excited state followed by a two-photon transition to the ground state (which gives two photons of lower energy which are not problematic), and gradual redshifting of distribution of resonance photons so that they can eventually escape leaving a net recombination. More detailed analyses show that recombination happens ($x_e \sim 1/2$) at around $z_* \sim 1090$, significantly later than predicted by the Saha result ($z_* \sim 1300$). For more details see e.g. *Peebles: Principles of Physical Cosmology.*

It should be stressed are several distinct events take place around recombination. First, most of the protons and electrons combine to form hydrogen. Then, the process of recombination stops, leaving a small fraction of free electrons and protons, when the interaction rate for $p + e \leftrightarrow H + \gamma$ drops below H. Around the same time the photon mean-free path, given by Γ_γ^{-1}, also becomes larger than H^{-1}, and photons decouple from matter (so we see CMB photons coming directly from the last-scattering

surface, where the photons decoupled from the matter). However since there are far more photons than baryons, the baryon gas decouples from the photons later ($z \sim 300$) than the photons decouple from the baryons ($z \sim 1100$): at $z > 300$ scattering of the residual free electrons with the large bath of photons kept the gas temperature in equilibrium with the photon temperature. However, eventually the density of photons thinned further and the matter then decoupled from the radiation. The baryon gas therefore has $T_b \sim T_\gamma \propto 1/a$ till redshift $z \sim 300$ and only then, after decoupling from the photons, does it fall adiabatically as $T_b \propto 1/a^2$ (as you'd expect from $1/a$ redshifting of non-relativistic baryon velocities, with $T \sim m_b v^2$).

I. Big-Bang Nucleosynthesis (BBN)

Or how to produce light elements (nuclei) from free protons and neutrons in an expanding Universe. At temperatures of order of 1 MeV or larger, protons and neutrons are free and they are kept in equilibrium by the weak interactions that are faster than the expansion rate. However, as the temperature drops and weak processes are not able to maintain equilibrium between neutrons and protons, other nuclear species are going to be thermodynamically favoured. That is, neutrons and protons start producing light nuclei: deuterium (D), tritium (T), helium (^{3}He, ^{4}He), beryllium (^{7}Be) and lithium (^{7}Li), through the following sequence of two-body reactions:

$$\begin{array}{ll} \mathrm{n} \longrightarrow \mathrm{p} + \mathrm{e}^- + \bar{\nu}_e & \mathrm{p} + \mathrm{n} \longrightarrow \mathrm{D} + \gamma \\ \mathrm{D} + \mathrm{p} \longrightarrow {}^3\mathrm{He} + \gamma & \mathrm{D} + \mathrm{D} \longrightarrow {}^3\mathrm{He} + \mathrm{n} \\ \mathrm{D} + \mathrm{D} \longrightarrow \mathrm{T} + \mathrm{p} & \mathrm{T} + \mathrm{D} \longrightarrow {}^4\mathrm{He} + \mathrm{n} \\ \mathrm{T} + {}^4\mathrm{He} \longrightarrow {}^7\mathrm{Li} + \gamma & {}^3\mathrm{He} + \mathrm{n} \longrightarrow \mathrm{T} + \mathrm{p} \\ {}^3\mathrm{He} + \mathrm{D} \longrightarrow {}^4\mathrm{He} + \mathrm{p} & {}^3\mathrm{He} + {}^4\mathrm{He} \longrightarrow {}^7\mathrm{Be} + \gamma \\ {}^7\mathrm{Li} + \mathrm{p} \longrightarrow {}^4\mathrm{He} + {}^4\mathrm{He} & {}^7\mathrm{Be} + \mathrm{n} \longrightarrow {}^7\mathrm{Li} + \mathrm{p} \end{array}$$

There are no stable isotopes with mass number 5 so nothing heavier than ^{4}He can be produced by proton or neutron capture on hydrogen and helium. The reactions are sometimes written in a condensed notation, e.g.

$$\begin{array}{lll} \mathrm{p(n,\gamma)D} & & \\ \mathrm{D(p,\gamma)^3He} & \mathrm{D(D,n)^3He} & \mathrm{D(D,p)T} \\ \mathrm{T(D,n)^4He} & \mathrm{T(^4He,\gamma)^7Li} & \\ \mathrm{^3He(n,p)T} & \mathrm{^3He(D,p)^4He} & \mathrm{^3He(^4He,\gamma)^7Be} \\ \mathrm{^7Li(p,^4He)^4He} & \mathrm{^7Be(n,p)^7Li} & \end{array}$$

The theory of BBN (Big-bang nucleosynthesis) predicts the *primordial* abundances of these light elements once the reaction rate has effectively dropped to zero. The process is dominated by two particle interactions since by the time helium is produced the density is too low for multi-body interactions on the expansion timescales, so for example almost no carbon is produced by interaction of three ^{4}He, and as a result only very light elements are formed.

Here we only sketch the physics involved in the calculation. The detail calculation of these abundances is done by solving numerically Boltzmann equations (which describe in detail the breakdown of equilibrium), taking into account among other things the precise nuclear interactions rates. Our discussion proceeds in three steps. First we will discuss the initial conditions, specifically the neutron abundance which determines how many heavier nuclei can be created. In a second step we discuss the equilibrium theory, and at the end we describe how nucleosynthesis proceeds in reality.

1. Neutron to proton ratio

As we have seen before, neutrinos decouple at around $T \sim 1$ MeV; until then the interaction rate for weak reactions is higher than the expansion rate, and neutrons and protons are kept in equilibrium

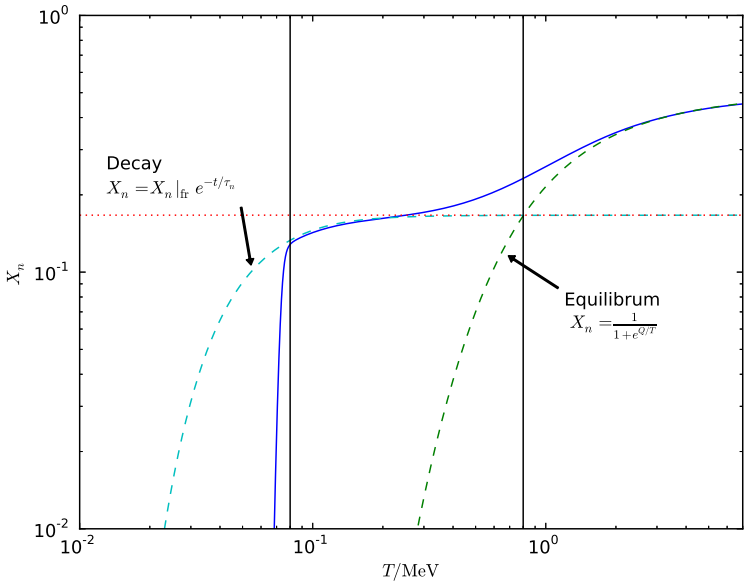

FIG. 9: Evolution of the neutron fraction $X_n = n_n/n_B$ as the universe cools (time goes left in this plot). Initially they are in equilibrium, but freeze out to the equilibrium value at around $T_{\rm fr} \sim 0.8$MeV (horizontal dotted line). They then decay until nucleosynthesis starts (very rapidly) at around $T_{\rm NS} \sim 0.08$MeV. Almost all the neutrons available when nucleosynthesis starts end up in Helium.

through the weak reactions:

$$\begin{aligned} \mathrm{n} &\leftrightarrow \mathrm{p} + \mathrm{e}^- + \bar{\nu_e}\,, \\ \nu_e + \mathrm{n} &\leftrightarrow \mathrm{p} + \mathrm{e}^-\,, \\ \mathrm{e}^+ + \mathrm{n} &\leftrightarrow \mathrm{p} + \bar{\nu_e}\,. \end{aligned}$$

Neutrons, protons and nuclei are all very non-relativistic. In chemical equilibrium, we have: $\mu_n + \mu_\nu = \mu_p + \mu_e$, and the ratio of neutrons to protons in equilibrium is given by:

$$\left(\frac{\mathrm{n}}{\mathrm{p}}\right)_{(\mathrm{eq})} = \left(\frac{n_n}{n_p}\right)_{(\mathrm{eq})} \approx \exp\left(\frac{\mu_n - \mu_p}{T} - \frac{m_n - m_p}{T}\right) = e^{-\frac{Q}{T}} e^{\frac{\mu_e - \mu_\nu}{T}}\,, \tag{132}$$

where $Q = m_n - m_p = 1.293$ MeV.

What do we know about the chemical potentials? Since electrons and positrons rapidly interact with $\mathrm{e}^- + \mathrm{e}^+ \leftrightarrow \gamma + \gamma$, and $\mu_\gamma = 0$, the chemical potentials of electrons and positrons are equal and opposite: $\mu_{\mathrm{e}^-} = -\mu_{\mathrm{e}^+}$. Also since the universe is very accurately charge neutral, $n_{\mathrm{p}} + n_{\mathrm{e}^+} = n_{\mathrm{e}^-}$. Since $n_\gamma \gg n_{\mathrm{p}}$ and at the energies of interest there is a significant fraction of unannihilated electrons and positions, $n_{\mathrm{p}} \ll n_{\mathrm{e}^+}$. In equilibrium $n_{\mathrm{e}^-}/n_{\mathrm{e}^+} \propto e^{(\mu_{\mathrm{e}^-} - \mu_{\mathrm{e}^+})/T} = e^{2\mu_e/T}$, which implies that $|\mu_{\mathrm{e}^-}|/T \ll 1$ at the energies of interest, and hence that term can be neglected. We don't know the chemical potential of the neutrinos, but standard BBN calculations assume zero net neutrino number so that $\mu_\nu = 0$. (It is possible to turn the argument around and use the observations to constrain μ_ν). We thus have

$$\left(\frac{\mathrm{n}}{\mathrm{p}}\right)_{(\mathrm{eq})} \approx e^{-\frac{Q}{T}}\,. \tag{133}$$

At temperatures below $T_{\rm fr} \simeq 0.8$ MeV, weak interactions become slower than H^{-1}, and the ratio of neutrons-to-protons freeze out to the equilibrium value at that temperature:

$$\left.\frac{\mathrm{n}}{\mathrm{p}}\right|_{\rm fr} = e^{-\frac{Q}{T_{\rm fr}}} \simeq \frac{1}{5}\,. \tag{134}$$

However, because of neutron decay, $\mathrm{n} \to \mathrm{p} + \mathrm{e}^- + \bar{\nu}_e$, the ratio of neutrons-to-protons slowly decreases as $n \propto n_{\rm fr}\, e^{-\Delta t/\tau_n}$, where $\tau_n = 880$s is the neutron life-time. By the time nucleosynthesis starts (we will see later that this is only at about $T_{\rm NS} \simeq 80$ keV) the neutrino fraction $X_n = n/(n+p)$ has decreased to

$$X_n|_{\rm NS} \simeq X_n|_{\rm fr}\, e^{-\Delta t/\tau_n} \simeq \frac{1}{6} e^{-\Delta t/\tau_n} \simeq \frac{1}{8}\,. \tag{135}$$

(corresponding to $n/p \approx 1/7$). See Fig. 10 for comparison with the full numerical result used by BBN codes.

Remember that when comparing the rate of the interactions and the expansion rate, the latter is that of a radiation dominated Universe, $H(T) \propto g_*^{1/2} T^2/M_P$, with g_* the total effective number of relativistic degrees of freedom. Since freeze out happens at around the same time as electron-positron annihilation, detailed analyses have to track the evolution of g_* numerically.

2. Weak interaction rates

All matrix elements for weak processes have in common a factor from the matrix element for the neutron decay:

$$|\mathcal{M}|^2 \propto G_F^2 \propto \tau_n^{-1}\,. \tag{136}$$

Therefore, weak interaction rates are usually expressed in terms of the neutron life-time τ_n. For example, for the process $\mathrm{p} + \mathrm{e}^- \leftrightarrow \mathrm{n} + \nu$, we will have:

$$\Gamma_{pe} = \begin{cases} \tau_n^{-1} \left(\frac{T}{m_e}\right)^3 \exp(-\frac{Q}{T})\,, & T \ll Q\,, \\ \sim G_F^2 T^5 & T \gg Q\,. \end{cases} \tag{137}$$

Therefore, the predictions for the abundances of light elements will depend on the precise value of τ_n.

3. Nuclear statistical equilibrium (NSE)

Given a nuclear species, with mass number $A = n + p$ and charge $Z = p$, where n = number of neutrons and p= number of protons in the nucleus, its equilibrium number density is given by:

$$n_A = g_A \left(\frac{m_A T}{2\pi}\right)^{3/2} \exp(\frac{\mu_A - m_A}{T})\,. \tag{138}$$

If nuclear reactions occur at a rate higher than the expansion rate H, then chemical equilibrium is obtained, and the chemical potential for the nucleus is related to that of protons and neutrons through:

$$\mu_A = Z\mu_p + (A-Z)\mu_n\,. \tag{139}$$

Using Eq. (139), we can re-express the number densities in Eq. (138) in terms of the number densities of protons and neutrons:

$$n_A = g_A \frac{A^{3/2}}{2^A} \left(\frac{m_N T}{2\pi}\right)^{3(1-A)/2} n_p^Z n_n^{A-Z} e^{\frac{B_A}{T}}\,, \tag{140}$$

where $B_A = Z m_p + (A-Z) m_n - m_A$ is the binding energy of the nucleus, and in the prefactor in Eq. (140) we have approximated the nucleon mass $m_N \simeq m_p \simeq m_n \simeq m_A/A$. In order to get rid of the effect of the expansion on the number densities, $n_i \propto a^{-3}$, it is useful to define the mass fraction X_A as the number density of the species per total nucleon density $n_N = n_p + n_n + \sum(A n_A)$, normalised

such that $\sum_A X_A = 1$,

$$\begin{aligned} X_A &= \frac{n_A A}{n_N} \\ &= g_A \frac{A^{5/2}}{2^A} \left(\frac{m_N T}{2\pi}\right)^{3(1-A)/2} \frac{n_p^Z n_n^{A-Z}}{n_N} \exp(\frac{B_A}{T}) \\ &= g_A \frac{A^{5/2}}{2^A} \left(\frac{m_N T}{2\pi}\right)^{3(1-A)/2} X_p^Z X_n^{A-Z} n_\gamma^{A-1} \eta^{A-1} \exp(\frac{B_A}{T}) \\ &= g_A C A^{5/2} \left(\frac{T}{m_N}\right)^{3(A-1)/2} X_p^Z X_n^{A-Z} \eta^{A-1} \exp(\frac{B_A}{T}) \,, \end{aligned} \tag{141}$$

where in the last line we have used $n_\gamma = 2\zeta(3)T^3/\pi^2$, $C \equiv \zeta(3)^{A-1}\pi^{(1-A)/2}2^{(3A-5)/2}$, and

$$\eta \equiv \frac{n_N}{n_\gamma} \simeq 2.68 \times 10^{-8} \Omega_B h^2 = 10^{-10}\eta_{10} \tag{142}$$

is the baryon-to-photon ratio.

At this point we are therefore able to build a system of equilibrium equations,

$$X_n/X_p = \exp(-Q/T) \tag{143}$$
$$X_2 = \bar{C}(T/m_N)^{3/2} X_p X_n \eta \exp(B_2/T) \tag{144}$$
$$X_3 = \dots \tag{145}$$
$$\dots \tag{146}$$
$$1 = X_n + X_p + X_2 + X_3 + \dots \tag{147}$$

A species becomes thermodynamically favoured when $X_A \simeq 1$, which depends on $\eta^{A-1}\exp(B_A/T)$. The binding energies for the four first light elements are (2=deuterium, 3=tritium):

$$\begin{aligned} B_2 &= 2.22 \text{ MeV} \,, \quad B_3 = 6.92 \text{ MeV} \,, \\ B_{^3\text{He}} &= 7.72 \text{ MeV} \,, \quad B_{^4\text{He}} = 28.3 \text{ MeV} \,. \end{aligned}$$

Notice that:

- The equilibrium abundances of all bound nuclei are negligible as long as free nucleons are in equilibrium due to the small η factor.
- The binding energies are all of the order of MeV, so their mass fractions are suppressed at least until the temperature is of the order of MeV. Indeed, it has to be lower than MeV to compensate the "suppression" given by the η^{A-1} factors, reflecting the fact that there are many high-energy photons that can immediately destroy nuclei that are formed at around the binding energy.
- The largest binding energy is that of ^{4}He, so almost all the neutrons will end in ^{4}He nuclei. And $X_{^4\text{He}}$ becomes of order one at a temperature $T_{^4\text{He}} \simeq 0.28$ MeV.
- However, in order to produce ^{4}He we need first the deuterium, which has the smallest binding energy. Deuterium is only produced when $X_2 \sim 1$ at a significantly lower temperature $T_2 \sim 0.08$ MeV. Only then can the nuclear reactions proceed beyond deuterium, and nucleosynthesis starts. This is called the *deuterium bottleneck*. Higher element abundances are not in equilibrium (they are much lower than the predictions from NSE) until the deuterium as formed and the reaction chains can proceed.

4. The primordial abundances of light elements

The synthesis of the light elements thus proceeds roughly as follows:

1. At $T \approx 10$ MeV all nuclear processes are in equilibrium, $X_n, X_p \approx 1/2$, $X_{A\geq 2} \ll 1$.

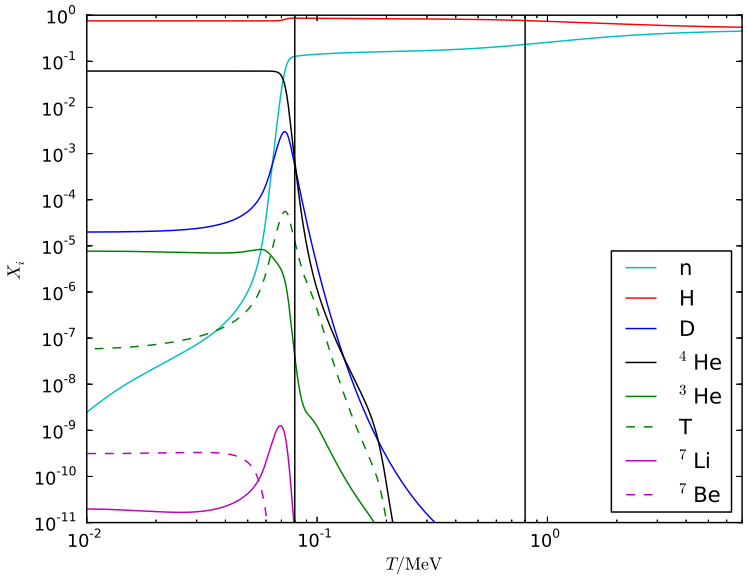

FIG. 10: Evolution of the element abundances through nucleosynthesis (time goes left in this plot), for $\eta \approx 6 \times 10^{-10}$. Nucleosynthesis starts (very rapidly) at around $T_{\rm NS} \sim 0.08$MeV (left vertical line) as deuterium starts to form and is rapidly burnt into other elements, ultimately mostly ^{4}He. The dashed lines correspond to unstable isotopes: Tritium (T) which ultimately decays to ^{3}He, and ^{7}Be which decays to ^{7}Li (so the later-time BBN predictions are the sum of the dashed and solid lines).

2. At $T \approx 1$ MeV the processes keeping neutrons and protons in equilibrium freeze out, $n/p \approx 1/5 \to X_n \approx 1/6, X_p \approx 5/6$. NSE is still a good approximation, and $\eta \ll 1$ ensures that $X_{A\geq 2} \ll 1$.

3. At $T \approx 0.1$ MeV, X_n falls to about 1/8 because of neutron decay. The abundance of helium, $X_{^4\rm He}$, falls below its equilibrium value (which would be about 1) because of the delay in forming deuterium. NSE is now no longer valid. When ^{4}He is finally produced, the temperature (and so the kinetic energy of the nucleons) is too low for particles to break through the Coulomb barrier of the even heavier elements (metals) like carbon and oxygen. This together with the absence of strongly bound stable nuclei with $A = 5 \ldots 8$ means that nearly all neutrons end up in helium.

As we said before, ^{4}He cannot be synthesised until enough deuterium has formed, and effectively delays the start of nucleosynthesis until $T_{NS} \simeq 80$ keV, by which time some of the neutrons present at freeze out will have decayed. Assuming that all the neutrons are bound in ^{4}He, we can estimate its primordial abundance as

$$X_{^4\rm He} = 4\frac{n_4}{n_N} \simeq 4\frac{n_n/2}{n_n + n_p} \simeq 2X_n \simeq 2X_n|_{\rm NS} \simeq 2/8 \simeq 0.25\,. \tag{148}$$

Some deuterium and ^{3}He is left unburnt (of the order of 10^{-5}), and a little ^{7}Li and ^{7}Be is synthesised. All the remaining protons will end up as hydrogen. Residual ^{7}Be gradually decays to ^{7}Li via $^7\text{Be}+e^- \to {}^7\text{Li} + \nu_e$, and residual tritium also decays via $\text{T} \to {}^3\text{He} + e^- + \bar{\nu}_e$ (half life $\sim$ 12 years), so the end products are simple: mostly hydrogen and helium, with small residuals of deuterium, ^{3}He and ^{7}Li.

5. *Predictions*

The primordial abundances of light elements depends on the following parameters:

- *Neutron life-time*: which is used to express the rate of the weak interactions. Nowadays, this is pretty well-known. If τ_n were larger, for example, then we would have a smaller interaction

rate, which would imply a larger freeze-out temperature and larger neutron-to-proton ratio, and therefore a larger ^{4}He abundance.

- *Baryon-to-photon ratio η*: The mass fraction for the nuclear species is proportional to this parameter, so the smaller η is, the smaller X_A (photon-destruction of formed nuclei remains effective until lower temperatures when there are more photons per baryon). A higher η means that the mass fractions for the deuterium, tritium and ^{3}He build up earlier, so that the synthesis of ^{4}He is more efficient, we get more helium and less is left of the lighter nuclei. This number can be constrained both from comparison of BBN predictions with data, and also independently through the detailed shape of the spectrum of anisotropies in the CMB.

- *Effective number of relativistic degrees of freedom*: or equivalently, the total energy density at the time of BBN. With 3 generations of neutrino species, we have $g_* = 10.75$ at $T \sim 1$ MeV. If there is an extra relativistic degree of freedom (like extra neutrinos), they will contribute:

$$g_* = 10.75 + \sum_B g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8}\sum_F g_i \left(\frac{T_i}{T}\right)^4 . \tag{149}$$

  An extra relativistic degree of freedom means more energy density in radiation, and a larger expansion rate. Therefore, freeze-out occurs earlier and there is less time for neutrons to decay. The neutron-to-proton ratio is therefore larger, and so is the final primordial helium abundance. Observations constrain the extra number of neutrino families to be less than one.

6. Abundances of the light elements, BBN and observations

The predicted abundances of light elements in BBN (no extra neutrinos, no chemical potentials) as a function of η (the baryon to photon ratio) are plotted in Fig. 11 (from a numerical code that evolves the full hierarchy of interaction equations).

The problem with using observations to compare with BBN is that the early universe is not the only place where the number of elements are changed: nucleosynthesis in stars forms heavy elements and can burn up primordial light elements. Nonetheless we can often tell when this is a problem, because heavy elements are only produced in significant numbers in stars, so places where we see many heavy elements are clearly contaminated. Instead constraints can be placed by looking for low metallicity systems – those with few heavy elements so that the observed abundances may better reflect the primordial abundances ("metals" in astrophysics are all the elements except hydrogen and helium). If the metallicity dependence of the abundance of interest can be judged, it may also be possible to extrapolate from observations to zero metallicity, even though we see no entirely uncontaminated systems.

- *Deuterium*: The deuterium can be easily burnt in stars to produce ^{3}He, so that its present abundance provides only a lower bound to the primordial one. Limits can be placed by observing quasar absorption lines from gas clouds with little evidence of contamination from stellar nucleosynthesis (e.g. arXiv:0805.0594). The present limit on D/H is of the order of 2.5×10^{-5}.

- *Helium-4*: This is a very strongly bound nuclei, so it is not destroyed by stellar processes, but is produced in stars together with other elements: the observed abundance should be larger than the primordial one. Observations give $Y_P = 0.253 \pm 0.01$ (from arXiv:1112.3713; Y_P is the standard symbol for the primordial mass fraction[12]/ This compares well with the BBN prediction of around 0.248 for Planck baryon density parameters.

We can use one of the abundances to measure the baryon to photon ratio $\eta = n_B/n_\gamma$, or this can be determined from independent data (CMB anisotropies). The other measurements then provide a

---

[12] Note that because the mass of helium is not quite 4 times the mass of hydrogen, for BBN studies people often define the mass fraction as the ratio as $Y_P = 4n_{^4\mathrm{He}}/n_B$. CMB codes use the actual mass fraction in Helium.

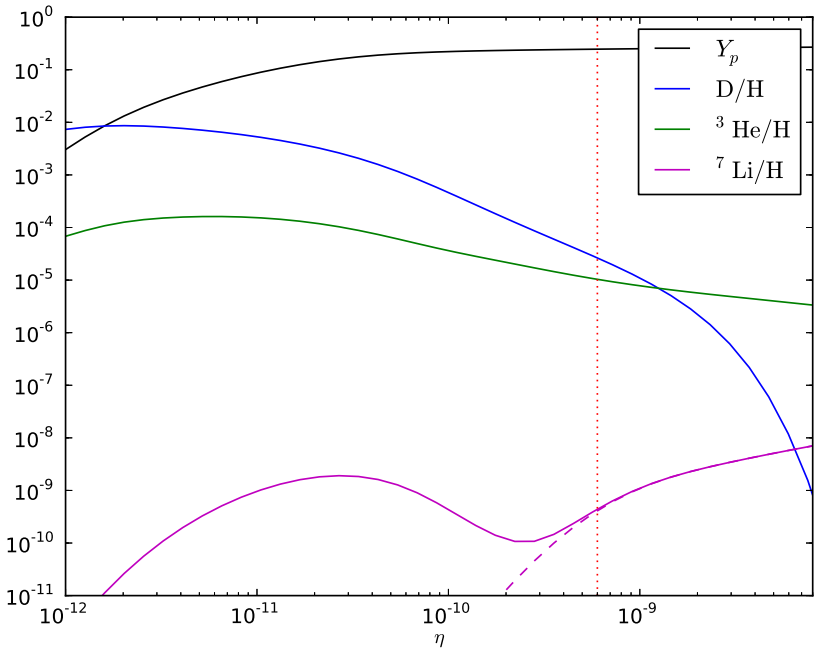

FIG. 11: Elemental abundances as a function of η, from the AlterBBN code The shape of the lithium result is complicated because on the left it first increases as more is directly produced, then decreases again as it starts to be destroyed by more proton collisions; finally for even higher baryon densities the residual is dominated by indirect production by decay of ^{7}Be (shown dashed). The vertical dotted line shows the value of η measured by Planck.

highly non-trivial consistency check on BBN predictions. They also constrain the number of additional light neutrinos or any other contribution to the energy densities at very high redshifts that would have changed the expansion rate during BBN.

The value of η that is observed, giving $\Omega_b \simeq 0.05$, is broadly consistent with independent measurements at a redshift of $z \sim 1100$ from the anisotropies in the CMB. Once Ω_b/η is fixed, we can then test observations of other abundances against the BBN predictions. These are broadly consistent for deuterium and helium-4. For lithium, observations suggest lithium is present at the order of magnitude predicted by BBN, but in detail significantly less is observed than we'd expect; this is sometimes called the 'lithium problem' (see e.g. arXiv:1203.3551). Understanding whether this represents a problem interpreting the data, or whether standard BBN needs to be modified (either resonant nuclear interactions that have not been included in the numerical codes, or new physics), is an open question.

X. PROBLEMS WITH THE HOT BIG BANG MODEL

A. The horizon problem

We know from CMB observations that the universe at recombination was the same temperature in all directions to better than 10^{-4}. Unless the universe has been able to smooth itself out dynamically, this means that the universe must have started in a very special initial state, being essentially uniform over the entire observable universe. Furthermore if we look at the small fluctuations in the temperature we see correlations on degree scales: there are 10^{-4} hot and cold spots of larger than degree size on the sky. Since random processes cannot generate correlations on scales larger than the distance light can travel, these fluctuations must have been in causal contact at some time in the past. This is telling us something very interesting about the evolution of the early universe, as we shall now see.

Let's investigate the distance travelled by a photon from the big bang to recombination and then from recombination to us. A photon travels a proper distance cdt in interval dt. Setting $c = 1$ the comoving distance travelled is $d\chi = dt/a$. The comoving distance light can travel from the big bang to a time t, the *particle horizon*, is therefore

$$\eta_* = \int d\chi = \int_0^t \frac{dt'}{a(t')}. \tag{150}$$

So starting at a given point, light can only reach inside a sphere of comoving radius η_* by the time of recombination t_*. *If* we assume the early universe contained only matter and radiation we can solve for η_*, about 300Mpc. So we only expect to see correlations over patches with comoving radius smaller than about 300Mpc at recombination.

What angular size would we expect to see this area on the sky? Today the comoving distance from recombination is $\sim \eta_0 \sim 14000$Mpc, so we would only expect correlations over an angular radius $\sim 300/14000 \sim 1°$. No causal process could have generated the correlations seen on larger scales!

We must have made an incorrect assumption. The most likely suspect is the assumption that the universe only contained matter and radiation in the early universe: perhaps some different content could change the evolution at early times so that $\eta_* \gg 300$Mpc, as required to generate correlations on large scales? For a constant equation of state, from Eq. 27 we know $H^2 \propto \rho \propto a^{-3(1+w)}$ so

$$\eta \propto \int \frac{da}{a^2 H} \propto \int \frac{da}{a^2 H} \propto \int da a^{(3w-1)/2} \propto \left[a^{(3w+1)/2}\right]_{a_i}^1 \tag{151}$$

so large conformal times can be obtained from the lower limit $a_i \to 0$ if $w < -1/3$. If this is the case very large distances could have been in causal contact. If $w \sim -1$ ($a \sim e^{Ht}$, $H \sim$ const) for some early-universe epoch so that $\ddot{a} > 0$, this is termed *inflation*, the most popular solution to the horizon problem. We use the term *hot big bang* to describe the time after inflation (usually taken to be $t = 0$) at which the normal radiation-dominated universe started.

A useful rule of thumb is that the universe approximately doubles in size in a 'Hubble time' $\sim 1/H$. The Hubble distance $1/H$ sets approximately the distance light can travel in this time. A scale larger than $\gtrsim 1/H$ is therefore expanded faster by the expansion than the extra distance light can travel in an expansion time, and so becomes out of causal contact with itself. Physical scales grow with the expansion, $\lambda_{\rm phys} \propto a \sim e^{Ht}$, so in inflation where $H \sim$ const every scale eventually becomes larger than $1/H$ and goes out of causal contact with itself, when $\lambda_{\rm phys} \gtrsim H^{-1}$. This is what is required to solve the horizon problem: going back in time from the hot big bang, a physical scale $\lambda_{\rm phys}$ shrunk rapidly, and was inside the horizon (in causal contact with itself) at some earlier point in inflation.

Equivalently in terms of comoving length, a comoving scale $\lambda = \lambda_{\rm phys}/a$ goes out causal contact with itself when $\lambda \sim (aH)^{-1}$, the comoving Hubble length. During inflation $H \sim$ const but $a \sim e^{Ht}$ grows rapidly, so $(aH)^{-1}$ is rapidly shrinking. In one Hubble time $\Delta t = H^{-1}$ light travels the comoving distance

$$\frac{\Delta t}{a} = \frac{1}{aH}, \tag{152}$$

so a comoving scale λ is in causal contact with itself when $\lambda \lesssim (aH)^{-1}$ (termed *inside the horizon*), and out of causal contact with itself when $\lambda \gtrsim (aH)^{-1}$ (termed *outside the horizon*).

Any solution to the horizon problem must have some time in the past (prior to recombination) at which all relevant scales have $\lambda < (aH)^{-1}$ so they are in causal contact. An alternative to inflation would be to have some contracting phase before the hot big bang, so that there is a time in the contracting phase when the observable universe was all in causal contact. Inflation solves the horizon problem because a grows rapidly and H is nearly constant, so that early in inflation $(aH)^{-1}$ was very large and then shrank rapidly. In a pre-big bang contracting universe light can travel a large comoving distance during the contracting phase, and the horizon problem is also solved. In the pre-big bang a can be decreasing, but if $|H|$ shrinks rapidly enough it's still possible to have $\lambda < |aH|^{-1}$ early in the contracting phase and also $\lambda > |aH|^{-1}$ as the big bang is approached, making it in some ways remarkably similar to inflation; such models have been called *ekpyrotic* models (see e.g. astro-ph/0404480) and can be motivated as resulting from the collision of 4-D 'branes' colliding in a higher-dimensional space. However the contracting universe has the disadvantage of being much harder to study, since it requires knowing how to go through $a = 0$ without singularities. We shall not discuss it further here, and focus on the more popular model of inflation.

B. The flatness problem

Recall the Friedmann equation

$$H^2 + \frac{K}{a^2} = \frac{8\pi G}{3}\rho, \tag{153}$$

and the critical density ρ_c defined so that

$$H^2 = \frac{8\pi G}{3}\rho_c. \tag{154}$$

In general $\rho = \Omega(t)\rho_c$, with $\Omega(t) \neq 1$ if the universe is not flat, and $\Omega(t)$ is a function of time.

The curvature radius of the FRW universe $R_K^2 \equiv a^2/|K|$ is given by

$$R_K^{-2} = \frac{|K|}{a^2} = \left|\frac{8\pi G\rho}{3} - H^2\right| = H^2|\Omega(t) - 1|, \tag{155}$$

so the ratio of the Hubble radius to the curvature scale is determined by $\Omega(t)$:

$$\frac{H^{-1}}{R_K} = |\Omega(t) - 1|^{1/2}. \tag{156}$$

Today, the energy density of the universe is close to the critical density, $\Omega_0 = \Omega(t_0) \sim 1$ (observations measure this to about 1%). Equivalently, the curvature radius is significantly larger than the current Hubble radius. Is this surprising?

Let's consider how the relative energy density $\Omega(t)$ evolves as a function of time when the background equation of state is constant. We shall assume $\Omega(t) \sim 1$, and see if that is stable, so we can use the flat-space result for $a(t)$, which gives $H \propto 1/t \propto a^{-(3+3w)/2}$ so that

$$\Omega(t) - 1 = \frac{K}{a^2H^2} \propto Ka^{1+3w}. \tag{157}$$

During matter and radiation domination $|\Omega(t)-1|$ grows, and hence had to be much smaller in the past to give a small value today. If for example $\frac{H^{-1}}{R_K} \sim 1$ at some early epoch, it would be enormous today, not small as observed. Why was the universe initially so flat? We have seen that $|\Omega(t) - 1| \propto a^{1+3w}$. So if the universe was at some point in the past dominated by a fluid with $w < -1/3$, then $|\Omega(t) - 1|$ was growing smaller instead of larger: the $\Omega(t) = 1$ is a stable fixed point. If this phase persisted for long enough it would explain the observed flatness: if inflation started in a universe that was slightly non-flat, the exponential expansion expands the curvature scale to be much larger than the current horizon, so the universe we see is very flat.

C. The monopole problem

Phase transitions are often associated with symmetry breaking. In a supersymmetric model, all running coupling constants seem to reach the same magnitude at about 10^{16}GeV, suggesting a unification of the gauge groups $U(1) \times SU(2) \times SU(3) \to G$. If this is indeed the case, and if G is a simple group, then a general theorem states that monopoles are formed when G is broken into subgroups containing $U(1)$. This is what happens in GUT (grand unification) theories when the universe cools below the critical temperature T_c at which the unification happens.

The characteristic mass of the monopoles is the critical temperature of the phase transition, $m_M \approx T_c \sim 10^{16}$GeV. We expect generically to form about one monopole per Hubble volume, so their number density is

$$n_M(T_c) \sim H(T_c)^3 \sim g_*^{3/2}\frac{T_c^6}{M_P^3}. \tag{158}$$

Using the entropy density $s \sim g_* T_c^3$, where from conservation of entropy n_M/s is constant, we can write this as

$$\frac{n_M}{s} \sim \left(\frac{T_c}{M_P}\right)^3 . \tag{159}$$

As the monopoles are "hidden" from each other by their associated gauge field, they freeze out very rapidly and no monopole-antimonopole annihilation occurs. We can write their contribution to the total energy density in terms of the entropy per baryon s/n_b at late times as

$$\rho_M = m_M n_M = \frac{m_M}{m_b}\frac{n_M}{s}\frac{s}{n_b} m_b n_b \sim \frac{m_M}{m_b}\frac{n_M}{s}\frac{n_\gamma}{n_b}\rho_b \tag{160}$$

and therefore $\rho_M/\rho_b \sim 10^{16} \cdot 10^{-6} \cdot 10^9 \sim 10^{19}$ – far too much mass in monopoles compared to baryons to be consistent with observation.

This problem can be avoided by postulating that there is no unification of the gauge groups to a simple group. Or there could be a rapid expansion: in this case all the matter is diluted as $1/a^3$, so a sufficiently large increase in the scale factor can get rid of all the monopoles. All the other matter and radiation is also redshifted away, but the energy in the expansion is released at the end of the "inflationary" phase (a process called reheating), generating the radiation and baryons that we see. Afterwards, the thermal history can proceed as in the standard model. But if $T_c > T_{\rm reheat} \gg m_x$ where x is the normal kind of matter, then the monopoles are no longer a problem.

D. Initial conditions

If we have a hot big bang model, the radiation density $\propto 1/a^4$, and so as $a \to 0$ the density becomes very high: at some point in the past physics at energy scale much higher than we know about in the lab will become important. It might look as though $1/a^4 \to \infty$ —- i.e. there is a *singularity* at the start of the hot big bang — but all it's really saying is that at early epochs physics that we don't know will become important. In particular once photons have energies comparable to the Planck mass ($M_P \sim 10^{18}$ GeV) we should no longer trust General Relativity: quantum gravity effects are likely to become important. Inflation models are an example where new physics enters at a somewhat lower energy scale, and remove the original singularity of the hot big bang model since the hot big bang only started at some finite energy density corresponding to the energy density at which inflation ended.

XI. INFLATION

As we have seen, various problems with the hot big bang model can be resolved by having a period of accelerated expansion with equation of state $w < -1/3$ before the start of the hot big bang. The most popular way to achieve this is with scalar field inflation: we conjecture the existence of one (or more) scalar fields, called *inflatons*, which are evolving slowly as the universe expands. The expansion rate is nearly constant, so it looks very similar to the de Sitter (cosmological constant dominated) model, but the slow evolution allows inflation to eventually end, and *reheating* to start the hot big bang. [just cosmological constant model would expand for ever, which is no good since we do need a hot big bang too!] Unfortunately there are no very good particle physics candidates for the inflaton. Even in supersymmetry theories it is quite hard to find a potential that is flat enough to give $w < -1/3$ for a significant period of time (and quantum corrections also tend to make potentials non-flat even if they are classically).

A. Amount of inflation

For inflation with $w \approx -1$, we have exponential expansion $a \propto e^{Ht}$ where H is nearly constant, $|\dot{H}|/H^2 \ll 1$ (this is often called *slow roll* - the energy density is only slowly evolving - rolling - down to lower values where inflation ends when $w > -1/3$).

The amount by which the universe inflates is measured as the number of *e-foldings* N, so the scale factor before and after inflation are related by

$$\frac{a(t_f)}{a(t_i)} = e^N. \tag{161}$$

The number of e-foldings is just given by the integral of the expansion rate, for nearly-constant H given by

$$N = \int_{t_i}^{t_f} H dt \sim H(t_f - t_i) \tag{162}$$

To solve the horizon, flatness and monopole problems, N has to be large enough.

For the monopoles to dilute sufficiently the density has to be diluted by at least a factor $\mathcal{O}(10^{-19}) \sim e^{-44}$ in order not to dominate. The primordial abundance that needs to be diluted is 1 per hubble volume of size H_i^{-1}, assuming inflation happened shortly after the phase transition at T_c. At the end of inflation the number density after dilution is $H_i^3 e^{-3N} \sim H_f^3 e^{-3N}$, so inflation suppresses the number density by $\mathcal{O}(e^{-3N})$, corresponding to $N \gtrsim 15$. In fact observations indicate that acceptable levels of monopoles or other relics is significantly lower than this, so we need N significantly more than this[13].

For the flatness and horizon problem, the number of e-foldings required depends on when inflation ended. Crudely approximating the universe as radiation dominated so that $\Omega(t) - 1 \propto Ka^2$ we have

$$\frac{\Omega_0 - 1}{\Omega_f - 1} \sim \left(\frac{1}{a_f^2}\right), \tag{163}$$

where a_f is the scale factor at the end of inflation. During inflation with $w \approx -1$ we have $\Omega(t) - 1 \propto K/a^2$, hence

$$\frac{\Omega_f - 1}{\Omega_i - 1} \sim \left(\frac{a_i}{a_f}\right)^2 = e^{-2N}. \tag{164}$$

Hence the ratio of $\Omega - 1$ today to the beginning of inflation is

$$\frac{\Omega_0 - 1}{\Omega_i - 1} \sim \frac{e^{-2N}}{a_f^2}. \tag{165}$$

For general initial conditions with $\Omega_i \sim \mathcal{O}(1)$ we need $e^{-N}/a_f \ll 1$ in order to have $\Omega_0 \sim 1$ today. The scale factor at the end of inflation depends on when inflation ended, often parameterized by the *reheating temperature* - the temperature T_f at which the hot big bang started. Note that since $H^2 \sim \rho/M_P^2 \sim T_f^4/M_P^2$ we have $T_f \sim \sqrt{HM_P}$, so for $H \ll M_P$ the reheating temperature is much higher than the energy scale of inflation.

Recalling that from conservation of entropy $a \propto g_{*S}^{-1/3} T^{-1}$ we have

$$\frac{1}{a_f} \approx \left(\frac{g_{*S}^f}{g_{*S}^0}\right)^{1/3} \left(\frac{T_f}{T_0}\right). \tag{166}$$

The prefactor should be $\mathcal{O}(1)$ or a bit more, and the absolute maximum value of the reheating temperature is $T_f \sim 10^{16}$ GeV (lower than the GUT scale so that inflation has diluted the monopoles), giving

$$\frac{1}{a_f} \sim \frac{10^{16}\text{ GeV}}{k_B 2K} \sim 10^{29} \sim e^{66}. \tag{167}$$

---

[13] For magnetic monopoles, an observational upper limit is around 10^{-30} per baryon, or a fraction 10^{-14} by mass. i.e. $N \gtrsim 25$

On the other hand the reheating temperature could be much lower, down to the electroweak scale at $T_f \sim 1\text{TeV}$, implying $\frac{1}{a_f} \sim 10^{16} \sim e^{36}$. Depending on the reheating temperature, inflation with $N \gtrsim 36$–66 is enough to make the universe today approximately flat even if inflation started with slight curvature.

The horizon problem will also be solved if the entire observable last-scattering surface was in causal contact at the start of inflation. We therefore require the comoving horizon $1/(aH)_i$ at the start of inflation to be larger than $\sim 10\text{Gpc}$. More explicitly the comoving distance travelled by light during inflation, taking H nearly constant is

$$\chi \sim \int_{a_i}^{a_f} \frac{da}{a^2 H} \approx H^{-1}\left(\frac{1}{a_i} - \frac{1}{a_f}\right) \approx \frac{1}{H a_i} \approx \frac{e^N}{a_f H}. \tag{168}$$

The expansion rate during inflation is determined by the Friedmann equation $H^2 \sim \kappa\rho/3$, and hence from energy conservation roughly to the energy density at the start of the hot big bang $H^2 \sim H_f^2 \sim \kappa T_f^4$. We want the current Hubble volume to have been in causal contact, hence $\chi \gtrsim \lambda \sim 10\text{Gpc}$, hence depending on the reheating temperature (and dropping lots of numerical prefactors)

$$e^N > \lambda a_f H_f \sim \frac{\lambda T_0 \sqrt{\kappa T_f^4}}{T_f} \sim \frac{\lambda T_0 T_f}{M_P} \sim e^{32}\text{–}e^{63}. \tag{169}$$

In all cases the problems are resolved as long as inflation gave more expansion than the expansion between the start of the hot big bang and today, typically $N \gtrsim 40$–60 is enough for consistency with data. N could be much larger, but the fluctuations we observe originate as quantum fluctuations from just before they left the horizon at $N \gtrsim 40$–60, and the earlier evolution of the universe is therefore largely irrelevant for observations. Note that smaller scales left the horizon later ($1/aH$ decreases as inflation progresses), and since we can only observe a range of scales spanning a few e-folds in e.g.the CMB, observations of fluctuations only probe a relatively small fraction of the inflationary epoch around $N \sim 40$–60.

Since inflation must dilute any monopoles, it must have energy scale $< 10^{16}$ GeV, which is much less than the Planck mass so quantum gravity effects should not be important. In this sense inflation regularises the start of the hot big bang. However of course it only defers the question of initial conditions to some earlier epoch, and inflation has its own problems in that respect (for example inflation has to start with a nearly smooth and homogeneous energy density over a Hubble radius sized patch, which is much larger than a Planck length - why wasn't it much more random?).

XII. STRUCTURES IN THE UNIVERSE AND COSMOLOGICAL CONSTRAINTS

A detailed analysis of perturbation generation and structure formation is beyond the scope of this course (but will be covered in Early Universe); here we only give a qualitative discussion to understand some of the most important features and terminology.

We know the early universe had small $\mathcal{O}(10^{-5})$ perturbations because we can see them in the CMB. We can see hot and cold spots of many different sizes, so there must have been perturbations on a wide range of scales. Inflation models can in fact predicts a nearly scale-invariant spectrum of perturbations, meaning the amplitude of the perturbations was the same on all different scales at the beginning of the hot big bang.

The perturbations we observe are consistent with *adiabatic* primordial perturbations - again, what you would expect from simple inflation models. This means that there was only one degree of freedom locally, so that perturbations in the dark matter, baryons, neutrinos and photons are all related to each other. This is what you expect from a gravitational perturbation: each fluid responds in a predictable way, where there is an overdensity of baryons there will also be an overdensity of dark matter, photons and neutrinos. As the universe expands an overdensity will exert gravitational attraction on the surrounding fluid, and hence tend to grow (become a deeper potential well) by pulling in more material. However this is compensated by the expansion of the universe: a perturbation of physical size r grows with the scale factor $\propto a(t)$, and this tends to reduce the GM/r gravitational potential as the perturbations becomes physically wider because of the expansions. During matter

domination it turns out that $\delta\rho/\rho$ grows $\propto a(t)$, and hence the gravitational potentials actually remain constant in time. Nonetheless the $\delta\rho/\rho$ can still become much larger, and indeed at late time (and small scales) can become $\gg 1$, at which point a simple perturbation theory analysis breaks down and numerical simulations have to be performed instead.

During radiation domination the picture is a bit more complicated, because the radiation fluid has a large pressure, $P_r = \rho_r/3$. As the fluid starts gravitational collapse the pressure increases, and this acts to oppose the gravitational force, preventing further collapse. In fact what happens is that perturbations in the radiation oscillate: first they start collapsing, pressure increases and overshoots that required to support gravity, so the pressure then pushes the densities out again reducing the density. This makes the gas underdense, so low pressure, and this pulls the gas back in again, and the cycle continues: these are called *acoustic oscillations*, because the perturbations behave like sound waves. These will effect perturbations on all scales that can feel pressure, which means they must be perturbations smaller than the horizon (more precisely, the *sound horizon* size, which depends on the speed of the pressure waves $c_s \lesssim c/\sqrt{3}$). There is therefore a characteristic scale, the sound horizon at recombination, corresponding to the largest perturbation that can undergo gravitational collapse by that time. Smaller scale perturbations will have had time to oscillate, and could be in any phase of the oscillation cycle at recombination. So when we look at the CMB, we see different amounts of perturbations of different (sub-horizon) sizes, depending on whether the perturbations were at an extremum, or near a null, at the time of recombination.

A. The CMB power spectrum

The anisotropies on the CMB are usually quantified by their *power spectrum*. For the CMB we can only observe the anisotropy as a function of direction, so the observation can be thought of as being on the surface of a sphere (corresponding to us looking at the last-scattering surface in all directions). In this case the eigenfunctions of the Laplacian on the sphere are the spherical harmonics Y_{lm} (the analogue of Fourier modes for the sphere), with $\nabla^2 Y_{lm} = -l(l+1)Y_{lm}$, we expand the data in spherical harmonics with

$$T(\hat{\mathbf{n}}) = \sum_{lm} T_{lm} Y_{lm}(\hat{\mathbf{n}}) \tag{170}$$

$$T_{lm} = \int d\Omega_{\hat{\mathbf{n}}} T(\hat{\mathbf{n}}) Y_{lm}(\hat{\mathbf{n}})^*. \tag{171}$$

The l index quantifies the scale, with low l corresponding to long-wavelength modes. The index m satisfies $-l \leq m \leq l$, so there are $2l+1$ different m values for each l, corresponding to different $e^{im\phi}$ modes. The lowest mode $l = 0$ corresponds to the monopole, the uniform component of the observed temperature (the average). The $l = 1$ modes correspond to a dipole pattern, and the m values correspond to the three numbers required to specify the direction and magnitude of the dipole. The sound horizon scale is around $l \sim 200$.

The CMB power spectrum is then defined to quantify the variance in the T_{lm} as a function of l, with

$$C_l \equiv \langle |T_{lm}|^2 \rangle. \tag{172}$$

For a statistically isotropic distribution the power spectrum is only a function of l not m, and $\langle T_{lm} T^*_{l',m'} \rangle = \delta_{ll'}\delta_{mm'} C_l$. If the temperature on the sky is measured, in terms of the observed T_{lm} one can estimate the power spectrum using

$$\hat{C}_l = \frac{1}{2l+1} \sum_m |T_{lm}|^2, \tag{173}$$

where the expectation value give the true power spectrum, $\langle \hat{C}_l \rangle = C_l$. It is important to distinguish between $\hat{C}_l$ and C_l: the latter is the average value of $\hat{C}_l$ if one could observe an infinite ensemble of universes. In our own universe we can only observe $\hat{C}_l$, which differs from C_l by *cosmic variance*.

For low l (large-scales), this significantly restricts the precision with which we can measure C_l, an unavoidable consequence of only having access to observations from one position in one universe.

The C_l spectrum has a peak at $l \sim 200$ corresponding to the largest scale that has time to collapse at recombination. Large scales have not had time to full collapse, and on very large scales $l \lesssim 50$ the spectrum is flat, which is just proportional to the nearly scale-invariant spectrum at the start of the hot big bang. At $l > 200$ there are a series of acoustic oscillation peaks, with the second peak corresponding to overdensities perturbations that are maximally rarefied at recombination (after pressure expansion), and the third peak corresponds to a scale that has had time to do a full oscillation back to maximum density at the recombination time.

The full numerical power spectrum C_l can be calculated using various standard codes such as CMBFAST, CAMB or CLASS. These take in a set of values for the cosmological parameters and produce the spectrum C_l as a function of l (and also polarization and other related spectra). By comparing the spectra to data it's possible to find which sets of cosmological parameters fit the data well, and hence strongly constrain various combination s of cosmological parameters.

B. The matter power spectrum

The matter power spectrum $P(k, z)$ gives the variance of the total density perturbations as a function of wavenumber k (by statistical homogeneity and isotropy it can only be a function of k, not $\mathbf{k}$). The wavenumber $\mathbf{k}$ is obtained by the Fourier transform of the density as a function of moving position at any fixed redshift, with $k = 2\pi/\lambda$ where λ is the comoving wavelength of a perturbation. Note that unlike the CMB power spectrum this is not directly observable — we can only observe our light cone! — but is still useful to consider from a theoretical point of view; relating to actual observations is more complicated.

We already described how matter perturbations grow during matter domination. What happens to them during radiation domination? Here the radiation fluid is undergoing acoustic oscillations: pressure is preventing the perturbations from continuously growing. However the universe is continuously expanding, so the GM/r potential is falling off $\propto 1/a(t)$ with no corresponding growth in M. This means that the gravitational potentials rapidly fall to nearly zero on sub-horizon scales: there are no strong gravitational forces acting on the dark matter. What happens is that the dark matter starts falling into the potential wells, but the potentials rapidly fall to zero and the dark matter then feels no force and only continues to fall in because of the infall velocity already built up, and this infall velocity also gradually redshifts away. The result is that sub-horizon scale dark matter perturbations only grow slowly during radiation domination, and only start to grow quickly again when the universe becomes matter dominated (at which point the radiation pressure is irrelevant, so growth is restored and potentials then remain constant). There is a characteristic scale, the *turnover* in the matter power spectrum, corresponding to the largest scale that entered the horizon before matter domination (the detailed shape of the spectrum requires more work to understand). However there are no acoustic oscillations in the dark matter, so this does not lead to oscillations in the power spectrum directly, just a change in shape at the characteristic scale.

There is also another characteristic scale in the matter due to the baryons. When the baryons were ionized, they were strongly coupled to the photons, and hence also underwent acoustic oscillations as they were dragged around by the photons. Only around recombination, when the baryons become neutral, could they freely fall into the dark matter potential wells. The baryons therefore *do* have acoustic oscillations in their amplitude at recombination, and this imprints itself into the total matter spectrum at late times (suppressed by a factor of ρ_b/ρ_c because most of the matter - the dark matter - does not have acoustic oscillations). These are called *baryon oscillations* (BAO), and are associated with nearly the same physical scale as the acoustic peaks in the CMB. Since the acoustic oscillation scale is imprinted at recombination, the scale can be used as a fixed comoving *standard ruler*: measurements of the BAO scale as a function of redshift can be used to measure the comoving angular diameter distance as a function of redshift.

C. Galaxy clustering

Of course we can't measure the matter power spectrum directly, because most of the matter is dark. What we can do instead is measure the galaxy power spectrum: if we count the local number density of galaxies, how does this fluctuate as a function of scale? It turns out that the galaxy power spectrum $P_g(k)$ is in general rather different to $P(k)$ of the total matter because of the complicated physics of galaxy formation, but on very large scales the result is simple: $P_g \propto P(k)$. The constant of proportionality is given by b^2 where b is called the *bias*.

This can be understood as follows. Galaxies are going to form at the peaks of the density fluctuations (those places above some threshold density that they can undergo gravitational collapse despite baryon pressure), so P_g is actually measuring the spectrum of the peaks in the underlying density field. If you put in a large-scale overdensity, this will push lots of just-below threshold lumps over the limit, and you will produce more galaxies. How many you push over the threshold depends on how high the critical threshold is, so the response in the number of galaxies generally varies depending on the value of the threshold, which can depend on the redshift and the type of galaxy you are looking at. Generally $b \geq 1$, reflecting the fact that rare overdensity peaks are rather sensitive to small changes in the large-scale density perturbations, and hence are more clustered than the underlying matter density perturbations.

Since on large scales $P_g(k) \propto P(k)$ it is possible to use galaxy clustering observations to constrain $P(k)$, albeit over a limited range of scales and with an additional bias parameter. The baryon oscillation signal is however rather simpler: since most of the information is in the *scale* of the ('standard ruler') oscillation scale, it is much less sensitive to bias (which just moves the amplitude of the spectrum up and down), and is thought to provide a fairly robust way to constrain cosmological models. In particular by measuring the BAO scale as a function of redshift, the late-time expansion history can be constrained and hence background cosmological parameters and test for evolution in the dark energy density.

D. Gravitational lensing

One way to probe the total matter density is via the gravitational lensing effect that it produces. In a perturbed universe light paths bend, and this leads to the shape of distant objects being distorted. If we can measure this distortion, it constrains the amount of lensing, and hence the amount and distribution of matter. In practice this is usually done by observing large numbers of distant galaxies, and arguing that *on average* galaxies should have independent circular profiles, so that any correlation observed between galaxy shapes must be due to lensing. With this assumption the lensing can be measured statistically (limited in precision by the fact that the galaxies have a rather wide dispersion in intrinsic shapes, which gives lots of chance non-lensing alignments). This is a way to probe a combination of the total matter density and background geometry along different lines of sight. It is intrinsically an integral constraint because the distortion we observe depends on the total effect of all the lenses between us and the source galaxy. In practice it can also be complicated by the fact that galaxy shapes may not be statistically independent, for example nearby galaxies may have their shapes aligned with the local gravitational tidal field.

E. Velocities: redshift distortions

Another indirect way to probe the total matter densities is via velocities. The idea is that the gravitational acceleration $\nabla\Phi$ depends on the gravitational potential, and hence on the total matter. For linear perturbations we can solve for how the velocities relate to the potentials, so by measuring the velocities we can probe the total matter. The line-of-sight component of the velocities can be measured from their Doppler effect on the observed redshift, so called *redshift distortions*. If there is a big overdensity, an galaxy on the far side will typically be falling in to it, and hence appear slightly Doppler blue shifted. If, a priori, we don't know the velocity and just convert the redshift into a distance, we would infer that the galaxy is actually slightly nearer than it really is. Likewise a galaxy on the near side of the overdensity would be moving away from us, and we would infer that

is slightly further away than it really is. I.e. galaxies appear to move along the line of sight towards the overdensity. This distortion in the apparent number densities along the line of sight compared to transverse to the line of sight is what makes redshift distortions observable, and hence allows them to be used to constrain cosmology.