Lecture Notes on

Cosmology

Kevin Zhou

kzhou7@gmail.com

These notes cover introductory cosmology. Nothing in these notes is original; they have been compiled from a variety of sources. The primary sources were:

- Daniel Baumann's Cosmology lecture notes.
- Ryden, *Introduction to Cosmology.* A well-written undergraduate-level introduction to cosmology, assuming very little background; ideal for a quick pass to get the intuition.
- Mukhanov, *Physical Foundations of Cosmology.* An introduction to cosmology and astroparticle physics with a distinctly Russian flavor. Prerequisites are kept to a minimum; the text provides very brief but self-contained introductions to general relativity and field theory. Many calculations usually done numerically are performed analytically, leading to the appearance of many special functions.
- Kolb and Turner, *The Early Universe.* A classic textbook on cosmology and astroparticle physics, notably devoting an entire chapter to axions. However, many observational statements are out of date.

The most recent version is here; please report any errors found to kzhou7@gmail.com.

Contents

1 Geometry and Dynamics

1.1 Introduction

We begin with some useful numbers from astrophysics.

- The astronomical unit is the distance from the Earth to the Sun,

$$1\,\mathrm{AU} = 1.5 \times 10^{11}\,\mathrm{m}.$$

- One parsec is the distance at which an astronomical unit subtends an arcsecond, which is $1/60^2$ of a degree, about the angular resolution of an amateur telescope. In particular,

$$1\,\mathrm{pc} = 3 \times 10^{16}\,\mathrm{m} = 3\,\mathrm{ly}.$$

- Distances between stars are on the order of parsecs. Galactic distances are in kpc, intergalactic distances are in Mpc, and galaxies are arranged into superclusters separated by voids, both of which have sizes on the order of 100 Mpc. The width of the observable universe is on the order of Gpc, as the age of the universe is measured in gigayears (Gyr).
- The sun has mass $M_\odot = 2 \times 10^{30}\,\mathrm{kg}$, and the galaxy has mass about $10^{12} M_\odot$.
- The sun has luminosity $L_\odot = 4 \times 10^{26}\,\mathrm{W}$, and the galaxy has luminosity about $10^{10} L_\odot$.

Next, we consider some fundamental cosmological observations.

- The cosmological principle states that the universe is isotropic and homogeneous at scales above 100 Mpc. These two conditions are independent; neither implies the other. This can be viewed as merely a convenient simplifying approximation, as structures on the 100 Mpc scale and larger have been observed.
- For comparison, the observable patch of the universe has scale 3000 Mpc. This does not set a bound on the total size of the universe, which could be infinite. For instance, most inflationary theories predict a breakdown of homogeneity on scales much larger than 3000 Mpc, which are unobservable. As such, we'll focus on our observable patch.
- This is related to Olbers' paradox: the night sky is not infinitely bright, though naively it would be assuming a homogeneous, infinite universe. The resolution is that the age of the universe is finite, so light from very distant stars can't have reached us yet.
- Light from distant galaxies is redshifted. The redshift is defined as

$$z = \frac{\lambda_{\mathrm{ob}} - \lambda_{\mathrm{em}}}{\lambda_{\mathrm{em}}} > 0$$

  and Hubble's law is the observation that, for nearby galaxies,

$$z = \frac{H_0}{c} r, \quad H_0 = 100h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}, \quad h = 0.67 \pm 0.01.$$

  In the nonrelativistic limit $z \ll 1$, this means the galaxies are receding with velocity $v = H_0 r$. Note that while intragalactic distances can be measured with parallax, intergalactic distances must be measured by galactic luminosity or standard candles.

- Hubble's law is consistent with homogeneity and isotropy, as every galaxy observes recession obeying the law. In fact, it is the only expansion law consistent with homogeneity and isotropy, and as such remains unchanged when accounting for relativistic effects.

- We take the convention that a zero subscript denotes the current time. If we naively assume the galactic velocities are constant, then Hubble's law suggests all galaxies were at the same point in a "Big Bang" at time $t = 0$, where

$$t_0 = H_0^{-1} \approx 14\,\text{Gyr}.$$

  This is only a rough estimate, because we expect gravity to slow down the expansion, and dark energy to accelerate the expansion.

- We've also measured the constituents of the universe. Baryonic matter consists mostly of protons and neutrons, though we also count electrons as "baryonic", so the word really means "the stuff ordinary stuff is made of". Baryonic matter is about 3/4 hydrogen by mass, and most of the rest is helium. About 2% consists of heavier elements, generally called "metals".

- Several independent measurements indicate dark matter, i.e. massive components of the universe which can't be detected readily. Dark matter includes stellar remnants and substellar objects such as brown dwarfs, and possibly new particles.

- Note that Hubble's law can also be explained by a 'steady state' model. This model assumes the 'perfect' cosmological principle, which assumes homogeneity in time; that means H is constant, and distances grow exponentially. The matter density can be kept constant by assuming that new matter is created at a constant rate per unit volume.

- However, the universe also contains light with a blackbody spectrum at temperature

$$T_0 = 2.725 \pm 0.001\,\text{K}$$

  called the CMB. Its existence can be explained by the Big Bang model but not the steady state model, and was a key piece of evidence in the historical debate.

- Specifically, suppose distance at time t are scaled by $a(t)$, where we conventionally set $a(t_0) = 1$. The cosmological redshift means the temperature of the CMB is $T \propto a^{-1}$, so the CMB points to an era where universe was much hotter than currently.

- The redshift of distant galaxies can be derived from the scale factor; we have

$$1 + z = \frac{1}{a(t_1)}$$

  by cosmological redshift, where t_1 is the emission time. Taylor expanding gives Hubble's law,

$$z \approx H_0 d, \quad H_0 \equiv \frac{\dot{a}(t_0)}{a(t_0)}$$

  where we have set $c = 1$ and will do so henceforth.

- The relatively recent observation that the acceleration of the universe is accelerating points to the presence of dark energy, which makes up about 2/3 of the energy density of the universe.

Note. There are several ways to establish $T \propto a^{-1}$ for a photon gas. Formally, the geodesic equation for photons in the FRW metric shows they are redshifted by a factor of a. By Planck's law, this maps a blackbody spectrum to another, with a temperature shrunk by a.

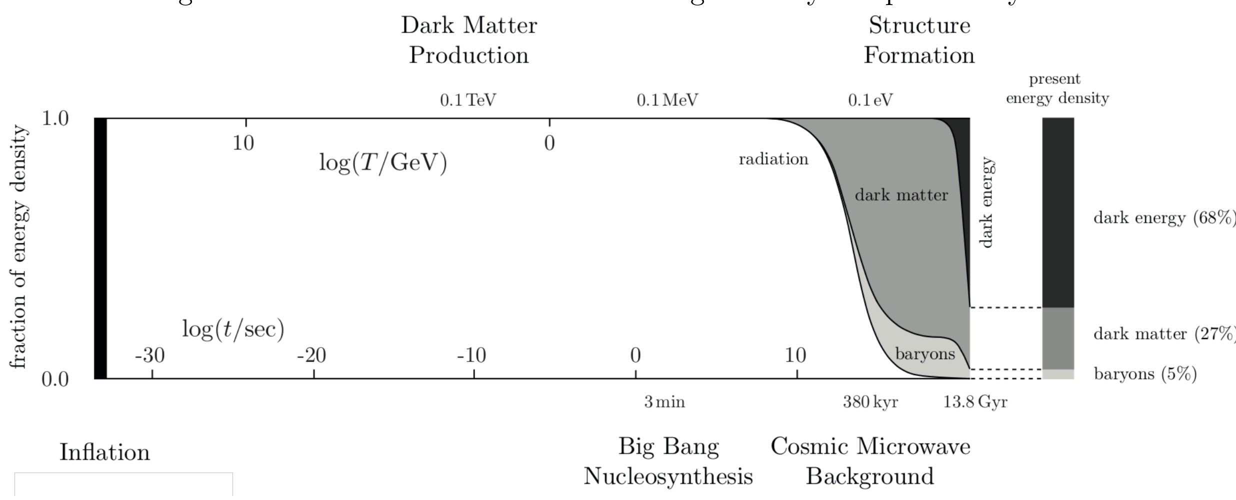
A slick method is to consider consecutive wavecrests of light. Consider two wavecrests emitted at times t_1 and $t_1 + \delta t_1$, and absorbed at times t_0 and $t_0 + \delta t_0$. Since null paths in the radial direction obey $dr/dt = 1/a$,

$$\int dr = \int_{t_1}^{t_0} \frac{dt}{a} = \int_{t_1+\delta t_1}^{t_0+\delta t_0} \frac{dt}{a}, \quad \delta t_0 = \frac{\delta t_1}{a(t_1)}$$

which implies the frequency is redshifted by $a(t_1)$ as before.

A final argument is from thermodynamics. For a photon gas, $N \propto VT^3$, and in an adiabatic expansion, no photons are absorbed or emitted by the walls. Then VT^3 must be constant, so $T \propto a^{-1}$ as before. Physically, the photons are redshifted by bouncing off the walls.

Note. The diagram above shows the standard cosmological story accepted today.



In this picture, an early inflationary era occurred, with quantum fluctuations inflated into large-scale fluctuations in the matter density of the universe. Within the first three minutes, the temperature cooled enough to form nuclei. About 380 kyr afterwards, the universe cooled enough to form neutral atoms and became transparent to radiation, forming the CMB. The initial inhomogeneities were imprinted on the CMB and amplified by gravity, creating the large-scale structure of the universe.

1.2 The Metric

We now introduce the Friedmann-Robertson-Walker (FRW) metric, beginning with spatial metrics.

- Spatial homogeneity and isotropy mean that spacetime can be foliated into spatial hypersurfaces, each of which are homogeneous and isotropic. Thus we begin by classifying these three-dimensional surfaces.

- Such spaces must have uniform curvature, and there are only three options.
  - Zero curvature space is three-dimensional Euclidean space E^3, $d\ell^2 = d\mathbf{x}^2$.
  - Positively curved space may be represented as a sphere S^3 embedded in E^4,

    $$d\ell^2 = d\mathbf{x}^2 + du^2, \quad \mathbf{x}^2 + u^2 = a^2.$$

    Homogeneity and isotropy result from rotational symmetry of the sphere.

- Negatively curved space may be represented as a hyperboloid H^3 embedded in Minkowski space with signature $(-+++)$,

$$d\ell^2 = d\mathbf{x}^2 - du^2, \quad \mathbf{x}^2 - u^2 = -a^2.$$

Then homogeneity and isotropy result from Lorentz symmetry. Note that the popular depiction of negatively curved space is a saddle in *Euclidean* space, but in this case the curvature is not uniform.

- In the last two cases, we rescale $\mathbf{x}$ and u by a to get

$$d\ell^2 = a^2(d\mathbf{x}^2 \pm du^2), \quad \mathbf{x}^2 \pm u^2 = \pm 1.$$

Then $\mathbf{x}$ and u are dimensionless while a carries units of length.

- Next, we can eliminate the parameter u, which gives

$$d\ell^2 = a^2 \left(d\mathbf{x}^2 \pm \frac{(\mathbf{x} \cdot d\mathbf{x})^2}{1 \mp \mathbf{x}^2} \right)$$

for the last two cases. We can combine all three cases by writing

$$d\ell^2 = a^2 \gamma_{ij} dx^i dx^j, \quad \gamma_{ij} = \delta_{ij} + k \frac{x_i x_j}{1 - k(x_k x^k)}, \quad k = \begin{cases} 0 & \text{Euclidean}, \\ 1 & \text{spherical}, \\ -1 & \text{hyperbolic}. \end{cases}$$

Now we don't have extra coordinates, though the homogeneity and isotropy is less manifest.

- To recover some of the manifest symmetry, we define spherical coordinates as usual, so

$$d\mathbf{x}^2 = dr^2 + r^2 d\Omega^2, \quad \mathbf{x} \cdot d\mathbf{x} = r dr.$$

Then the metric simplifies to

$$d\ell^2 = a^2 \left(\frac{dr^2}{1 - kr^2} + r^2 d\Omega^2 \right)$$

which at least makes isotropy manifest.

- Finally, one may simplify the radial component by defining a new radial coordinate

$$d\chi = \frac{dr}{\sqrt{1 - kr^2}}, \quad d\ell^2 = a^2 \left(d\chi^2 + S_k^2(\chi) d\Omega^2 \right)$$

where integration yields

$$S_k(\chi) = \begin{cases} \sinh \chi & k = -1, \\ \chi & k = 0, \\ \sin \chi & k = 1. \end{cases}$$

Note that while $S_k(\chi)$ is defined piecewise, the metric varies "smoothly" as the curvature is varied; we just can't see this since we've factored out the curvature scale.

Next, we introduce the FRW metric.

- To get the FRW metric, we simply add a time dimension,

$$ds^2 = dt^2 - a^2(t)\gamma_{ij}dx^i dx^j = dt^2 - a^2(t)\left(\frac{dr^2}{1-kr^2} + r^2 d\Omega^2\right)$$

  where we allow the parameter a to depend arbitrarily on time. Note that we can rescale

$$a \to \lambda a, \quad r \to r/\lambda, \quad k \to \lambda^2 k$$

  while preserving the metric. Conventionally, we use this freedom to make a dimensionless with $a(t_0) = 1$, so that r and $k^{-1/2}$ gain dimensions of length.

- This form is the most general form; we can always set $g_{00} = 1$ by rescaling the time coordinate. There are no cross terms g_{0i}, as this would break isotropy. However, note that the FRW spacetime is not stationary or static.

- In principle, we could also consider spacetimes that are not homogeneous and isotropic, but few exact solutions are known. As one example, the Lemaitre-Tolman-Bondi metric is isotropic but not homogeneous, describing an observer at the center of a large void or cluster. However, we will not consider these complications below.

- The coordinates x^i are called comoving coordinates. Physically, observers stationed at these coordinates will see the other observers moving away from them isotropically, and the time they measure is t.

- More specifically, define the 'physical' coordinates $x_p^i = a(t)x^i$, so differences in the physical coordinate better reflect physical distances. Then

$$v_p^i = \frac{dx_p^i}{dt} = a(t)\frac{dx^i}{dt} + \frac{da}{dt}x^i = a(t)\frac{dx^i}{dt} + Hx_p^i.$$

  These terms are called the peculiar velocity and Hubble flow, where we defined $H = \dot{a}/a$. Hence we have derived Hubble's law.

- Finally, it is convenient to work in conformal time. Letting $d\tau = dt/a(t)$, we have

$$ds^2 = a^2(\tau)\left(d\tau^2 - (d\chi^2 + S_k^2(\chi)d\Omega^2)\right).$$

  This is especially useful for analyzing the paths of light rays, since the overall scale factor doesn't matter. We see that for a radial path, the conformal time elapsed is equal to the change in the (scaled) radial coordinate, $\Delta\tau = \Delta\chi$. It's also obvious that two successive wavecrests for light separated by $\Delta\tau$ arrive separated by $\Delta\tau$, giving $\delta t \propto a(t)$ as before.

- Note that FRW spacetimes with $k = 0$ are only *spatially* flat; the Riemann tensor does not vanish! However, the Riemann tensor pulled back to a slice of constant t does vanish.

Next, we consider geodesics in the FRW spacetime.

- We will use the geodesic equation in the form

$$p^\nu \partial_\nu p^\mu = -\Gamma^\mu_{\nu\rho} p^\nu p^\rho$$

and we evaluate the Christoffel symbols in the coordinates where

$$ds^2 = dt^2 - a^2(t)\gamma_{ij} dx^i dx^j.$$

Some explicit calculation gives

$$\Gamma^0_{ij} = a\dot{a}\gamma_{ij}, \quad \Gamma^i_{0j} = \frac{\dot{a}}{a}\delta^i_j, \quad \Gamma^i_{jk} = \frac{1}{2}\gamma^{i\ell}(\partial_j \gamma_{k\ell} + \partial_k \gamma_{j\ell} - \partial_\ell \gamma_{jk})$$

with Γ^i_{j0} related by symmetry.

- Since the FRW metric is homogeneous, $\partial_i p^\mu = 0$, so we have

$$p^0 \frac{dp^\mu}{dt} = -\Gamma^\mu_{\nu\rho} p^\nu p^\rho = -\left(2\Gamma^\mu_{0j} + \Gamma^\mu_{ij} p^i\right) p^j.$$

In particular, for massive particles at rest in the comoving frame, $p^i = 0$, we have $dp^\mu/dt = 0$.

- Next, consider the $\mu = 0$ component. Here we have

$$E\frac{dE}{dt} = -\Gamma^0_{ij} p^i p^j = -\frac{\dot{a}}{a} p^2.$$

Since $E^2 = p^2 + m^2$, we have $E\,dE = p\,dp$, so this equation reduces to

$$p\,dp = \frac{da}{a} p^2, \quad p \propto \frac{1}{a}.$$

Hence for a massless particle, we have derived $E \propto 1/a$ as anticipated earlier.

- For massive particles, we instead have

$$p = \frac{mv}{\sqrt{1-v^2}} \propto \frac{1}{a}$$

where v^2 is defined with respect to the spatial metric γ_{ij}. This means that v^2 decreases as a increases, so freely falling particles converge onto the Hubble flow.

Note. The result that $T \propto 1/a^2$ for nonrelativistic matter can also be understood in the Newtonian picture. We imagine a population of particles exploding out from the origin. Once the particle cloud expands by a factor of a, the local dispersion in the velocities within a box of fixed size decreases by $1/a$. Since this defines the temperature, $T \propto \sigma_v^2 \propto 1/a^2$.

It is tricky to define distance on cosmological scales, and we give a few ways below.

- We consider the FRW metric in the form

$$ds^2 = dt^2 - a^2(t)\left(d\chi^2 + S_k^2(\chi) d\Omega^2\right), \quad S_k(\chi) = \begin{cases} R_0 \sinh(\chi/R_0) & k = -1, \\ \chi & k = 0, \\ R_0 \sin(\chi/R_0) & k = 1, \end{cases}$$

where we have rescaled so that $a(t)$ is dimensionless and $a(t_0) = 1$.

- Consider a point at the origin and a point at coordinate χ. We define the comoving distance between them by χ, and the metric distance by

$$d_m = S_k(\chi)$$

and the two agree for $k = 0$. Neither can be directly measured.

- The metric distance behaves a bit strangely; for example for $k = 1$ it hits a maximum and decreases. We introduce it because it is simply related to empirical distance measures, and because it is close to the comoving distance for scales less than the scale of the entire universe.

- The comoving distance is the current proper distance between the points. Since a radial light ray satisfies $dt = a(t)\,d\chi$, the comoving distance obeys

$$\chi(z) = \int_{t_1}^{t_0} \frac{dt}{a(t)} = \int_0^z \frac{dz}{H(z)}$$

where $H(z)$ is the Hubble parameter at the time when light with redshift z was emitted.

- On solar system scales, we measure distances using the speed of light. Since galaxies are on the scale of parsecs, we can use parallax to measure galactic distances, since the angles involved are on the order of arcseconds. Cosmological effects are negligible on these scales.

- For larger distances, we typically use standard candles. If a standard candle has luminosity L and the flux is F, we define the luminosity distance by

$$F = \frac{L}{4\pi d_L^2}$$

where d_L is the luminosity distance.

- For $k \neq 0$, the flux is spread over a surface area $4\pi S_k^2(\chi)$. The rate of arrival of photons is redshifted by $1 + z$, while the energy of each photon is also redshifted by $1 + z$. Then

$$d_L = d_m(1 + z).$$

That is, d_L overestimates d_m because of the cosmological redshift.

- Another option is to use a 'standard ruler' of known proper length D, defining the angular diameter distance

$$d_A = \frac{D}{\delta\theta}.$$

If the ruler lies along the tangential direction at comoving distance χ, the FRW metric gives

$$D = a(t_1)S_k(\chi)\delta\theta, \quad d_A = \frac{d_m}{1 + z}$$

The expansion of the universe makes d_A underestimate d_m.

Next, we briefly discuss how these distances are measured in practice.

- Standard yardsticks are difficult to use. Galaxies and galaxy clusters have been candidates, but they don't have well-defined borders, and their size can change over time. Instead, we mainly use standard candles.

- Cepheid variables are supergiant stars about 400 to $40,000$ times more luminous than the sun. They pulsate radially, with a period on the order of days to months. By studying clusters of Cepheids in the Large Magellanic Cloud, a simple relationship between the mean flux and period was found, make them standard candles.
- The main problem with Cepheids is calibrating the relationship: the closest Cepheid is hundreds of parsecs away, with a high distance uncertainty. Another problem is that they aren't bright enough to go to the 100 Mpc scales where the universe is homogeneous and isotropic, so results from them must be corrected for local peculiar velocities.
- Galaxies can be standard candles, since they are bright enough to go beyond 1 Gpc. However, their brightness is hard to predict; the Tully-Fisher gives some approximate information.
- Type 1a supernova work as standard candles at very high distances; they are standardized because they are all produced by white dwarfs in binary star systems passing the Chandrasekhar limit. Studies of supernova in the late 90's established that the expansion of the universe was accelerating, pointing to the existence of dark energy.

1.3 Dynamics

Next, we consider the matter content of the universe.

- We begin with a simpler case. Consider a set of particles with four-velocity u^μ. We define the number current N^μ, where $n = N^0$ is the number density and N^i is the number flux. Then
$$N^\mu = nu^\mu$$
and the conservation law $\nabla_\mu N^\mu = 0$ implies $n(t) \propto a^{-3}$.
- Next, consider a perfect fluid, with energy-momentum tensor
$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu - pg^{\mu\nu}.$$
The components of the tensor with mixed indices are simple,
$$T^\mu_{\ \nu} = \mathrm{diag}(\rho, -p, -p, -p).$$
- The conservation of the stress-energy tensor, $\nabla_\mu T^{\mu\nu} = 0$, gives four conservation equations. The component $\nu = 0$ gives the energy conservation equation
$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + p) = 0$$
in the FRW metric. This may also be derived using $dU = -pdV$, where $U = \rho V$ and $V \propto a^3$.
- Most cosmological substances obey the equation of state $p = w\rho$, where $\rho \propto a^{-3(1+w)}$ by the above. Matter refers to anything with $w \approx 0$, including baryonic matter and dark matter. Then $\rho \propto a^{-3}$, which can also be found using number conservation.
- Radiation refers to anything with $w \approx 1/3$, which holds for gases of ultrarelativistic particles, such as neutrinos in the early universe, and gases of massless particles, such as photons and gravitons; this result may be derived for photons by noting that $T^{\mu\nu}$ is traceless, because the Maxwell action is conformally invariant. Then $\rho \propto a^{-4}$, with the extra factor due to cosmological redshift.

- Dark energy is predicted to have $w = -1$, and a constant energy density $\rho \propto a^0$. It may arise in QFT from the vacuum energy, which is not diluted by the expansion of the universe. As we'll see below, anything with $w < -1/3$ accelerates the expansion of the universe.
- Fluids with $p = w\rho$ have linear dispersion relations, with a speed of sound of $\sqrt{w}$. Hence we require $w \leq 1$ to preserve causality. Alternatively, the NEC requires $|w| \leq 1$.
- Evidently sound waves do not exist for $w < 0$, as the medium would be unstable; this is not a problem for vacuum energy since it doesn't allow fluctuations in p and ρ in the first place.
- Note that energy densities due to ordinary matter must be positive, since negative energy densities would imply that vacuum would be unstable against decay. On the other hand, given this fact, the vacuum energy can be negative, giving a negative dark energy, because by definition it cannot decay.

In order to find how the scale factor evolves, we have to evaluate the Einstein tensor.

- The calculation is simplified using symmetry. By isotropy, we know that $R_{i0} = 0$, or else it would give a distinguished 3-vector. Similarly, R_{ij} must be proportional to g_{ij}, since there are no distinguished tensors, and by homogeneity the proportionality constant must be the same everywhere.
- Using the Christoffel symbols computed earlier easily gives
$$R_{00} = -3\frac{\ddot{a}}{a}.$$
- To compute R_{ij}, we work at $\mathbf{x} = 0$. The spatial metric is
$$\gamma_{ij} = \delta_{ij} + \frac{kx_i x_j}{1 - k(x_k x^k)} = \delta_{ij} + kx_i x_j + O(x^4)$$
and we only need to maintain terms up to quadratic order in x, because the Ricci tensor only contains second derivatives of the metric. We then have
$$\Gamma^i_{jk} = \frac{1}{2}\gamma^{i\ell}(\partial_j \gamma_{k\ell} + \partial_k \gamma_{j\ell} - \partial_\ell \gamma_{jk}) = \frac{1}{2}(\delta^{i\ell} - O(x^2))(2k\delta_{jk}x_\ell) = kx^i \delta_{jk}$$
where we again threw away higher-order terms in x because the Ricci tensor only contains first derivatives of the connection. Straightforwardly plugging in gives
$$R_{ij} = -\left(\frac{\ddot{a}}{a} + 2\left(\frac{\dot{a}}{a}\right)^2 + 2\frac{k}{a^2}\right) g_{ij}.$$
- Therefore, the Ricci scalar is
$$R = -6\left(\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2}\right).$$
This implies the Einstein tensor with mixed indices is
$$G^0{}_0 = 3\left(\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2}\right), \quad G^i{}_j = \left(2\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2}\right)\delta^i_j.$$

- The Einstein field equations are the Friedmann equations,

$$\boxed{\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{k}{a^2}, \quad \frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p).}$$

  The second is also called the acceleration equation; we've already seen it in general relativity, where it is a special case of the fact that energy density and pressure cause geodesics to contract. The first Friedmann equation is also called *the* Friedmann equation.

Note. A fake derivation of the Friedmann equation. Consider space filled with matter of density ρ, and a mass a distance r from an arbitrary center. Then conservation of energy gives

$$\frac{1}{2}m\dot{r}^2 - \frac{Gm\rho}{r}\frac{4\pi r^3}{3} = E.$$

Defining $r(t) = a(t)R_0$, where $R_0 = r(0)$, we may eliminate r to find,

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho + \frac{2E}{mR_0^2}\frac{1}{a^2}$$

which is simply the Friedmann equation up to some redefinitions. The reason this fake derivation works is that r drops out of the Friedmann equation, as it must by the homogeneity of Hubble expansion. Hence we can consider the limit of small r, where relativistic corrections are negligible.

In this Newtonian picture, closed and open universes correspond to negative and positive total energy, which makes it clear that a closed universe can't have $a(t) \to \infty$. This remains true when radiation is included, but is *not* true if there is dark energy.

The idea that "density is destiny" is a bit confusing in the Newtonian picture, because the density only determines the potential energy, while the total energy is what matters. We focus on the density because the "kinetic energy" was found much earlier, as it is determined by H_0.

Note. In a closed universe, the total electric charge must be zero, because the electric field lines have nowhere to end. Alternatively, one can cover the universe with two patches, and the electric flux going out of one patch equals the flux going into the other; then the total charge vanishes by Gauss's law. One can also argue similarly that the total energy vanishes in a closed universe. To see this here, rearrange the Friedmann equation by multiplying by Ma^2 for

$$M\dot{a}^2 + M - \frac{8\pi}{3}G\rho M a^2 = 0$$

where M is the total mass. Using $M = \rho V$ and $V = 2\pi^2 a^3$, we have

$$M\dot{a}^2 + M - \frac{4}{3\pi}\frac{GM^2}{a} = 0.$$

Then the Friedmann equation can be interpreted as the statement that the kinetic energy, mass energy, and gravitational potential energy sum to zero. We didn't get this result in the Newtonian picture, because there we didn't count the mass energy.

Note. One needs to be careful to keep track of the dimensions. In our relativistic derivation of the Friedmann equation, the curvature term was $-k/a^2$, the scale factor had dimensions of length, and

$k \in \{-1, 0, 1\}$. If we rescale the scale factor to be dimensionless, with $a(t_0) = 1$, then the Friedmann equation must pick up factors of R_0 to balance the dimensions,

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{k}{R_0^2 a^2}.$$

which is just what we saw in our Newtonian derivation. We can thus either maintain $k \in \{-1, 0, 1\}$, or perform a further redefinition $\tilde{k} = k/R_0^2$ so that $\tilde{k}$ has dimensions. Unfortunately, many sources do this wrong and write down equations that are dimensionally inconsistent.

Next, we consider the dynamics of some model universes, using dimensionless $a(t)$ and k.

- It is conventional to define

  $$H_0 = 100\,h\,\mathrm{km s^{-1} Mpc^{-1}} = 2.17 \times 10^{-20}\,\mathrm{s}^{-1}, \quad h = 0.67 \pm 0.01.$$

  A flat universe has the "critical density"

  $$\rho_{c,0} = \frac{3H_0^2}{8\pi G} = 1.9 \times 10^{-29} h^2\,\mathrm{g/cm^3} = 8.5 \times 10^{-27}\,\mathrm{kg/m^3}$$

  and universes with higher or lower density are closed or open, respectively. Note that the critical density is a function of time.

- We define dimensionless density parameters by dividing by the critical density, giving

  $$\boxed{H^2(a) = H_0^2\left(\Omega_{r,0} a^{-4} + \Omega_{m,0} a^{-3} + \Omega_{k,0} a^{-2} + \Omega_{\Lambda,0}\right).}$$

  We conventionally drop the '0' subscripts on the density parameters. Here Ω_k is the effect of the curvature,

  $$\Omega_k = -\frac{k}{(a_0 H_0 R_0)^2}$$

  which we formally treat as a contribution to the density with $w = -1/3$, and

  $$\Omega_r + \Omega_m + \Omega_k + \Omega_\Lambda = 1.$$

  However, note that the total energy density is actually $\Omega = 1 - \Omega_k$.

- Rearranging, we have

  $$\dot{a} = H_0\sqrt{\Omega_r a^{-2} + \Omega_m a^{-1} + \Omega_k + \Omega_\Lambda a^2}.$$

  Thus for the purposes of intuition, we can use the Newtonian picture above, where Ω_m behaves like Newtonian matter, and Ω_Λ is a repulsive spring.

- Note that we've defined the density parameters to be time-independent above for convenience. It's also sensible to talk about time-dependent density parameters, by dividing by $\rho_c(t)$.

- Current observations indicate

  $$|\Omega_k| \leq 0.01, \quad \Omega_r = 9.4 \times 10^{-5}, \quad \Omega_m = 0.32, \quad \Omega_\Lambda = 0.68$$

  and the matter splits into baryonic and cold dark matter (CDM) as

  $$\Omega_b = 0.05, \quad \Omega_c = 0.27.$$

  Only a small part of the radiation contribution is from stars; most is from the CMB and the cosmic neutrino background.

- We see the curvature parameter is unimportant now and hence was even less important in the past, so we will just set $\Omega_k = 0$ below.
- The fact that Ω_Λ has just recently passed Ω_m is called the 'cosmic coincidence problem', since the ratio $\Omega_m/\Omega_\Lambda \propto 1/a^3$ varies by many orders of magnitude.
- If we only have a single matter component, then
$$\frac{\dot{a}}{a} = H_0\sqrt{\Omega}a^{-(3/2)(1+w)}.$$
This implies the following dependence:
$$a(t) \propto \begin{cases} t^{2/3} & \text{matter} \\ t^{1/2} & \text{radiation} \\ e^{Ht} & \Lambda \end{cases} \qquad a(\tau) \propto \begin{cases} \tau^2 & \text{matter} \\ \tau & \text{radiation} \\ -1/\tau & \Lambda \end{cases}$$
This is often good enough, because the history of the universe can be divided into a time where radiation was dominant, a time where matter was dominant, and a time where the cosmological constant is dominant. We have only recently entered this third era, which is a minor theoretical puzzle.
- Note that in the case of Λ only, the universe would be infinitely old. This is the steady state theory of cosmology in another guise, replacing the spontaneously created matter replaced with dark energy.

Example. Consider a universe, not necessarily flat, with matter and dark energy, which is a good model for our universe today. If $\Omega_\Lambda > 0$, then the expansion is accelerated; in particular, the scale factor can grow arbitrarily large even if the universe is closed. On the other hand, if $\Omega_\Lambda < 0$, then the scale factor will always reach a maximum and begin to decrease, regardless of whether the universe is open or closed, ending in a 'big crunch'. For our universe, Ω_Λ is high enough that, regardless of the sign of the small quantity Ω_k, the universe will continue to expand forever.

Example. Consider a flat universe with matter and radiation; this describes the crossover period in the early universe. The simplest method is to work with conformal time, where the Friedmann equations become
$$(a')^2 = \frac{8\pi G}{3}\rho a^4, \quad a'' = \frac{4\pi G}{3}(\rho - 3p)a^3$$
which is convenient because radiation does not contribute to a'' at all. The density is
$$\rho = \rho_m + \rho_r = \frac{\rho_{\text{eq}}}{2}\left(\left(\frac{a_{\text{eq}}}{a}\right)^3 + \left(\frac{a_{\text{eq}}}{a}\right)^4\right), \quad a_{\text{eq}} = \frac{\Omega_r}{\Omega_m} \approx 3\times 10^{-4}.$$
Now the second equation can be simply solved,
$$a'' = \frac{2\pi G}{3}\rho_{\text{eq}}a_{\text{eq}}^3, \quad a(\tau) = \frac{\pi G}{3}\rho_{\text{eq}}a_{\text{eq}}^3\tau^2 + C\tau + D.$$
We impose $a(\tau = 0) = 0$, so $D = 0$, and by using the first Friedmann equation,
$$a(\tau) = a_{\text{eq}}\left(\left(\frac{\tau}{\tau_\star}\right)^2 + 2\frac{\tau}{\tau_\star}\right), \quad \tau_\star = \left(\frac{\pi G}{3}\rho_{\text{eq}}a_{\text{eq}}^2\right)^{-1/2} = \frac{\tau_{\text{eq}}}{\sqrt{2}-1}.$$
Thus for low and high τ we recover the appropriate limits.

Note. All of our examples have a Big Bang singularity where the scale factor is zero. One might think this is just an artifact of demanding homogeneity and isotropy; however, cosmological singularity theorems indicate that a singularity is generic, assuming certain conditions on the matter. Of course this does not mean a singularity necessarily exists, since at this point quantum gravity takes over.

Example. In the Einstein static universe, $a(t)$ is constant. This requires the right-hand sides of both Friedmann equations to vanish, so $\rho + 3p = 0$. Assuming the pressure and density are nonzero, Einstein could satisfy this equation by letting

$$\Omega_\Lambda = \frac{1}{2}\Omega_m.$$

The resulting spatial curvature parameter k is nonzero. The Einstein static universe is not realistic, but it remains a useful tool for constructing conformal diagrams.

Example. The Milne universe. Consider an open universe with $k = -1$ and nothing else. The Friedmann equation reduces to $\dot{a}^2 = 1$, which has a solution $a(t) = t$. The metric is then

$$ds^2 = dt^2 - t^2(d\chi^2 + \sinh^2 \chi \, d\Omega^2).$$

On the other hand, we would expect that an isotropic matter-free solution to Einstein's equations must be Minkowski space. In fact, we can regard the Milne spacetime as a subset of Minkowski space. Starting with

$$ds^2 = d\tau^2 - dr^2 - r^2 d\Omega^2$$

we arrive at the Milne universe if

$$\tau = t\cosh\chi, \quad r = t\sinh\chi.$$

To interpret this, consider observers which start from the origin in Minkowski space, each with constant velocity. For the observer with velocity v,

$$v = \frac{r}{\tau} = \tanh\chi.$$

The proper time elapsed for this observer is

$$\sqrt{1 - v^2}\,\tau = t.$$

Then the Milne universe describes observers moving uniformly outward from the origin of Minkowski space, with proper time t and their velocity labeled by χ.

Note that constant timeslices of the Milne universe have constant 3-curvature. However, since it is part of Minkowski space, it has vanishing 4-curvature. This is an important lesson to keep in mind: the 3-curvature is not determined by the spatial part of the 4-curvature.

Stepping back, it seems strange that we could describe the same spacetime as either $k = -1$ with $a(t) = t$, or as Minkowski space, which has $k = 0$ and $a(t) = 1$. The reason this is strange is that homogeneity restricts the choice of foliation of a spacetime to those with uniform energy density. In this unusual case there is zero energy density, and hence much greater freedom; the other spacetimes above generally can't be understood this way.

Example. A similar situation occurs for de Sitter space, which contains only a cosmological constant. By the acceleration equation alone, we have

$$\ddot{a} = H_0^2 a, \quad H_0^2 = \frac{8\pi G\rho}{3}.$$

This gives the general solution

$$a(t) = C_1 e^{H_0 t} + C_2 e^{-H_0 t}.$$

Plugging this into the Friedmann equation, we have

$$4H_0^2 C_1 C_2 = k.$$

This means that the form of the solution is different for different k,

$$ds^2 = dt^2 - H_0^{-2}(d\chi^2 + S_k^2(\chi)d\Omega^2) \times \begin{cases} \sinh^2(H_0 t) & k = -1 \\ e^{2H_0 t} & k = 0 \\ \cosh^2(H_0 t) & k = 1 \end{cases}.$$

However, because the energy density due to the cosmological constant is constant in time, all three of these are merely different foliations of the same spacetime, by homogeneous and isotropic hypersurfaces with different curvatures. Another way to construct de Sitter space is to embed it as a hyperboloid in five-dimensional Minkowski spacetime, as shown in the notes on General Relativity.

2 Inflation

2.1 Motivation

To understand the motivation for inflation, we consider the causal structure of our universe.

- We work in conformal time, so light rays are at 45° when plotting χ and τ. Let the initial and final conformal time of our universe be τ_i and τ_f. Note that τ_f may be finite even if the universe never ends in ordinary time; it is indeed finite for the standard model of our universe.

- The particle horizon at time τ is bounded by

$$\chi_{\text{ph}}(\tau) = \tau - \tau_i = \int_{t_i}^{t} \frac{dt}{a(t)}.$$

  Only events inside the particle horizon could have affected us. In other words, we can only see effects from events inside the particle horizon. (If we are interested in the furthest distance we could see light from, the lower bound should be the recombination time, as before this time the universe was opaque to photons; however, in practice this makes little difference.)

- The event horizon at time τ is bounded by

$$\chi_{\text{eh}}(\tau) = \tau_f - \tau = \int_{t}^{t_f} \frac{dt}{a(t)}.$$

  It is analogous to the event horizon for black holes, and bounds the spatial regions we can affect in the future. A numerical calculation shows that for the concordance model, objects that we can just reach in the future currently have a redshift of about 1.8.

- Note that we may write

$$\chi_{\text{ph}}(\tau) = \int_{t_i}^{t} \frac{dt}{a} = \int_{a_i}^{a} \frac{da}{a\dot{a}} = \int_{\log a_i}^{\log a} \frac{d\log a}{aH}.$$

  We call the comoving Hubble radius $(aH)^{-1}$, and it defines a comoving Hubble sphere.

- To understand its physical meaning, consider sending signals to an observer a comoving distance d away. If d is small, then the signal takes about time d to arrive, during which time the observer moves away by $d^2\dot{a}$. This is a small correction if $d^2\dot{a} \ll d$, which implies $d \ll 1/\dot{a} = (aH)^{-1}$. Hence the Hubble sphere contains observers we can actively communicate with "now", sending messages back and forth in roughly less than a Hubble time. Then it is intuitive that χ_{ph} is found by summing it over Hubble times.

- For a perfect fluid, we have

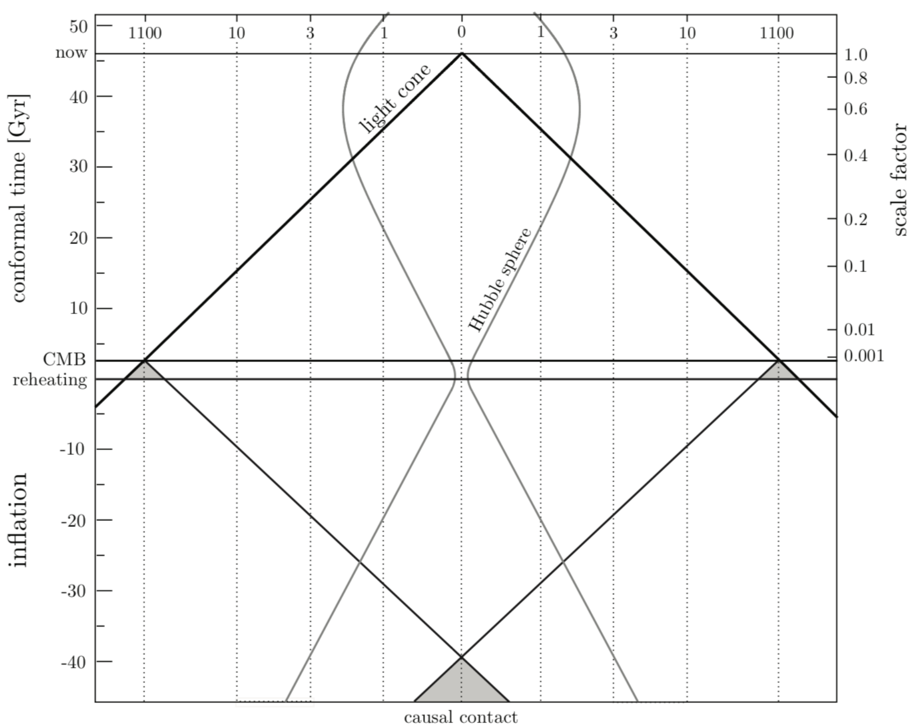
$$(aH)^{-1} = H_0^{-1} a^{(1+3w)/2}.$$

  All familiar forms of matter obey the SEC, which states $1 + 3w > 0$. Dark energy does not, but it is less important in the early universe. Hence the comoving Hubble radius should have monotonically increased in the early universe.

- In particular, the integral for $\chi_{\rm ph}$ is dominated by its upper bound. Explicitly, we have
$$\chi_{\rm ph}(a) = \frac{2H_0^{-1}}{1+3w}\left(a^{(1+3w)/2} - a_i^{(1+3w)/w}\right) \equiv \tau - \tau_i$$
where $\tau_i = 0$ since $a_i = 0$, giving
$$\chi_{\rm ph} \sim (aH)^{-1}$$
up to small constant prefactors; the latter is sometimes called the "Hubble horizon". This leads to a problem, because then $\chi_{\rm ph}$ is too small for different regions of the CMB to have ever been in causal contact; instead there are about 10^4 causally disconnected patches. The 'horizon problem' asks why they 'know' to be at the same temperature.

These problems can be addressed by a shrinking Hubble sphere.

- The essence of the problem is that we want to set $\chi_{\rm ph} \gg (aH)^{-1}$, which means we need a shrinking Hubble sphere, which requires a fluid that violates the SEC.
- Given an era with a shrinking Hubble sphere, it is perfectly possible for opposite ends of the CMB to have been in causal contact, because as $a \to 0$, we have $\tau \to -\infty$, giving a large range of conformal time to work with.
- Inflation postulates such an era, occupying negative conformal time. The "Big Bang" corresponds to the end of inflation, conventionally zero conformal time, at which point the Hubble sphere begins to expand again. A Minkowski diagram of the situation is shown below.



Recently, the Hubble sphere has begun to shrink again due to dark energy.

- A standard criterion to check if there is 'enough' inflation is
$$(a_I H_I)^{-1} > (a_0 H_0)^{-1}$$

which means that anything we are in causal contact with now, we were also in contact with in the past. This is not quite as strong as the actual condition, $(a_I H_I)^{-1} > \chi_{\text{ph}}$, but easier to evaluate.

- To estimate the amount of inflation needed, note that $H \propto a^{-2}$ during radiation domination; we focus on this period since most of the expansion occurred during it. Letting an E subscript stand for the exit of inflation,
$$\frac{a_0 H_0}{a_E H_E} = \frac{a_E}{a_0} = \frac{T_0}{T_E} \sim 10^{-28}$$
where we used the numbers
$$T_E \sim 10^{15}\,\text{GeV}, \quad T_0 = 10^{-3}\,\text{eV} = 2.7\,\text{K}.$$

- For simplicity, suppose H is approximately constant during inflation. Then we require
$$\frac{a_E H_E}{a_I H_I} \approx \frac{a_E}{a_I} > 10^{28}$$
which implies at least
$$N_{\text{tot}} > \log 10^{28} = 64$$
e-foldings during inflation. A more accurate accurate comes from the largest scales observed in the CMB, which have to be created $N_{\text{cmb}} = 60$ e-foldings before the end of inflation. Hence inflation must last for at least 60 e-foldings. In principle, it is not hard for it to last much longer, but it is the last 60 e-foldings which create the cosmological perturbations we see today.

Note. The flatness problem is that the universe is observed to be very close to flat today, with $|\Omega_k| \ll 1$. The severity of this problem is more apparent if we consider how the density parameters evolve in time. If the universe is very nearly flat, then
$$\Omega_k \propto \frac{1}{\dot{a}^2} \propto \begin{cases} t^{2/3} & \text{matter domination} \\ t & \text{radiation domination} \end{cases}.$$
That is, Ω_k is constantly growing, so if we extrapolate back naively to the Planck scale, we must have an initial condition of $\Omega_k \sim 10^{-60}$, which seems unnatural. Inflation solves this problem because, during the period of inflation, Ω_k shrinks exponentially.

Next, we consider some equivalent conditions for inflation.

- We note that
$$\frac{d}{dt}(aH)^{-1} = \frac{d}{dt}\frac{1}{\dot{a}} = -\frac{\ddot{a}}{(\dot{a})^2}.$$
Hence inflation corresponds to a period of accelerating expansion.

- Alternatively, note that
$$\frac{d}{dt}(aH)^{-1} = -\frac{\dot{a}H + a\dot{H}}{(aH)^2} = -\frac{1}{a}(1-\epsilon), \quad \epsilon = -\frac{\dot{H}}{H^2}.$$
Hence inflation occurs for $\epsilon < 1$, which corresponds to a slowly decreasing Hubble parameter. In fact, as we'll see below, we actually have $\epsilon \ll 1$ in most inflationary models.

- In the case of perfect inflation, $\epsilon = 0$, the scale factor increases exponentially, and
$$ds^2 = dt^2 - e^{2Ht} d\mathbf{x}^2.$$
Hence during inflation, the spacetime is approximately de Sitter. A bit more carefully, the very early, inflationary universe is approximately the same as a small slice of de Sitter space; the two have very different global structure.

- Using the continuity equation, we can show
$$\epsilon = \frac{3}{2}\left(1 + \frac{P}{\rho}\right) < 1$$
so inflation requires $w < -1/3$, as we saw earlier, and hence negative pressure.

- Again using the continuity equation, we can show
$$\left|\frac{d\log\rho}{d\log a}\right| = 2\epsilon < 1$$
so for small ϵ, the energy density is nearly constant.

- Let N be the cumulative number of e-folds. Using $dN = d\log a = H\,dt$, we can rewrite ϵ in the useful form
$$\epsilon = -\frac{d\log H}{dN}.$$
We already know $\epsilon < 1$ during inflation, and that inflation requires about $N = 60$, so ϵ itself must also change slowly. That is, we require
$$\eta = \frac{d\log\epsilon}{dN} = \frac{\dot{\epsilon}}{H\epsilon}, \quad |\eta| \ll 1.$$
Our task below will be to construct a model with both ϵ and η appropriately small.

Note. A notational subtlety. When we say the expansion of the universe is accelerating, we mean $\ddot{a} > 0$, and equivalently the Hubble sphere is shrinking. However, this is not equivalent to $\dot{H} > 0$. In fact, at the current moment we have $\ddot{a} > 0$ but $\dot{H} < 0$. The former means that a fixed object is moving away from us faster and faster, while the latter means that the objects present at a fixed distance will move away from us slower and slower.

2.2 Slow Rolling

We now consider a simple explicit model of inflation.

- We postulate a scalar field, called the inflaton ϕ, with potential $V(\phi)$ and stress-energy tensor
$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi - g_{\mu\nu}\left(\frac{1}{2}g^{\alpha\beta}\partial_\alpha\phi\partial_\beta\phi - V(\phi)\right).$$
By homogeneity, ϕ can only depend on t, so we have
$$T^0{}_0 = \rho_\phi = \frac{1}{2}\dot{\phi}^2 + V(\phi), \quad T^i{}_j = -p_\phi\delta^i_j, \quad p_\phi = \frac{1}{2}\dot{\phi}^2 - V(\phi).$$
Therefore, if the potential energy dominates over the kinetic energy, the scalar field behaves like a fluid with $w = -1$, which may lead to accelerated expansion.

- Substituting into the Friedmann equations with the substitution $M_{\text{pl}} = \sqrt{\hbar c/8\pi G}$ in units where $\hbar = c = 1$, we have

$$H^2 = \frac{\rho}{3M_{\text{pl}}^2} = \frac{1}{3M_{\text{pl}}^2}\left(\frac{1}{2}\dot{\phi}^2 + V\right), \quad \dot{H} = -\frac{\rho+p}{2M_{\text{pl}}^2} = -\frac{1}{2}\frac{\dot{\phi}^2}{M_{\text{pl}}^2}.$$

We note that $\dot{H}$ is sourced by the kinetic energy alone.

- Next, differentiating the first equation gives

$$2H\dot{H} = \frac{1}{3M_{\text{pl}}^2}\left(\dot{\phi}\ddot{\phi} + V'\dot{\phi}\right)$$

where $V' = dV/d\phi$, and using the second Friedmann equation gives the equation of motion

$$\boxed{\ddot{\phi} + 3H\dot{\phi} + V' = 0.}$$

The expansion of the universe adds "Hubble friction". Intuitively this is because it is "diluting" the field momentum.

- Substituting into the definition of ϵ, we have

$$\epsilon = \frac{\dot{\phi}^2/2}{M_{\text{pl}}^2 H^2}$$

so we again see that inflation occurs if the kinetic energy is small; this situation is called slow-roll inflation.

- For inflation to persist, we requires $|\eta| \ll 1$, and it is convenient to define

$$\delta = -\frac{\ddot{\phi}}{H\dot{\phi}}$$

which is the dimensionless acceleration per Hubble time. By rearranging our equations, we find $\eta = 2(\epsilon - \delta)$, so $|\eta|$ is indeed small if ϵ and $|\delta|$ are.

In order to make more analytic progress, we take the slow roll approximation.

- In the slow roll approximation, we take the Hubble parameter to be entirely determined by the potential energy because $\epsilon \ll 1$, so the Friedmann equation becomes

$$H^2 = \frac{V}{3M_{\text{pl}}^2}.$$

Also, since $|\delta| \ll 1$, we approximate the field's equation of motion as

$$3H\dot{\phi} \approx -V'.$$

- Next, we use these assumptions to compute ϵ and δ. Combining the equations, we have

$$\epsilon \approx \frac{M_{\text{pl}}^2}{2}\left(\frac{V'}{V}\right)^2.$$

Differentiating the equation of motion gives

$$3\dot{H}\dot{\phi} + 3H\ddot{\phi} \approx -V''\dot{\phi}$$

which implies that

$$\delta + \epsilon \approx M_{\text{pl}}^2 \frac{V''}{V}.$$

Therefore, the approximation is self-consistent as long as the slow-roll parameters

$$\boxed{\epsilon_v = \frac{M_{\text{pl}}^2}{2}\left(\frac{V'}{V}\right)^2, \quad |\eta_v| = M_{\text{pl}}^2 \frac{|V''|}{V}}$$

are both small. In this case, $\epsilon_v \approx \epsilon$ and $\eta_v \approx 2\epsilon - \eta/2$, so ϵ and $|\eta|$ are both small as desired. These parameters are useful because they can be computed from the inflaton potential alone.

- The number of e-folds of accelerated expansion is given by

$$N_{\text{tot}} = \int d\log a = \int H\, dt = \int \frac{H}{\dot{\phi}}\, d\phi = \int \frac{1}{\sqrt{2\epsilon}} \frac{d\phi}{M_{\text{pl}}}.$$

In order to simplify this, we replace ϵ with ϵ_v for

$$N_{\text{tot}} = \int \frac{V}{M_{\text{pl}}^2 V'} d\phi$$

which may be computed from the inflaton potential alone. We take inflation to end when $\epsilon_v = 1$, and take an initial field value ϕ_I.

Example. Inflation driven by a mass term,

$$V(\phi) = \frac{1}{2} m^2 \phi^2.$$

The slow-roll parameters are

$$\epsilon_v = \eta_v = 2\left(\frac{M_{\text{pl}}}{\phi}\right)^2$$

so inflation requires super-Planckian values for the inflaton. Specifically, the number of e-folds is

$$N(\phi_I) = \frac{\phi_I^2}{4M_{\text{pl}}^2} - \frac{1}{2}$$

and the fluctuations observed in the CMB are created at $\phi_{\text{CMB}} \approx 2\sqrt{N_{\text{cmb}}} M_{\text{pl}} \sim 15 M_{\text{pl}}$. We require the energy density to not be near the Planck scale, or else nonperturbative quantum gravity effects may become relevant; this yields the constraint $m \ll M_{\text{pl}}/15$.

Next, we briefly discuss what happens after inflation.

- During inflation, most of the energy density of the universe is in the form of the inflaton potential. At the end of inflation, this energy has been mostly converted to the kinetic energy of the inflaton field. Reheating is the process by which energy is transferred from the inflaton field to the particles of the SM, thereby starting the hot Big Bang.

- First, note that after inflation, the inflaton field begins to oscillate at the minimum of its potential. Approximating $V(\phi) = m^2\phi^2/2$ near the minimum, we have

$$\ddot{\phi} + 3H\dot{\phi} = -m^2\phi.$$

  The expansion timescale soon becomes much larger than the oscillation period, $H^{-1} \gg m^{-1}$, so we can neglect the friction term; the field then oscillates with frequency m.

- The energy continuity equation gives

$$\dot{\rho}_\phi + 3H\rho_\phi = -3Hp_\phi = -\frac{3}{2}H(m^2\phi^2 - \dot{\phi}^2)$$

  and the right-hand side averages to zero over one oscillation period. Then the oscillating field behaves like pressureless matter, with $\rho_\phi \propto a^{-3}$. This is not surprising, because a spatially uniform massive field can be viewed as a condensate of massive particles at rest. The fall in the energy density is reflected by a decrease of the oscillation amplitude.

- Note that there are other possibilities, depending on the potential. For example, if the potential has the form $V \propto \phi^n$, then $w \approx (n-2)/(n+2)$. If the potential is convex, such as $V \sim \log(|\phi|/\phi_c|)$, then we have $w \approx -1$ even after slow roll ends. This is because the oscillating scalar field spends most of its time near the potential walls, where the kinetic energy is negligible. If the potential is exponential, then the slow-roll conditions are either always or never satisfied, so such a potential is unsuitable.

- In order to transfer energy to SM particles, the inflaton must couple to other fields, so it can decay. Supposing the decay is slow, we have

$$\dot{\rho}_\phi + 3H\rho_\phi = -\Gamma_\phi\rho_\phi.$$

  However, if the inflaton can decay into bosons, the decay can be very rapid, involving a mechanism called parametric resonance due to Bose condensation effects. This kind of rapid decay is called preheating, since the bosons are created far from thermal equilibrium.

- The next step is thermalization. The particles created by inflaton decay interact, and perform further decays, until we arrive at a thermal soup of particles at temperature T_{rh}. This marks the start of the hot Big Bang.

- Note that some particles, such as gravitinos, might never reach thermal equilibrium. However, as long as their energy density is high, they will behave like radiation regardless. We only require thermalization of the baryons, photons, and neutrinos.

Note. Inflation also solves the monopole problem. Grand unified theories (GUTs) generally predict monopoles should be formed during the GUT phase transition in the early universe, as they are topological defects. We expect one topological defect per Hubble sphere, which yields a huge amount in the naive Big Bang model; it implies most matter should be made of magnetic monopoles today. Inflation solves this problem because, if the GUT phase transition occurs before or during inflation, then the monopoles are diluted away. Note that this requires that reheating does not yield a temperature above the GUT scale, but this is not a stringent constraint on most models.

Of course, if one doesn't believe in GUTs, then the monopole problem is not an issue. Note that all three problems that inflation solves can be phrased as naturalness problems; the flatness problem is closely related to the cosmological constant problem and has a similar anthropic solution.

Note. We've been careful to avoid conflating dark energy, vacuum energy, and the inflaton. Dark energy is a substance with $w = -1$ that drives the current accelerating expansion of the universe. Vacuum energy is a natural candidate for dark energy, but there are also dynamical models of dark energy (quintessence) where it can be sourced by a slowly rolling scalar field. This is the same basic mechanism as inflation, though the inflaton field is likely not related. In this case, we only have $w \approx -1$, though the value of w is not measured very precisely.

Note. What is the inflaton? In the simplest models, it is simply a single new scalar field. It could also be a scalar condensate of fermionic particles. It is also possible to have Higgs inflation, where the inflaton is the Higgs field itself. Yet another possibility is an extension of general relativity. For example, in $f(R)$ gravity, one replaces the term R in the Einstein-Hilbert action with a general function $f(R)$. Often such theories are conformally equivalent to general relativity with an additional scalar field, which can serve as the inflaton. For example, in Starobinsky inflation we have

$$f(R) = R + \frac{R^2}{6M^2}.$$

Axion-like particles such as the QCD axion could serve as the inflaton. More generally, string theory also provides many inflaton candidates.

2.3 Models of Inflation

First, we cover the historical development of inflation.

- In Guth's original model of inflation, now called "old inflation", the inflaton begins trapped in a false vacuum, during which its energy density drives inflation. Inflation ends when the field tunnels out of the vacuum, leading to bubble nucleation. Reheating occurs when the bubbles collide. However, bubble collisions are too rare in this model, and when they do occur they produce inhomogeneities that are too strong. This is the "graceful exit problem".
- Old inflation was motivated by phase transitions in the $SU(5)$ GUT. Generically, such GUTs have a Higgs field which has a global potential minimum at $\varphi = 0$ for high temperatures. For lower temperatures, this minimum becomes only a local minimum, leading to a first order phase transition.
- In "new inflation", which is the framework we've used above, inflation occurs during a phase of slow rolling, without the need for bubble nucleation. For example, it can occur if we start from a very flat maximum neat $\varphi = 0$.
- In both cases, the universe was regarded as existing in thermal equilibrium throughout inflation, beginning with a standard Big Bang. In the case of new inflation, the inflaton ends up at the top of a maximum because of symmetry restoration due to high temperature.
- The first models that broke away from this assumption were chaotic inflation models, which simply used polynomial potentials. For inflation to work, one requires very high initial field values, which are justified by positing "chaotic" initial conditions with Planckian energy densities. The regions with high field values inflate more, extracting a homogeneous universe from inhomogeneous initial conditions. All models of inflation are viewed this way today; more detailed investigation of the initial conditions is done by "quantum cosmologists".

- Today, "chaotic inflation" and "new inflation" are typically used to describe the shape of the potential; they are pure polynomials or have an extended flat region, respectively. The graceful exit problem is solved by the theory of reheating.

- Many models of inflation generically predict eternal inflation. To understand eternal inflation, we consider the fate of one Hubble patch during an e-fold of expansion, where the volume increases by a factor of $e^3 \propto 20$. Quantum fluctuations mean that some of the new patches have a higher value of the inflaton field than the original patch. Hence inflation can continue in these patches, and extends infinitely into the future. The universe acquires a 'fractal' structure.

- To be more quantitative, we consider our explicit model above. During a Hubble time,
$$\Delta\phi \sim \frac{\dot{\phi}}{H} \sim \frac{V'}{H^2} \sim M_{\text{pl}}^2 \frac{V'}{V} \sim \frac{M_{\text{pl}}^2}{\phi}.$$
On the other hand, the scale of quantum fluctuations is
$$|\delta\phi| \sim \frac{H}{2\pi} \sim \frac{m\phi}{M_{\text{pl}}}.$$
Therefore, the quantum fluctuations dominate when
$$\phi > M_{\text{pl}} \sqrt{\frac{M_{\text{pl}}}{m}}.$$
Hence we can avoid eternal inflation if we simply take m small and postulate an initial condition as above. However, if we take chaotic initial conditions with Planckian energy densities, the inequality above must be satisfied, leading to eternal inflation. (This is compatible with observations, because observable quantities only come from the last 60 e-folds of inflation.) Depending on taste, one can ensure eternal inflation to facilitate anthropics or forbid it.

- Eternal inflation comes with issues in calculating probabilities, called the measure problem. For example, naively $1/2$ of the integers are even, but if one orders the integers in an alternate way, we can have any fraction of the first N integers be even. The same applies for the 'pocket universes' of eternal inflation; since they are spacelike separated, the time ordering is arbitrary.

- If one simply fixes a naive time ordering, e.g. in synchronous gauge, one runs into the youngness paradox: almost all universes are very young. For example, if one conditions on our existence, with a uniform prior on all universes created to date, then we should have evolved "as quickly as possible", and there can be no older alien civilizations. One can avoid this by using a prior that weights on volume, or various others, but there is no canonical prescription.

Next, we consider more recent news.

- The CMB was mapped by COBE in the 1990s. COBE confirmed that the CMB spectrum was very close to a blackbody spectrum, and detected small anisotropies.

- In the 2000s, WMAP measured the CMB more precisely, confirming more predictions of inflation. In particular, the CMB perturbations are consistent with being adiabatic (vs. isocurvature) and gaussian, with an approximately scale-invariant power spectrum. Curvature fluctuations, which are measured by the multipole moments of the CMB's fluctuations, were also found to be consistent with the inflationary paradigm. Finally, WMAP measured the parameters in the standard ΛCDM model and found the universe to be nearly flat, consistent with inflation.

- Two parameters that describe the fluctuations in the CMB are the tensor-to-scalar ratio r and the primordial tilt n_s, which quantifies scale-invariance. WMAP found that $r < 0.6$ and $n_s \approx 1$. Inflation generically predicts $r \neq 0$ and $n_s \approx 1$.
- A nonzero value for r indicates the presence of primordial gravitational waves, thought of as a unique signature of inflation. In 2012, the telescope BICEP2 reported a measurement of $r \approx 0.2$, disfavoring $r = 0$ at 7σ, but the significance was removed once galactic dust was accounted for.
- Note that there also exist more complicated inflationary models where the perturbations are not adiabatic, not gaussian, or not nearly scale-invariant. In fact, there even exist inflationary models that don't predict flatness. We focus only on the simpler models, which don't have these features. The value for r varies strongly between different models, even simple ones.
- The Planck satellite measured the CMB in the 2010s, and found parameters that were uncomfortable for the simplest models in inflation, involving $V \propto \phi^p$ for a power p. Instead, concave potentials such as those in new inflation are favored. Simple models such as Starobinsky R^2 inflation, historically the first model of inflation, and Higgs inflation, also fit well.
- There are also alternatives to inflation, such as cyclic cosmology, which involve a 'big bounce'. However, such theories run into the 'singularity problem'. They cannot predict what will happen at a cosmological singularity, where nonperturbative quantum gravity can play a role.

3 Thermal History

3.1 The Hot Big Bang

We begin with a basic overview of the first three minutes of the universe.

- Let Γ be the rate of interaction for a particle, and define timescales $t_c = 1/\Gamma$ and $t_H = 1/H$. Then when $t_c \ll t_H$, local thermal equilibrium is reached before the effect of the expansion becomes relevant, and when $t_c \sim t_H$, the particle decouples from the thermal bath. Different particle species may decouple at different times.

- The contribution to Γ due to scattering off another particle species is

$$\Gamma = n\sigma v$$

where n is the number density of that species, σ is the scattering cross section, and v is the relative velocity.

- We focus on the case $T \gtrsim 100\,\mathrm{GeV}$, where all known particles are ultrarelativistic. Then $v = 1$, and by dimensional analysis we have $n \sim T^3$ for every species, since the masses play no role. If we assume the scattering is primarily by tree-level exchange of a massless gauge boson, we have $\sigma \propto \alpha^2$, and dimensional analysis gives $\sigma \sim \alpha^2/T^2$. Therefore

$$\Gamma \sim \alpha^2 T.$$

- By using dimensional analysis again, we have

$$H \sim \frac{\sqrt{\rho}}{M_{\mathrm{pl}}} \sim \frac{T^2}{M_{\mathrm{pl}}}, \quad \frac{\Gamma}{H} \sim \frac{\alpha^2 M_{\mathrm{pl}}}{T} \sim \frac{10^{16}\,\mathrm{GeV}}{T}$$

where we used $\alpha \gtrsim 0.01$, valid for charged or strongly interacting particles. Hence when $100\,\mathrm{GeV} \ll T \ll 10^{16}\,\mathrm{GeV}$, all particles are ultrarelativistic and in local thermal equilibrium.

- When a particle species is in equilibrium, it obeys the Fermi-Dirac or Bose-Einstein distribution,

$$f(E) = \frac{1}{e^{E/T} \pm 1}.$$

In particular, once the particles become nonrelativistic, the density falls exponentially since $f \sim e^{-m/T}$. Hence we can approximate the energy density by summing only over relativistic particle species, giving

$$\rho_r = \frac{\pi^2}{30} g_*(T) T^4$$

where $g_*(T)$ is the number of relativistic degrees of freedom. It ranges from 106.75 at early times to 3.38 at the present day, where only photons and perhaps neutrinos are still relativistic.

- If equilibrium persisted forever, then the universe would be mostly photons, since massive particles species would be exponentially suppressed. Consider a species of massive particle with an equal number of antiparticles. As the particles annihilate, at some point they ‘freeze out’, as the annihilation rate falls below H, leaving behind a ‘relic density’. The annihilation process is thus not in chemical equilibrium.

- This is distinct from the question of 'decoupling', when a particle species is no longer in thermal equilibrium with the radiation bath. However, for most massive species decoupling and freeze out happen at around the same time, at $m \sim T$. Exceptions occur for particles that don't interact strongly or electromagnetically.
- For example, neutrinos decouple while they are still relativistic, because they only interact weakly. Electroweak symmetry breaking occurs at $T \leq 100\,\text{GeV}$, below which weak cross sections are suppressed as

$$\sigma \sim G_F^2 T^2, \quad G_F \sim 10^{-5}\,\text{GeV}^{-2}.$$

and hence we have

$$\frac{\Gamma}{H} \sim \left(\frac{T}{1\,\text{MeV}}\right)^3.$$

Thus the neutrinos decouple at around $1\,\text{MeV}$.

- More speculatively, gravitons have $\sigma \sim G_N^2 T^2$ and hence decouple when $T \sim M_{\text{pl}}$, leaving a graviton background with temperature approximately $1\,\text{K}$ today.

The events in the history of the universe are summarized in the table below.

Event	time t	redshift z	temperature T
Inflation	10^{-34} s (?)	–	–
Baryogenesis	?	?	?
EW phase transition	20 ps	10^{15}	100 GeV
QCD phase transition	20 μs	10^{12}	150 MeV
Dark matter freeze-out	?	?	?
Neutrino decoupling	1 s	6×10^9	1 MeV
Electron-positron annihilation	6 s	2×10^9	500 keV
Big Bang nucleosynthesis	3 min	4×10^8	100 keV
Matter-radiation equality	60 kyr	3400	0.75 eV
Recombination	260–380 kyr	1100–1400	0.26–0.33 eV
Photon decoupling	380 kyr	1000–1200	0.23–0.28 eV
Reionization	100–400 Myr	11–30	2.6–7.0 meV
Dark energy-matter equality	9 Gyr	0.4	0.33 meV
Present	13.8 Gyr	0	0.24 meV

- The first event is baryogenesis, which seeks to explain why the universe has net baryon number. Note that one could simply postulate an initial baryon asymmetry, but that wouldn't be satisfying. We will black box the process of baryogenesis since not much is known about it.
- All known particle species are in thermal equilibrium until the electroweak phase transition. Particles acquire masses from the Higgs mechanism at this point.
- If dark matter is a WIMP with a mass around the electroweak scale, then around this time, dark matter freezes out. However, it doesn't decouple from the thermal bath until around $T \sim 1\,\mathrm{MeV}$, by the same argument as for neutrinos. This has observational consequences because it affects the temperature of the dark matter, which falls as $1/a^2$ while decoupled, but only as roughly $1/a$ when in equilibrium with a radiation-dominated thermal bath.
- At a temperature of $150\,\mathrm{MeV}$, the QCD phase transition occurs; the quark gluon plasma hadronizes into baryons and mesons.
- The next event is neutrino and WIMP decoupling, which occur around $T \sim 1\,\mathrm{MeV}$. Shortly afterwards, electrons and positrons annihilate. This energy heats up the photons but not the neutrinos, since they have decoupled, causing their temperatures to be different today.
- After about three minutes, at temperature $T \sim 100\,\mathrm{keV}$, Big Bang nucleosynthesis (BBN) occurs, forming primarily deuterium, helium, and lithium. This is later than one would expect, given nucleon binding energies of about $1\,\mathrm{MeV}$, because photons greatly outnumber nuclei, so photons in the high-energy tail tend to break them apart.
- Not much happens until recombination, when neutral hydrogen forms by the reaction $e^- + p^+ \to H + \gamma$, with the reverse reaction energetically disfavored. This again happens somewhat later than one would expect because of the relatively large number of photons.
- Since photons mostly interact by Thomson scattering $e^- + \gamma \to e^- + \gamma$ at this point, photons decouple shortly afterward and "free stream" through the universe, forming the CMB. Note that we say the photons decoupled from matter, rather than vice versa, because by this point the universe has become matter-dominated.
- Afterward, there is a period called the "cosmic dark ages", named because the radiation background no longer contains visible light, and stars haven't formed yet. When stars do form, they cause hydrogen gas in space to reionize; black holes also contribute through X-ray emission.

It should be noted that many elements of the story above are quite speculative; the earliest element with good direct support is BBN, which strongly constrains many alternative models.

3.2 Equilibrium

We now consider the equilibrium aspects of the story above, starting by reviewing basic equilibrium statistical mechanics.

- For a gas in a box of volume V, the density of states is g/h^3 in phase space, where g is the number of internal degrees of freedom. In natural units, this is $g/(2\pi)^3$.

- By homogeneity, the distribution in position space is uniform, leaving a distribution in momentum space; by isotropy it only depends on the magnitude of the momentum. If $f(p)$ is this distribution function, then
$$n = \frac{g}{(2\pi)^3} \int d\mathbf{p}\, f(p).$$

- Ignoring interactions between the particles,
$$\rho = \frac{g}{(2\pi)^3} \int d\mathbf{p}\, f(p)E(p), \quad E(p) = \sqrt{p^2+m^2}.$$
Finally, the pressure is
$$p = \frac{g}{(2\pi)^3} \int d\mathbf{p}\, f(p) \frac{p^2}{3E}.$$
The factor of $p^2/3E$ is the usual $\langle p_x v_x \rangle = \langle \mathbf{p} \cdot \mathbf{v} \rangle / 3$ factor from kinetic theory, with $\mathbf{p} = E\mathbf{v}$.

- The distribution function is
$$f(p) = \frac{1}{e^{(E(p)-\mu)/T} \pm 1}$$
with the plus sign for fermions and the minus sign for bosons. The chemical potential μ changes as the universe expands; its evolution may be fixed by the continuity equations for energy and entropy.

- If species are in chemical equilibrium, then the chemical potential balances in every reaction. For example, if we have the reaction $1 + 2 \leftrightarrow 3 + 4$, then
$$\mu_1 + \mu_2 = \mu_3 + \mu_4.$$
At high temperatures, photons may be produced by, e.g. double Compton scattering
$$e^- + \gamma \leftrightarrow e^- + \gamma + \gamma$$
which sets $\mu_\gamma = 0$. In particular, by considering the process
$$X + \overline{X} \leftrightarrow \gamma + \gamma$$
we must have $\mu_X = -\mu_{\overline{X}}$. Note that chemical equilibrium is distinct from thermal equilibrium, which is when the species are at the same temperature.

Now we perform some explicit calculations.

- At early times, the chemical potential of all species is approximately zero. Neglecting it,
$$n = \frac{g}{2\pi^2} \int_0^\infty dp\, \frac{p^2}{\exp(\sqrt{p^2+m^2}/T) \pm 1}, \quad \rho = \frac{g}{2\pi^2} \int_0^\infty dp\, \frac{p^2\sqrt{p^2+m^2}}{\exp(\sqrt{p^2+m^2}/T) \pm 1}.$$
Defining $x = m/T$ and $\xi = p/T$, we find
$$n = \frac{g}{2\pi^2} T^3 I_\pm(x), \quad \rho = \frac{g}{2\pi^2} T^4 J_\pm(x)$$
defined in terms of the dimensionless integrals
$$I_\pm(x) = \int_0^\infty d\xi\, \frac{\xi^2}{\exp(\sqrt{\xi^2+x^2}) \pm 1}, \quad J_\pm(x) = \int_0^\infty d\xi\, \frac{\xi^2\sqrt{\xi^2+x^2}}{\exp(\sqrt{\xi^2+x^2}) \pm 1}.$$

- To make progress, we use the standard integrals

$$\int_0^\infty d\xi\, \frac{\xi^n}{e^\xi - 1} = \zeta(n+1)\Gamma(n+1), \qquad \int_0^\infty d\xi\, \xi^n e^{-\xi^2} = \frac{1}{2}\Gamma((n+1)/2)$$

which are derived by geometric series and integration by parts respectively.

- In the relativistic limit $x \to 0$, we have $I_-(0) = 2\zeta(3)$. As for the plus sign, note that

$$\frac{1}{e^\xi + 1} = \frac{1}{e^\xi - 1} - \frac{2}{e^{2\xi} - 1}, \qquad I_+(0) = I_-(0) - 2\left(\frac{1}{2}\right)^3 I_-(0) = \frac{3}{4} I_-(0).$$

Hence we have

$$n = \frac{\zeta(3)}{\pi^2} g T^3 \begin{cases} 1 & \text{bosons}, \\ 3/4 & \text{fermions}. \end{cases}$$

A very similar computation yields

$$\rho = \frac{\pi^2}{30} g T^4 \begin{cases} 1 & \text{bosons}, \\ 7/8 & \text{fermions} \end{cases}$$

where we used $\zeta(4) = \pi^4/90$. Doing the same computation for p gives the usual relation for a relativistic gas, $p = \rho/3$.

- We may also account for a chemical potential in the ultrarelativistic case. This doesn't make sense for massless bosons, since either n or $\overline{n}$ would diverge, but for massless or ultrarelativistic fermions,

$$n - \overline{n} = \frac{g}{2\pi^2} \int_0^\infty dp\, p^2 \left(\frac{1}{e^{(p-\mu)/T} + 1} - \frac{1}{e^{(p+\mu)/T} + 1} \right) = \frac{gT^3}{6\pi^2} \left(\pi^2 \left(\frac{\mu}{T}\right) + \left(\frac{\mu}{T}\right)^3 \right)$$

which may be shown by shifting $p \to p + \mu$ in the first integral and $p \to p - \mu$ in the second, then performing some cancellations and a contour integral.

- We can also work in the nonrelativistic limit $x \gg 1$, where for both fermions and bosons

$$I_\pm(x) \approx \int_0^\infty d\xi\, \frac{\xi^2}{e^{\sqrt{\xi^2 + x^2}}}.$$

Taylor expanding the denominator gives

$$I_\pm(x) \approx \int_0^\infty d\xi\, \frac{\xi^2}{e^{x + \xi^2/2x}} = (2x)^{3/2} e^{-x} \int_0^\infty d\xi\, \xi^2 e^{-\xi^2} = \sqrt{\frac{\pi}{2}} x^{3/2} e^{-x}$$

where we used our second standard integral, and $\Gamma(3/2) = \sqrt{\pi}/2$. Thus we have

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

This is of course just what we would expect from the Maxwell-Boltzmann distribution.

- As for the energy density, using $E(p) = \sqrt{p^2 + m^2} \approx mn + 3nT/2$ gives

$$\rho \approx mn + \frac{3}{2} nT.$$

  One may also compute $P = nT \ll \rho$, i.e. the ideal gas law, as expected.

- It is also straightforward to restore finite μ, which gives an extra prefactor,

$$n = g \left(\frac{mT}{2\pi} \right)^{3/2} e^{-(m-\mu)/T}, \quad n - \overline{n} = 2g \left(\frac{mT}{2\pi} \right)^{3/2} e^{-m/T} \sinh(\mu/T).$$

Finally, we consider the effective number of relativistic species.

- Let T be the temperature of the photon gas. Using our results, we may compute the number of relativistic degrees of freedom $g_*(T)$. In general we have

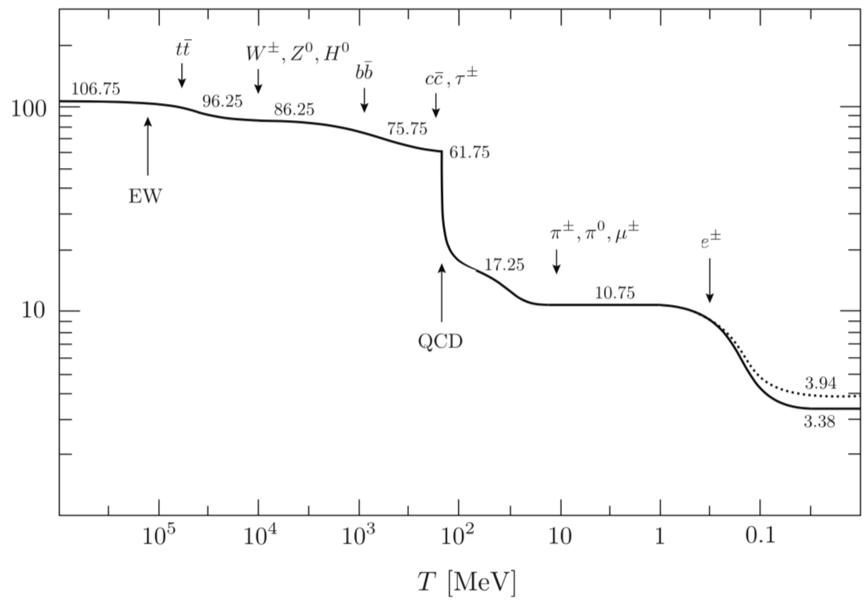
$$g_*(T) = \sum_i g_i \left(\frac{T_i}{T} \right)^4 \begin{cases} 1 & \text{bosons,} \\ 7/8 & \text{fermions} \end{cases}$$

  where the sum is over species with $m < T_i$, and T_i is the temperature of the species. For species in thermal equilibrium with the photons, which will be most of them, we simply have $T_i = T$.

- At high temperatures, the SM degrees of freedom are counted as follows.
  - Quarks have 2 spins and 3 colors; counting antiquarks gives another factor of 2. Hence they contribute $6 \times 12 = 72$ degrees of freedom.
  - Each massless gauge boson has two polarizations, so the gluons contribute 16.
  - Similarly, the photon contributes 2, while the $W^\pm$ and Z bosons contribute 9.
  - The Higgs boson is a real scalar and contributes 1.
  - The charged leptons contribute 4 each, from 2 spins and antiparticles. The neutrinos contribute only 2 each, since all neutrinos have negative helicity and all antineutrinos have positive helicity.

  This gives a total of 28 bosonic degrees of freedom and 90 fermionic degrees of freedom, for $g_* = 106.75$ for $T \geq 100\,\mathrm{GeV}$.

- The evolution of $g_*(T)$ in the early universe is shown below.



Following the electroweak phase transition, the top quark is the first to annihilate, followed by the weak bosons, the Higgs, the bottom quark, the charm quark, and the tau.

- After the QCD phase transition, the quarks condense into baryons and mesons, but only the pions ($\pi^\pm, \pi^0$) are relativistic, contributing 3 degrees of freedom. Hence we are left with pions, electrons, muons, neutrinos, and photons. The pions and muons annihilate next; then the neutrinos decouple and the electrons annihilate. The dotted line shows the effective number of degrees in entropy $g_{*S}(T)$, explained below.

- Since these particles interact by the strong and electromagnetic forces, this annihilation process is quite efficient. Almost all pions and muons annihilate; those that don't decay later, so we don't see a relic density. The remaining matter is in the form of protons and neutrons.

- The annihilation of antibaryons is especially efficient, due to the net baryon number created in baryogenesis, as the amount of baryons they have to annihilate against approaches a constant rather than zero.

- A useful expression for the Hubble constant during this period is

$$H^2 M_{\text{pl}}^2 = \frac{\pi^2}{90} g_*(T) T^4$$

  which is valid during radiation domination. The factor of M_{pl} leads to a large hierarchy between H and T, in contrast to $H \sim T$ during inflation.

We now discuss the conservation of entropy.

- A particle species in thermal equilibrium with $\mu = 0$ has constant entropy as the universe expands. To see this, we note that

$$\frac{\partial p}{\partial T} \propto \int d\mathbf{p}\, \frac{df}{dT} \frac{|\mathbf{p}|^2}{E} = -\frac{1}{T} \int d\mathbf{p}\, \frac{df}{dE} |\mathbf{p}|^2 \propto -\frac{1}{T} \int dE\, \frac{d|\mathbf{p}|}{dE} \frac{df}{dE} |\mathbf{p}|^4 = -\frac{1}{T} \int dE\, \frac{df}{dE} (|\mathbf{p}|^3 E)$$

where we used the fact that the distribution function depends only on E/T, then performed the angular integral and switched variables to E. Integrating by parts gives

$$\frac{\partial p}{\partial T} \propto \frac{1}{T} \int dE\, f(E) \left(|\mathbf{p}|^3 + 3|\mathbf{p}|E^2\right).$$

Converting this back to a $d\mathbf{p}$ integral and restoring constants yields the simple identity

$$\frac{\partial p}{\partial T} = \frac{\rho + p}{T}.$$

- Next, we use the first law of thermodynamics and the identity $U = \rho V$ for

$$dS = \frac{1}{T}(dU + p\, dV) = \frac{1}{T}\left(d((\rho+p)V) - V\, dp\right) = d\left(\frac{\rho + p}{T} V\right)$$

where the last equality follows from our identity above. Hence we have

$$s = \frac{\rho + p}{T}$$

where $s = S/V$ is the entropy density, where the constant of integration is fixed to be zero by the third law of thermodynamics.

- To show the entropy is conserved, note that the continuity equation is

$$\dot{\rho} + 3H(\rho + p) = 0, \quad d\rho + \frac{dV}{V}(\rho + p) = 0.$$

Then the differential of entropy is

$$dS = \frac{1}{T}((\rho + p)dV + V\, d\rho) = 0$$

as desired. This result generalizes the fact that a photon gas expands adiabatically in an expanding universe to nonrelativistic gases. It is perfectly intuitive, as there is no heat transfer possible from outside the system, since our system is the entire universe.

- We define the effective number of degrees of freedom in entropy by

$$s = \frac{2\pi^2}{45} g_{*S}(T) T^3.$$

By similar reasoning to the above, we have

$$g_{*S}(T) = \sum_i g_i \left(\frac{T_i}{T}\right)^3 \begin{cases} 1 & \text{bosons,} \\ 7/8 & \text{fermions} \end{cases}$$

where the sum is over all relativistic species; the only difference is that there is a cubic dependence on T rather than a quartic dependence.

- The conservation of entropy is a useful tool when species annihilate, because

$$S \propto g_{*S} T^3 a^3 = \text{constant}$$

which implies that $T \propto g_{*S}^{-1/3} a^{-1}$. This reproduces the usual $1/a$ falloff, but when a species annihilates, the temperature increases, as the entropy of the annihilating species is transferred to other relativistic species.

- Technically, the temperature actually monotonically decreases, since annihilation is a gradual process that occurs as a changes by an $O(1)$ factor. The temperature just declines slightly slower during this period.

- Note that entropy is not conserved in nonequilibrium processes. However, such processes are confined to the baryons, which are greatly outnumbered by photons, so the change is negligible.

- For a given species, we define the number of particles per comoving volume as
$$N_i = \frac{n_i}{s}.$$
This works because $N_i = n_i a^3 / s a^3$, where the numerator is the number of particles in some comoving volume and the denominator is the conserved entropy in that comoving volume. Hence dividing by s just rescales this volume, and eliminates the explicit dependence on a.

We now apply this to the case of neutrino decoupling.

- When neutrinos decouple, they are ultrarelativistic. After decoupling, they don't interact with anything, including each other, but they maintain a *relativistic* thermal distribution $f(p, T)$ with $T \propto a^{-1}$ since all neutrinos are redshifted equally. This remains true even when the "temperature" drops below the neutrino mass.

- Similarly, a species that decouples when it is nonrelativistic maintains a thermal distribution with $T \propto a^{-2}$ for all future times.

- Neutrinos are coupled to the thermal bath by weak interactions, such as
$$\nu_e + \overline{\nu}_e \leftrightarrow e^+ + e^-, \quad e^- + \overline{\nu}_e \leftrightarrow e^- + \overline{\nu}_e.$$
As we've seen, neutrinos decouple when $T \sim 1\,\mathrm{MeV}$. However, despite this decoupling, they maintain roughly the same temperature as the photons since both fall as $1/a$, until electron-positron annihilation warms up the photon bath. Without counting the neutrinos, the effective number of degrees of freedom in entropy is
$$g'_{*S} = \begin{cases} 2 + \frac{7}{8} \times 4 = \frac{11}{2} & T > m_e, \\ 2 & T < m_e. \end{cases}$$
Then the temperature of the photons increases by a factor of $(11/4)^{1/3}$.

- This ratio holds until the present day. One might imagine that the photon temperature should fall as $1/a^2$ rather than $1/a$ in the period of matter-domination preceding photon decoupling. However, in this period the photon energy is still much greater than the total *kinetic* energy of matter, even though it's much less than the rest energy, so the temperature still falls as $1/a$.

- The results above are only approximate, since neutrino decoupling is a gradual process; in reality some of the energy 'leaks' to the neutrinos. As a result, today we have
$$g_* = 2 + \frac{7}{8} \times 2N_{\text{eff}} \left(\frac{4}{11}\right)^{4/3} = 3.36, \quad g_{*S} = 2 + \frac{7}{8} \times 2N_{\text{eff}} \left(\frac{4}{11}\right) = 3.94$$

where instantaneous decoupling would give $N_{\mathrm{eff}} = 3$, but instead

$$N_{\mathrm{eff}} = \begin{cases} 3.36 \pm 0.34 & \text{experiment,} \\ 3.046 & \text{theory.} \end{cases}$$

Future refinements of the measurement of N_{eff} may be able to test for new physics.

- The number density of neutrinos is

$$n_\nu = \frac{3}{4} N_{\mathrm{eff}} \times \frac{4}{11} n_\gamma.$$

The energy density depends on the neutrino masses. It used to be believed that neutrinos were massless, in which case the energy density is closely related to the CMB energy density,

$$\rho_\nu = \frac{7}{8} N_{\mathrm{eff}} \left(\frac{4}{11}\right)^{4/3} \rho_\gamma.$$

However, experiments indicate that neutrinos have mass, with

$$\sum_i m_{\nu,i} > 0.06\,\mathrm{eV}.$$

- The temperatures of the CMB and "CνB" are

$$T_0 = 2.73\,\mathrm{K} = 0.24\,\mathrm{meV}, \quad T_\nu = 1.95\,\mathrm{K} = 0.17\,\mathrm{meV}.$$

This holds regardless of the neutrino masses, as the distribution remains formally relativistic.

- If the neutrino masses were too large, then they would overclose the universe by themselves, just by virtue of their rest energy. This yields the constraint

$$\sum_i m_{\nu,i} < 15\,\mathrm{eV}.$$

In fact, more stringent experimental tests show that

$$\sum_i m_{\nu,i} < 0.3\,\mathrm{eV}$$

which indicates that while neutrinos likely have more energy than the photons, they still are a small contribution overall, $\Omega_\nu < 0.01$.

- Given the results above, it's likely that no neutrino species remain relativistic today. However, it's logically possible that one neutrino species is massless.

3.3 The Boltzmann Equation

Next, we introduce the Boltzmann equation for nonequilibrium processes.

- In the absence of interactions, the number density of a particle species i evolves as

$$\dot{n_i} + 3n_i\frac{\dot{a}}{a} = \frac{1}{a^3}\frac{d(n_i a^3)}{dt} = 0$$

since the particles simply dilute with the expansion. The Boltzmann equation is

$$\frac{1}{a^3}\frac{d(n_i a^3)}{dt} = C_i[\{n_j\}]$$

where the collision term on the right-hand side accounts for all reactions. In general, both sides would have full phase space distributions, but for our calculations this will suffice.

- In general, reactions involving three or more particles are very unlikely, so we can restrict to decays or two-particle scatterings and annihilations. All reactions we study will be of the form

$$1 + 2 \leftrightarrow 3 + 4.$$

In this case the Boltzmann equation for species 1 is

$$\frac{1}{a^3}\frac{d(n_1 a^3)}{dt} = -\alpha n_1 n_2 + \beta n_3 n_4.$$

- The coefficients are thermally averaged cross sections, $\alpha = \langle \sigma v \rangle$, and β may be related by

$$\beta = \left(\frac{n_1 n_2}{n_3 n_4}\right)_{\text{eq}} \alpha$$

by detailed balance, where we use the equilibrium number densities calculated in the previous section. Hence we have

$$\frac{1}{a^3}\frac{d(n_1 a^3)}{dt} = -\langle \sigma v \rangle \left(n_1 n_2 - \left(\frac{n_1 n_2}{n_3 n_4}\right)_{\text{eq}} n_3 n_4 \right).$$

- It is useful to rewrite this in terms of $N_i = n_i/s$, for

$$\frac{d\log N_1}{d\log a} = -\frac{\Gamma_1}{H}\left(1 - \left(\frac{N_1 N_2}{N_3 N_4}\right)_{\text{eq}} \frac{N_3 N_4}{N_1 N_2}\right), \quad \Gamma_1 = n_2 \langle \sigma v \rangle.$$

The quantity in parentheses expresses the deviation from equilibrium. Then when $\Gamma_1 \gg H$, N_1 quickly approaches the equilibrium value, while for $\Gamma_1 \ll H$ we get a constant value of N_1, i.e. a relic density.

As a first example, we consider the freeze-out production of WIMP dark matter.

- We consider a reaction of the form

$$X + \overline{X} \leftrightarrow \ell + \bar{\ell}$$

where ℓ is a light (essentially massless) tightly coupled to the SM plasma, which always has its equilibrium density. We also assume $n_X = n_{\overline{X}}$. Then the Boltzmann equation becomes

$$\frac{dN_X}{dt} = -s\langle \sigma v \rangle \left(N_X^2 - (N_X^{\text{eq}})^2\right).$$

- It is most convenient to express the evolution in terms of $x = M_X/T$, so the interesting dynamics occurs near $x \sim 1$. To perform the change of variable, note that

$$\frac{dx}{dt} = -\frac{1}{T}\frac{dT}{dt}x = Hx$$

where we used $T \propto a^{-1}$ during radiation domination. Furthermore, during radiation domination

$$H = \frac{H(M_X)}{x^2}.$$

Plugging these into the Boltzmann equation, we have

$$\frac{dN_X}{dx} = -\frac{\lambda}{x^2}\left(N_X^2 - (N_X^{\text{eq}})^2\right), \quad \lambda = \frac{2\pi^2}{45} g_{*S} \frac{M_X^3 \langle \sigma v \rangle}{H(M_X)}.$$

During the process of freeze-out, λ is approximately constant.

- The differential equation above is the Riccati equation and has no closed-form solution. However, numerically we find departure from equilibrium occurs near $x \sim 10$. At this point N_X^{eq} is very small, so in the subsequent solution

$$\frac{dN_X}{dt} = -\frac{\lambda}{x^2}N_X^2$$

which integrates to

$$\frac{1}{N_X^\infty} - \frac{1}{N_X(x_f)} = \frac{\lambda}{x_f}.$$

Since typically $N_X(x_f) \gg N_X^\infty$, we have the simple approximation

$$N_X^\infty \approx \frac{x_f}{\lambda} \sim \frac{10}{\lambda}.$$

- The remaining dark matter density today is

$$\rho_{X,0} = M_X N_X^\infty s_0.$$

Substituting our result for N_X^∞ and $s_0 = s(T_0)$, we have

$$\rho_{X,0} = \frac{H(M_X)}{M_X^2} \frac{x_f}{\langle \sigma v \rangle} \frac{g_{*S}(T_0)}{g_{*S}(M_X)} T_0^3.$$

During radiation domination, the Hubble constant is

$$H^2 M_{\text{pl}}^2 = \frac{\pi^2}{90} g_*(T) T^4.$$

Plugging this in, using $\rho_{c,0} = 3M_{\text{pl}}^2 H_0^2$, and plugging in the currently measured values of H_0, T_0, and $g_{*S}(T_0) = 3.91$, we find

$$\Omega_X h^2 \sim 0.1 \frac{x_f}{10} \left(\frac{10}{g_*(M_X)}\right) \frac{10^{-8}\,\text{GeV}^{-2}}{\langle \sigma v \rangle}.$$

- This accounts for the observed dark matter density if
$$\sqrt{\langle \sigma v \rangle} \sim 10^{-4}\,\mathrm{GeV}^{-1} \sim 0.1\sqrt{G_F}$$
which is what we would expect for a weak-scale WIMP. This result is the WIMP miracle.

Next, we consider recombination and photon decoupling.

- In this case, we are concerned with the reaction
$$e^- + p^+ \leftrightarrow H + \gamma$$
which is in equilibrium for $T > 1\,\mathrm{eV}$. We begin with equilibrium considerations. Since all particles besides the photon are nonrelativistic,
$$n_i^{\mathrm{eq}} = g_i \left(\frac{m_i T}{2\pi}\right)^{3/2} e^{(\mu_i - m_i)/T}, \quad \mu_p + \mu_e = \mu_H.$$
- To remove the dependence on the chemical potential, we consider the ratio
$$\left(\frac{n_H}{n_e n_p}\right)_{\mathrm{eq}} = \frac{g_H}{g_e g_p}\left(\frac{m_H}{m_e m_p}\frac{2\pi}{T}\right)^{3/2} e^{(m_p+m_e-m_H)/T}.$$
The first factor is $4/(2\times 2) = 1$. The exponential factor is $e^{B_H/T}$ where $B_H = 13.6\,\mathrm{eV}$ is the binding energy of hydrogen. Since the universe is charge neutral, we have $n_e = n_p$, giving
$$\left(\frac{n_H}{n_e^2}\right)_{\mathrm{eq}} = \left(\frac{2\pi}{m_e T}\right)^{3/2} e^{B_H/T}$$
where we used $m_H \approx m_p$.
- Next, we define the free electron fraction
$$X_e = \frac{n_e}{n_b}$$
where n_b is the baryon density. We also define the baryon-to-photon ratio
$$\eta = \frac{n_b}{n_\gamma}, \quad n_\gamma = \frac{2\zeta(3)}{\pi^2}T^3, \quad \eta = 5.5\times 10^{-10}(\Omega_b h^2/0.020).$$
The total baryon density is approximately $n_b \approx n_p + n_H = n_e + n_H$. Note that the baryon-to-photon ratio is constant after photon decoupling, since both n_b and n_γ dilute as $1/a^3$. Hence the quoted value above is measured from the CMB temperature and baryon density today. Furthermore, it is constant before photon decoupling because $n_b \propto 1/a^3 \propto T^3$ during radiation domination, and also during the short window of matter domination before photon decoupling because photons still hold most of the kinetic energy.
- Since we have
$$\frac{1-X_e}{X_e^2} = \frac{n_H}{n_e^2} n_b$$
the equilibrium value of the free electron fraction obeys
$$\left(\frac{1-X_e}{X_e^2}\right)_{\mathrm{eq}} = \frac{2\zeta(3)}{\pi^2}\,\eta\left(\frac{2\pi T}{m_e}\right)^{3/2} e^{B_H/T}.$$
This is the Saha equation, which as expected predicts an exponential falloff of X_e at low temperature. We expect it should be reasonably accurate before X_e gets too small.

- We conventionally define the recombination temperature as the temperature where $X_e = 0.1$. Plugging this in and solving for T, we find

$$T_{\text{rec}} \approx 0.3\,\text{eV} \approx 3600\,\text{K}, \quad z_{\text{rec}} \approx 1320$$

which is, as expected earlier, much less than $B_H = 13.6\,\text{eV}$ because η is large.

- Now we consider photon decoupling. At this stage, photons are most strongly coupled to the thermal plasma by Thomson scattering,

$$e^- + \gamma \leftrightarrow e^- + \gamma, \quad \Gamma_\gamma = n_e \sigma_T = n_b X_e \sigma_T = \eta n_\gamma X_e \sigma_T$$

where $\sigma_T \approx 2 \times 10^{-3}\,\text{MeV}^{-2}$ is the Thomson cross section.

- We define photon decoupling to occur when $\Gamma_\gamma \sim H$. To evaluate this condition, note that Γ_γ may be evaluated in terms of quantities known at recombination, and since this time period is matter dominated,

$$H(T) \approx H_0 \sqrt{\Omega_m} \left(\frac{T}{T_0}\right)^{3/2}.$$

Putting this all together and using the Saha equation, we find

$$T_{\text{dec}} \sim 0.27\,\text{eV}, \quad z_{\text{dec}} \sim 1100$$

by which point we have $X_e \sim 0.01$. At this point, the CMB is formed.

- Finally, there is a relic density of free electrons and protons, which may be computed with the Boltzmann equation applied to the reaction $e^- + p^+ \leftrightarrow H + \gamma$. This is fairly similar to the computation for the WIMP relic density.

Note. A more accurate analysis of recombination. In reality, the Saha equation breaks down far before photon decoupling, because the photon field is not in equilibrium; the creation of a hydrogen atom in the ground state yields a photon of energy B_H, significantly changing the high-energy tail of the distribution. This photon quickly reionizes another hydrogen atom, resulting in no net change. Most of the recombination is due to processes where a hydrogen atom is formed in an $n = 2$ excited state which decays to the ground state, releasing a photon of energy $(3/4)B_H = L_\alpha$ in the process.

The large population of L_α photons causes most hydrogen atoms to be in $n = 2$ states, significantly delaying recombination since the energy gap is only effectively $1/4$ as large. The system can be investigated accurately numerically by taking $1s$, $2s$, and $2p$ hydrogen atoms, free electrons and positrons, and L_α photons as species in a set of coupled Boltzmann equations. Photon decoupling is irrelevant here, because there are effectively no thermal photons of energy L_α. The system is then in quasi-equilibrium, with a slow "leak" because the $2s \to 1s$ decay emits two photons. Eventually the residual ions "freezes out" of this system, leaving a similar relic density to the one computed more naively above.

3.4 Nucleosynthesis

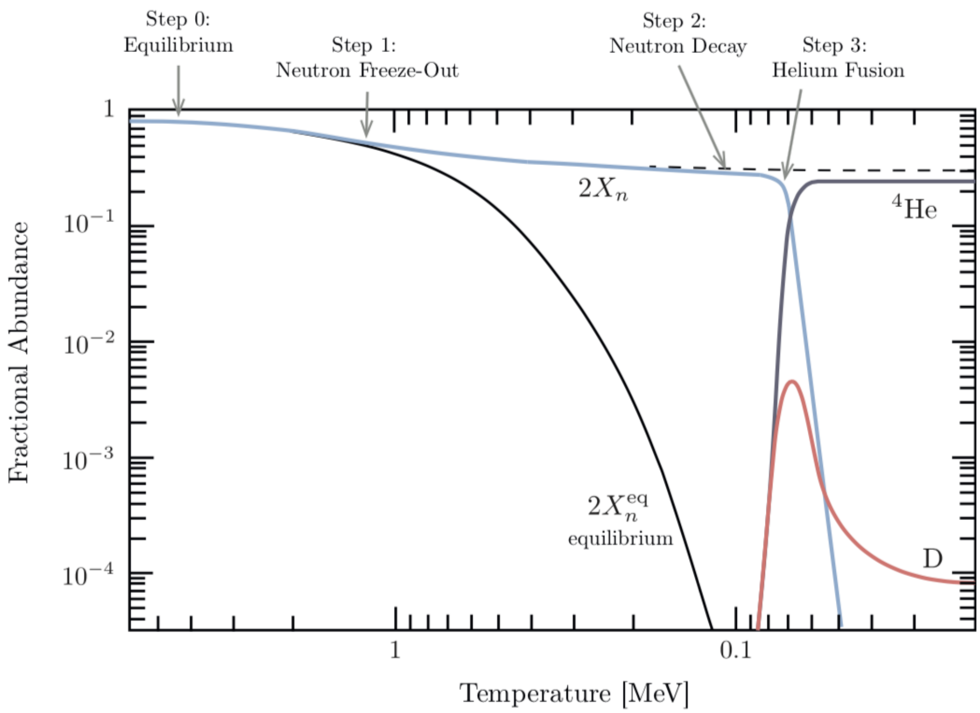
Next, we move backwards in time to cover nucleosynthesis.

- Nucleosynthesis occurs at energy scales of $T \sim 1\,\text{MeV}$. By this time, baryons have long since decoupled, but electrons and positrons have not. Weak nuclear reactions convert neutrons and protons into each other, and strong nuclear reactions build nuclei from them.

- We will concentrate on the light nuclei, which are

$$\mathrm{H} = p, \quad \text{deuterium } \mathrm{D} = pn, \quad \text{tritium } {}^3\mathrm{H} = pnn, \quad {}^3\mathrm{He} = ppn, \quad {}^4\mathrm{He} = ppnn.$$

Note that for this discussion, "hydrogen" is short for a hydrogen nucleus, i.e. a proton. All heavier nuclei are produced in much small quantities. An overview is shown below.



- We note that neutrons and protons are coupled by reactions like

$$n + \nu_e \leftrightarrow p^+ + e^-, \quad n + e^+ \leftrightarrow p^+ + \overline{\nu}_e.$$

We assume the chemical potentials of electrons and neutrinos are negligible, so $\mu_n \approx \mu_p$. Then

$$\left(\frac{n_n}{n_p}\right)_{\mathrm{eq}} \approx \left(\frac{m_n}{m_p}\right)^{3/2} e^{-(m_n - m_p)/T} \approx e^{-Q/T}, \quad Q = 1.30\,\mathrm{MeV}.$$

Hence for $T < 1\,\mathrm{MeV}$, the equilibrium neutron density falls rapidly. Around the same time neutrinos decouple, shutting down the weak processes above and allowing the neutron density to freeze out. This seems to be a coincidence, as neutrino decoupling is determined by the weak interaction while Q is determined by the strong and electromagnetic interactions. Electron-positron annihilation also occurs around this time, though it won't affect our story below.

- For comparison, deuterium can be produced in the reaction

$$n + p^+ \leftrightarrow \mathrm{D} + \gamma$$

and since $\mu_\gamma = 0$, we have $\mu_n + \mu_p = \mu_{\mathrm{D}}$. By the same reasoning as for the Saha equation,

$$\left(\frac{n_{\mathrm{D}}}{n_n n_p}\right)_{\mathrm{eq}} = \frac{3}{4}\left(\frac{m_{\mathrm{D}}}{m_n m_p}\frac{2\pi}{T}\right)^{3/2} e^{-(m_{\mathrm{D}} - m_n - m_p)/T}$$

since deuterium has spin 1 and hence $g_{\mathrm{D}} = 3$. Again we can approximate $m_{\mathrm{D}} \approx 2m_p \approx 2m_n$ in the prefactor, but must preserve the difference in the exponential,

$$\left(\frac{n_{\mathrm{D}}}{n_p}\right)_{\mathrm{eq}} = \frac{3}{4} n_n^{\mathrm{eq}} \left(\frac{4\pi}{m_p T}\right)^{3/2} e^{B_D/T}, \quad B_D = 2.22\,\mathrm{MeV}.$$

To get an order of magnitude estimate, note that $n_n \sim n_b = \eta n_\gamma \sim \eta T^3$, so

$$\left(\frac{n_{\mathrm{D}}}{n_p}\right)_{\mathrm{eq}} \sim \eta \left(\frac{T}{m_p}\right)^{3/2} e^{B_D/T}.$$

Because of the smallness of η, n_{D} is negligible for $T \gtrsim 0.1\,\mathrm{MeV}$, so we can ignore everything except for protons and neutrons before this point.

- Therefore, we will simply track the neutron fraction

$$X_n = \frac{n_n}{n_n + n_p}$$

until $T \sim 0.1\,\mathrm{MeV}$, at which point it will be used as an input for reactions producing heavier nuclei. We know that in equilibrium,

$$X_n^{\mathrm{eq}}(T) = \frac{e^{-Q/T}}{1 + e^{-Q/T}}.$$

Furthermore, as stated above, neutrons freeze out when neutrinos decouple at $T \sim 0.8\,\mathrm{MeV}$, at which point $X_n^{\mathrm{eq}} = 0.17$. We hence estimate $X_n^\infty \sim 1/6$.

- At temperatures below $0.2\,\mathrm{MeV}$, corresponding to $t \gtrsim 100\,\mathrm{s}$, we must account for the finite lifetime of the neutron,

$$X_n(t) = X_n^\infty e^{-t/\tau_n} = \frac{1}{6} e^{-t/\tau_n}, \quad \tau_n = 880.0 \pm 0.9\,\mathrm{s}.$$

- Now, heavier nuclei are primarily produced by two-body reactions, since the density is too low for three-body reactions. The primary ones are

$$n + p^+ \leftrightarrow \mathrm{D} + \gamma, \quad \mathrm{D} + p^+ \leftrightarrow {}^3\mathrm{He} + \gamma, \quad \mathrm{D} + {}^3\mathrm{He} \leftrightarrow {}^4\mathrm{He} \pm p^+.$$

These are strong interactions, and the first ensures that D tracks its equilibrium abundance. However, since B_D is relatively small and η is very small, the other interactions are not in equilibrium since there is too little deuterium. This is known as the deuterium bottleneck.

- As a rough estimate, note that $(n_{\mathrm{D}}/n_p)_{\mathrm{eq}} \sim 1$ at temperature

$$T_{\mathrm{nuc}} \sim 0.06\,\mathrm{MeV}, \quad t_{\mathrm{nuc}} \sim 330\,\mathrm{s}, \quad X_n(t_{\mathrm{nuc}}) \sim \frac{1}{8}$$

where the time is computed using the expression for $T(t)$ in a radiation-dominated universe. At this point we account for the reactions that produce ^{4}He. Since ^{4}He has such a high binding energy, almost all the deuterium is quickly converted to it, and since each ^{4}He requires two neutrons, we get

$$\frac{n_{\mathrm{He}}}{n_{\mathrm{H}}} \sim \frac{1}{16}$$

or alternatively, the mass fraction of helium is about 1/4.

- The amount of ^{3}He is smaller by a factor of about 10^4 because of its lower binding energy. Creation of heavier nuclei is restricted because there are no nuclei of atomic mass 5 or 8. Instead, the most common subsequent pathway for ^{4}He is fusion to create ^{7}Li or ^{7}Be in small amounts. These reactions can be treated numerically as a set of coupled Boltzmann equations. Nuclei heavier than ^{7}Be are negligible and are instead created much later inside stars.

- Experimentally, the primordial abundances of nuclei can be measured by observing dwarf galaxies, where little stellar nucleosynthesis has occurred, or very distant objects and hence younger objects, such as quasars. Within the context of BBN, the baryon to photon ratio is the only free parameter, but it has also been measured independently by CMB measurements. Using this value, the measured abundances of ^{3}He and ^{4}He are as expected, but the amount of ^{7}Li is not; this is called the lithium problem.

Note. Why does the deuteron have spin 1? One way to see this is to use isospin symmetry, which relates protons and neutrons. All three members of the isospin triplet $|nn\rangle, (|np\rangle + |pn\rangle)/\sqrt{2}, |pp\rangle$ have similar energies, and $|nn\rangle$ and $|pp\rangle$ are known to not be stable, so neither is the third state. The lowest energy states are in the s-wave, so the spin wavefunction must be antisymmetric to ensure overall antisymmetry of the wavefunction, so the spin is zero. Hence the spin zero deuteron is unstable.

Note. Since there is a sizable amount of ^{4}He, we should also account for helium recombination. This occurs much earlier than hydrogen recombination because the binding energies are much higher. For the first electron, the binding energy is $4 \times 13.6\,\text{eV} = 54.4\,\text{eV}$ and recombination occurs at $T \sim 15000\,\text{K}$. The second electron has a binding energy of only $24.62\,\text{eV}$ because of repulsion with the first, and recombines at $T \sim 5000\,\text{eV}$, at which point helium decouples from the photon bath.

Note. The final helium abundance is an important parameter that is a useful probe of new physics.

- The quantity g_* determines the Hubble parameter by $H \sim \sqrt{G_N g_*} T^2$ and hence the neutrino freeze-out temperature. Using the simple criterion $\Gamma \sim H$,

$$(G_F^2 T_f^2) T_f^3 \sim \sqrt{G_N g_*} T_f^2, \quad T_f \propto g_*^{1/6}.$$

  A larger value of g_* increases T_f, which increases the n/p ratio at freeze-out and hence increases the final helium abundance; this constrains models that change the number of light particles. Changing G_N and G_F would also affect the helium abundance by the same mechanism of changing T_f.

- A larger neutron lifetime τ_n would reduce the amount of neutron decay after freeze-out and hence would increase the final helium abundance.

- A larger mass difference Q between neutrons and protons would decrease the n/p ratio at freeze-out and hence would decrease the final helium abundance.

- A larger value for η allows synthesis of ^{4}He to begin earlier and hence increases its final abundance.

4 Cosmological Perturbation Theory

4.1 Newtonian Perturbation Theory

We now turn to the inhomogeneous universe. We use perturbation theory to describe the formation and evolution of large-scale structure. Since this theory is quite complicated, we begin with the Newtonian version, which applies for nonrelativistic matter on scales below the Hubble radius.

- Consider a nonrelativistic fluid with mass density ρ, pressure $P \ll \rho$, and velocity $\mathbf{u}$. Mass conservation gives the continuity equation
$$\partial_t \rho = -\nabla_{\mathbf{r}} \cdot (\rho \mathbf{u})$$
while momentum conservation gives the Euler equation
$$(\partial_t + \mathbf{u} \cdot \nabla_{\mathbf{r}})\mathbf{u} = -\frac{\nabla_{\mathbf{r}} P}{\rho} - \nabla_{\mathbf{r}} \Phi$$
where Φ is the gravitational potential, determined by the Poisson equation
$$\nabla_{\mathbf{r}}^2 \Phi = 4\pi G \rho.$$

- The fluid equations can be written more intuitively in terms of the convective derivative
$$D_t = \partial_t + \mathbf{u} \cdot \nabla_{\mathbf{r}}$$
which gives the rate of change of a quantity, following a fluid element as it moves. Then
$$D_t \rho = -\rho \nabla \cdot u, \quad D_t \mathbf{u} = -\frac{\nabla_{\mathbf{r}} P}{\rho} - \nabla_{\mathbf{r}} \Phi.$$

- We now consider a small perturbation about a constant background, e.g. $\rho(\mathbf{r}, t) = \overline{\rho}(t) + \delta\rho(t, \mathbf{r})$, where $\overline{\mathbf{u}} = 0$. Linearizing the perturbed equations and neglecting gravity, we find
$$\partial_t \delta\rho = -\nabla_{\mathbf{r}} \cdot (\overline{\rho} \mathbf{u}), \quad \overline{\rho} \partial_t \mathbf{u} = -\nabla_{\mathbf{r}} \delta P.$$
By combining these equations, we find
$$\partial_t^2 \delta\rho - \nabla_{\mathbf{r}}^2 \delta P = 0.$$

- For an adiabatic fluctuation, to be defined below, we have $\delta P = c_s^2 \delta\rho$, giving
$$(\partial_t^2 - c_s^2 \nabla_{\mathbf{r}}^2)\delta\rho = 0.$$
This is just the wave equation, where the parameter c_s is the speed of sound; fluctuations have constant amplitude.

- When we include gravity, we find an extra term,
$$(\partial_t^2 - c_s^2 \nabla_{\mathbf{r}}^2)\delta\rho = 4\pi G \overline{\rho} \delta\rho.$$
The dispersion relation is modified to
$$\omega^2 = c_s^2 k^2 - 4\pi G \overline{\rho}$$

for which the frequency is imaginary for wavenumbers between the

$$k_J = \frac{\sqrt{4\pi G\overline{\rho}}}{c_s}.$$

This indicates that fluctuations above the length scale $\lambda_J = 2\pi/k_J$, called the Jeans length, grow exponentially.

- Now, we need to account for the expansion of the universe. This can be handled in the Newtonian formalism by defining

$$\mathbf{r}(t) = a(t)\mathbf{x}$$

where $\mathbf{x}$ is a comoving coordinate. The velocity field is then

$$\mathbf{u}(t) = \dot{\mathbf{r}} = H\mathbf{r} + \mathbf{v}$$

where $\mathbf{v}$ is the proper velocity. We would now like to rewrite our above equations in terms of $\mathbf{x}$ and $\mathbf{v}$, and time derivatives at fixed $\mathbf{x}$ rather than at fixed $\mathbf{r}$.

- To handle the time derivatives, note that by the chain rule,

$$\partial_t|_\mathbf{r} = \partial_t|_\mathbf{x} + \left(\frac{\partial \mathbf{x}}{\partial t}\right)_\mathbf{r} \cdot \nabla_\mathbf{x} = \partial_t|_\mathbf{x} - H\mathbf{x} \cdot \nabla_\mathbf{x}.$$

From now on all time derivatives will hold $\mathbf{x}$ constant, and all gradients will be with respect to $\mathbf{x}$, so we'll suppress the subscripts.

- At zeroth order, the continuity equation becomes

$$\frac{\partial \overline{\rho}}{\partial t} + 3H\overline{\rho} = 0$$

which simply recovers $\overline{\rho} \propto a^{-3}$.

- At first order, defining the fractional density perturbation $\delta = \delta\rho/\overline{\rho}$, we find

$$\dot{\delta} = -\frac{1}{a}\nabla \cdot \mathbf{v}.$$

Similar manipulation of the Euler equations gives

$$\dot{\mathbf{v}} + H\mathbf{v} = -\frac{1}{a\overline{\rho}}\nabla\delta P - \frac{1}{a}\nabla\delta\Phi.$$

At zeroth order, this simply indicates that $\mathbf{v} \propto a^{-1}$, a familiar result.

- Finally, the Poisson equation easily becomes

$$\nabla^2\delta\Phi = 4\pi G a^2\overline{\rho}\delta.$$

- Considering adiabatic perturbations as before, we find that

$$\ddot{\delta} + 2H\dot{\delta} - \frac{c_s^2}{a^2}\nabla^2\delta = 4\pi G\overline{\rho}\delta.$$

The essential differences are that we now have a Hubble friction term, and that both $c_s(t)$ and $\overline{\rho}(t)$ have power-law time dependence. The result is that below the Jeans length, fluctuations oscillate with decreasing amplitude, while above the Jeans length, the fluctuations grow but only by a power law.

As a concrete example, we'll consider dark matter within the Hubble scale.

- Dark matter is special because it doesn't interact significantly with itself or with other matter or radiation, so the pressure term above can be dropped; note that this sets the Jeans length to zero, i.e. we lose all k-dependence. This means that dark matter behaves very differently from ordinary matter during structure formation, tending to clump up on smaller scales, and it is in fact required to match the observed results.

- During matter domination, we have
$$\ddot{\delta}_m + 2H\dot{\delta}_m - 4\pi G\overline{\rho}_m\delta_m = 0.$$
Since $a \propto t^{2/3}$, we have $H = 2/3t$ which implies
$$\ddot{\delta}_m + \frac{4}{3t}\dot{\delta}_m - \frac{2}{3t^2}\delta_m = 0.$$
Every term is the same power in t, so it is natural to guess a polynomial, giving solutions
$$\delta_m \propto \begin{cases} t^{-1} \propto a^{-3/2} \\ t^{2/3} \propto a \end{cases}.$$
A generic perturbation will contain both modes; the growing mode is the latter one and the only one we see at late times. Hence dark matter perturbations grow like the scale factor during matter domination. Below, we will always refer to the larger mode as the "growing mode", even if it is decaying.

- Now we consider radiation domination. In this case we must add an additional term to the equation above due to the contribution of $\delta\rho_r$ to $\delta\Phi$. We cannot analyze radiation fluctuations here since they are inherently relativistic, but it turns out that they oscillate as sound waves below the Hubble radius, so that $\delta\rho_r$ time averages to zero.

- Then the only change is the behavior of $\overline{\rho}_m$ and H,
$$\ddot{\delta}_m + \frac{1}{t}\dot{\delta}_m - 4\pi G\overline{\rho}_m\delta_m \approx 0.$$
Since δ_m evolves only on cosmological timescale, as there is no pressure term to make it do otherwise, we have $\partial_t \sim H$ by order of magnitude. But then the final term above is much smaller, because we are in radiation domination.

- The equation can then be solved straightforwardly to get
$$\delta_m \propto \begin{cases} \text{const} \\ \log t \propto \log a \end{cases}.$$
That is, the rapid expansion due to the presence of radiation reduces the growth of δ_m to logarithmic. The growth of dark matter perturbations almost pauses during radiation domination.

- Finally, consider the Λ-dominated era. We don't have to include a δ_Λ contribution to $\delta\Phi$, because dark energy is always uniform by definition. Then we have
$$\ddot{\delta}_m + 2H\dot{\delta}_m - 4\pi G\overline{\rho}_m\delta_m = 0.$$

By the same argument as for radiation domination, the final term is negligible, but by contrast H is now constant. Then the solutions are

$$\delta_m \propto \begin{cases} \text{const} \\ e^{-2Ht} \propto a^{-2} \end{cases}.$$

Evidently, matter fluctuations stop growing at this stage.

4.2 Relativistic Perturbation Theory

We now use a full relativistic treatment, which can handle all length scales and relativistic fluids.

- We consider small perturbations about the Friedmann metric, $g_{\mu\nu} = \overline{g}_{\mu\nu} + \delta g_{\mu\nu}$. The metric perturbations will be coupled to matter by the Einstein equations. For simplicity, we only expand about a flat FRW background,

$$ds^2 = a^2(\tau)(d\tau^2 - \delta_{ij} dx^i dx^j).$$

- Lorentz symmetry is broken by the background, but rotational symmetry is preserved, so it is useful to sort the perturbations by their rotational transformation properties. We write

$$ds^2 = a^2(\tau)\left((1+2A)d\tau^2 - 2B_i dx^i d\tau - (\delta_{ij} + h_{ij})dx^i dx^j\right)$$

  where A is a scalar, B_i is a vector, and h_{ij} is a tensor. We raise and lower the Latin spatial indices with δ_{ij}, because we are linearizing about a rotationally symmetric background.

- We can further decompose B_i into curl-free and divergence-free components,

$$B_i = \partial_i B + \hat{B}_i$$

  where hatted quantities have zero divergence, $\partial^i \hat{E}_i = 0$. Similarly, h_{ij} can be decomposed as

$$h_{ij} = 2C\delta_{ij} + 2\partial_{\langle i}\partial_{j\rangle}E + 2\partial_{(i}\hat{E}_{j)} + 2\hat{E}_{ij}$$

  where we have defined

$$\partial_{\langle i}\partial_{j\rangle}E = \left(\partial_i\partial_j - \frac{1}{3}\partial_{ij}\nabla^2\right)E$$

  and the symmetric tensor $\hat{E}_{ij}$ is divergenceless and traceless, $\partial^i \hat{E}_{ij} = \hat{E}^i{}_i = 0$.

- We have thus decomposed the ten degrees of freedom of the metric into four scalars, A, B, C, and E, the two vectors $\hat{B}_i$ and $\hat{E}_i$, and the tensor $\hat{E}_{ij}$, which carries two degrees of freedom. This is called the SVT decomposition. Furthermore, it turns out that the scalars, vectors, and tensor evolve independently under the linearized Einstein equation. It is useful to think of all these fields as propagating on top of a rotationally invariant background metric, rather than being part of the metric itself.

- Scalar perturbations are rather generic. Inflation does not predict vector perturbations, though these would decay quickly anyway. However, inflation notably predicts tensor perturbations, though these have not yet been observed.

Note. We would like to think of $\delta g_{\mu\nu}$ as an independent tensor field propagating on top of the background $\overline{g}_{\mu\nu}$, but how can we make this mathematically precise? We imagine two spacetime manifolds, a "physical" M_p with metric $g_{\mu\nu}$ and a "background" M_b with metric $\overline{g}_{\mu\nu}$, along with a diffeomorphism $\phi\colon M_b \to M_p$. Then $\delta g_{\mu\nu}$ may be defined as $((\phi^* g) - \overline{g})_{\mu\nu}$. The gauge symmetry addressed below arises from the fact that the diffeomorphism ϕ is ambiguous up to composition with a diffeomorphism $\psi\colon M_b \to M_b$.

Before continuing, we need to account for the gauge symmetry remaining in the metric.

- The metric perturbations depend on our choice of coordinates, or gauge choice. This can be interpreted as the combination of a choice of timeslicing for the spacetime, and a choice of spatial coordinates on those timeslices. Changing either can make fictitious perturbations appear, or make real perturbations vanish.

- Concretely, consider the coordinate transformation
$$X^\mu \to \tilde{X}^\mu = X^\mu + \xi^\mu(\tau, \mathbf{x}), \quad \xi^0 = T, \quad \xi^i = L^i = \partial^i L + \hat{L}^i.$$
The transformation of the metric is
$$g_{\mu\nu}(X) = \frac{\partial \tilde{X}^\alpha}{\partial X^\mu} \frac{\partial \tilde{X}^\beta}{\partial X^\nu} \tilde{g}_{\alpha\beta}(\tilde{X})$$
and we wish to compare the metric perturbations of $g_{\mu\nu}$ and $\tilde{g}_{\mu\nu}$. In practice, we will only be interested in coordinate transformations that take weakly perturbed metrics to other weakly perturbed metrics; other transformations exist, but then cosmological perturbation theory will not be applicable anymore. Hence we can take ξ^μ to be of the same order as the SVT parameters, and expand everything to linear order.

- For example, we have
$$g_{00}(X) = \left(\frac{\partial \tilde{\tau}}{\partial \tau}\right)^2 \tilde{g}_{00}(\tilde{X})$$
where all other terms are higher order. Then we have
$$a^2(\tau)(1 + 2A) = (1 + T')^2 a^2(\tau + T)(1 + 2\tilde{A}).$$
As usual, a prime denotes a derivative with respect to conformal time τ. Defining $\mathcal{H} = a'/a$ and expanding everything to linear order, we find
$$\tilde{A} = A - T' - \mathcal{H}T.$$
Similarly, we have
$$\tilde{B}_i = B_i + \partial_i T - L_i', \quad \tilde{h}_{ij} = h_{ij} - 2\partial_{(i} L_{j)} - 2\mathcal{H}T\delta_{ij}.$$

- In terms of the SVT parameters, we have
$$A \to A - T' - \mathcal{H}T, \quad B \to B + T - L', \quad C \to C - \mathcal{H}T - \frac{1}{3}\nabla^2 L, \quad E \to E - L$$
and
$$\hat{B}_i \to \hat{B}_i - \hat{L}_i', \quad \hat{E}_i \to \hat{E}_i - \hat{L}_i, \quad \hat{E}_{ij} \to \hat{E}_{ij}.$$
In accordance with rotational symmetry, the scalars T and L' can only affect the scalars, the vector $\hat{L}$ can only affect the vectors, and there is nothing that can affect the tensor.

- To avoid gauge problems, we can work in terms of the gauge-invariant Bardeen variables

$$\Psi = A + \mathcal{H}(B - E') + (B - E')', \quad \Phi = -C - \mathcal{H}(B - E') + \frac{1}{3}\nabla^2 E, \quad \hat{\Phi}_i = \hat{E}_i' - \hat{B}_i, \quad \hat{E}_{ij}.$$

- An alternative solution is to fix the gauge. For simplicity, from this point on we'll focus on only scalar perturbations, setting the rest to zero. We can set the values of only two scalar perturbations; other coordinate transformations will reintroduce the vector perturbations.

- In Newtonian gauge, we set

$$B = E = 0$$

in which case $A = \Psi$ and $C = -\Phi$, leaving a perturbed metric of the form

$$ds^2 = a^2(\tau)\left((1 + 2\Psi)d\tau^2 - (1 - 2\Phi)\delta_{ij}dx^i dx^j\right).$$

This is a complete gauge fixing for perturbations that decay at spatial infinity. The advantage is its similarity to the weak-field limit of GR about Minkowski space, where Ψ plays the role of the gravitational potential. We will see later that in the absence of anisotropic stress, $\Psi = \Phi$.

- Another gauge is spatially flat gauge, $C = E = 0$, which fixes the spatial part of the metric; this will be useful when considering how perturbations are sourced by the inflaton.

The next step is to parametrize perturbed matter.

- We recall that for a perfect fluid,

$$\overline{T}^\mu{}_\nu = (\overline{\rho} + \overline{P})\overline{U}^\mu \overline{U}_\nu - \overline{P}\delta^\mu_\nu.$$

We use the mixed form of the metric because its components are slightly easier to interpret, and because the last term has a slightly simpler form.

- Now we consider small perturbations,

$$T^\mu{}_\nu = \overline{T}^\mu{}_\nu + \delta T^\mu{}_\nu, \quad \delta T^\mu{}_\nu = (\delta\rho + \delta P)\overline{U}^\mu \overline{U}_\nu + (\overline{\rho} + \overline{P})(\delta U^\mu \overline{U}_\nu + \overline{U}^\mu \delta U_\nu) - \delta P \delta^\mu_\nu - \Pi^\mu{}_\nu$$

where $\Pi^\mu{}_\nu$ is called the anisotropic stress, which obeys $\Pi_{\mu\nu} = \Pi_{\nu\mu}$. It turns out that by redefining other variables, we may always choose

$$\Pi^i{}_i = 0, \quad \overline{U}^\mu \Pi_{\mu\nu} = 0$$

and, working in a frame where $\overline{U}^\mu = a^{-1}\delta^0_\mu$, the only surviving components of the anisotropic stress are Π_{ij}, representing shear forces. However, anisotropic stress will be negligible for all scenarios considered in these notes, so we will drop it later.

- Perturbations in the four-velocity will induce nonvanishing energy flux $T^0{}_j$ and momentum density $T^i{}_0$. To compute them, we need to parametrize the four-velocity. Note that

$$\delta g_{\mu\nu}\overline{U}^\mu \overline{U}^\nu + 2\overline{U}_\mu \delta U^\mu = 0$$

since the four-velocity must always have unit norm. This implies $\delta U^0 = -Aa^{-1}$, so

$$U^\mu = a^{-1}(1 - A, v^i), \quad v^i = \frac{dx^i}{d\tau}$$

where v^i is the coordinate velocity. By lowering both sides and keeping only linear terms,

$$U_\mu = a(1 + A, -(v_i + B_i)).$$

- Plugging this into our expression for $\delta T^\mu{}_\nu$ we have

$$\delta T^0{}_0 = \delta\rho, \quad \delta T^i{}_0 = (\overline{\rho}+\overline{P})v^i, \quad \delta T^0{}_j = -(\overline{\rho}+\overline{P})(v_j + B_j), \quad \delta T^i{}_j = -\delta P \delta^i_j - \Pi^i{}_j.$$

We hence define the momentum density $q^i = (\overline{\rho}+\overline{P})v^i$. For multiple components, each with an independent velocity, the perturbations simply add.

- Under coordinate transformations, we have

$$T^\mu{}_\nu(X) = \frac{\partial X^\mu}{\partial \tilde{X}^\alpha}\frac{\partial \tilde{X}^\beta}{\partial X^\nu}\tilde{T}^\alpha{}_\beta(\tilde{X})$$

which gives

$$\delta\rho \to \delta\rho - T\overline{\rho}', \quad \delta P \to \delta P - T\overline{P}', \quad q_i \to q_i + (\overline{\rho}+\overline{P})L_i', \quad v_i \to v_i + L_i'$$

with Π_{ij} invariant under coordinate transformations.

- Note that all of these quantities have SVT decompositions, such as

$$v_i = \partial_i v + \hat{v}_i.$$

Since we will only be considering scalar perturbations, we may toss away the second term **(is this right?)** , so that

$$v_i = \partial_i v, \quad q_i = \partial_i q.$$

- We may define gauges in terms of the matter perturbation. In uniform density gauge, we can use the freedom in the time slicing to set $\delta\rho = 0$. In comoving gauge, we set $q = 0$, where $q_i = \partial_i q + \hat{q}_i$. In both cases, we have the freedom to set one further scalar metric perturbation to zero, and we choose $B = 0$.

- One important gauge-invariant combination is

$$\boxed{\overline{\rho}\Delta \equiv \delta\rho + \overline{\rho}'(v+B).}$$

Note that in comoving gauge, $\overline{\rho}\Delta = \delta\rho$, so Δ is called the comoving gauge density perturbation.

- Simple, single-field inflation models predict initial fluctuations that are adiabatic, which means that the local state of matter at some spacetime point $(\tau, \mathbf{x})$ is the same as in the background universe at some slightly different time $\tau + \delta\tau(\mathbf{x})$. We can view some parts of the universe as being "ahead" in evolution compared to others.

- For a time shift $\delta\tau$, we have

$$\delta\rho_I = \overline{\rho}_I'\delta\tau(\mathbf{x}).$$

In particular, $\delta\rho_I/\overline{\rho}_I'$ is the same for each species I. Assuming the energy continuity equation is conserved for each species separately, $\overline{\rho}_I' = -3\mathcal{H}(1+w_I)\overline{\rho}_I$, we find

$$\frac{\delta_I}{1+w_I}$$

is the same for every species, where we have defined the fractional density contrast $\delta_I = \delta\rho_I/\overline{\rho}_I$.

- For example, all radiation perturbations are related to all matter perturbations by $\delta_r = (4/3)\delta_m$. Since all the δ_I are on the same order of magnitude, the total density perturbation

$$\delta\rho_{\text{tot}} = \overline{\rho}_{\text{tot}}\delta_{\text{tot}} = \sum_I \overline{\rho}_I \delta_I$$

is dominated by the dominant species. Also note that the δ_I are functions of $\mathbf{x}$.

- Perturbations that are not adiabatic are called isocurvature perturbations and parametrized by

$$S_{IJ} = \frac{\delta_I}{1+w_I} - \frac{\delta_J}{1+w_J}.$$

All present observational data is consistent with $S_{IJ} = 0$.

Note. Anisotropic stress is negligible for perfect fluids, which are characterized by strong interactions which keep the pressure isotropic. Decoupled or weakly interacting species such as neutrinos cannot be described in this way, and hence we must account for their anisotropic stress. Decoupled cold dark matter is collisionless with a negligible velocity dispersion; we can hence describe it as a pressureless perfect fluid, even though it has almost no interactions, and hence without using anisotropic stress.

4.3 Equations of Motion

Finally, we investigate the equations of motion. Since the calculations are tedious but straightforward, we simply quote the final results.

- We work in Newtonian gauge and continue to ignore vector and tensor perturbations, so

$$g_{\mu\nu} = a^2 \begin{pmatrix} 1+2\Psi & 0 \\ 0 & -(1-2\Phi)\delta_{ij} \end{pmatrix}.$$

We set the anisotropic stress to zero, $\Pi_{ij} = 0$. As we will see later, this implies $\Psi = \Phi$.

- The perturbed connection coefficients are

$$\Gamma^0_{00} = \mathcal{H} + \Psi', \quad \Gamma^0_{0i} = \partial_i \Psi, \quad \Gamma^i_{00} = \partial^j \Psi$$

and

$$\Gamma^0_{ij} = \mathcal{H}\delta_{ij} - (\Phi' + 2\mathcal{H}(\Phi+\Psi))\delta_{ij}, \quad \Gamma^i_{j0} = \mathcal{H}\delta^i_j - \Phi'\delta^i_j, \quad \Gamma^i_{jk} = -\delta^i_j\partial_k\Phi - \delta^i_k\partial_j\Phi + \delta_{jk}\partial^i\Phi$$

where, as usual, we raise spatial indices as $\partial^i = \delta^{ij}\partial_j$.

- Next, we consider the perturbed stress-energy conservation equation $\nabla_\mu T^\mu{}_\nu = 0$. At zeroth order, the zero component is

$$\overline{\rho}' = -3\mathcal{H}(\overline{\rho} + \overline{P})$$

which is the expected energy continuity equation. At first order, we have

$$\delta\rho' = -3\mathcal{H}(\delta\rho + \delta P) + 3\Phi'(\overline{\rho} + \overline{P}) - \nabla \cdot \mathbf{q}.$$

The first term is just the usual dilution due to the background expansion. The third term is familiar from Newtonian perturbation theory, while the second term is relativistic, corresponding to the density changes due to perturbations to the local expansion rate.

- It is useful to rewrite this equation as

$$\boxed{\delta' + \left(1 + \frac{\overline{P}}{\overline{\rho}}\right)(\nabla \cdot \mathbf{v} - 3\Phi') + 3\mathcal{H}\left(\frac{\delta P}{\delta \rho} - \frac{\overline{P}}{\overline{\rho}}\right)\delta = 0}$$

  where we have defined the fractional overdensity $\delta = \delta\rho/\overline{\rho}$.

- Next, the spatial components give the Euler equation

$$\boxed{\mathbf{v}' + \mathcal{H}\mathbf{v} - 3\mathcal{H}\frac{\overline{P}'}{\overline{\rho}'}\mathbf{v} = -\frac{\nabla \delta P}{\overline{\rho} + \overline{P}} - \nabla\Psi.}$$

  The second term on the left is the familiar redshifting due to expansion; the third term is an $O(\overline{P}/\overline{\rho})$ correction for relativistic fluids. The terms on the right are due to pressure gradients and gravitational infall, where the pressure gradient term again has a correction for the pressure.

- Next, we approach the Einstein field equation. The Ricci tensor is

$$R_{00} = -3\mathcal{H}' + \nabla^2\Psi + 3\mathcal{H}(\Phi' + \Psi') + 3\Phi'', \quad R_{0i} = 2\partial_i\Phi' + 2\mathcal{H}\partial_i\Psi$$

  and

$$R_{ij} = (\mathcal{H}' + 2\mathcal{H}^2 - \Phi'' + \nabla^2\Phi - 2(\mathcal{H}' + 2\mathcal{H}^2)(\Phi + \Psi) - \mathcal{H}\Psi' - 5\mathcal{H}\Phi')\delta_{ij} + \partial_i\partial_j(\Phi - \Psi).$$

  The Ricci scalar is then straightforwardly

$$a^2 R = -6(\mathcal{H}' + \mathcal{H}^2) + 2\nabla^2\Psi - 4\nabla^2\Phi + 12(\mathcal{H}' + \mathcal{H}^2)\Psi + 6\Phi'' + 6\mathcal{H}(\Psi' + 3\Phi').$$

- By more straightforward manipulations, the Einstein tensor is

$$G_{00} = 3\mathcal{H}^2 + 2\nabla^2\Phi - 6\mathcal{H}\Phi', \quad G_{0i} = 2\partial_i(\Phi' + \mathcal{H}\Psi)$$

  and

$$G_{ij} = -(2\mathcal{H}' + \mathcal{H}^2)\delta_{ij} + (\nabla^2(\Psi - \Phi) + 2\Phi'' + 2(2\mathcal{H}' + \mathcal{H}^2)(\Phi + \Psi) + 2\mathcal{H}\Psi' + 4\mathcal{H}\Phi')\delta_{ij} + \partial_i\partial_j(\Phi - \Psi).$$

  The first-order terms come from three places: the change in R_{ij}, the change in R, and the change in the metric g_{ij} that multiplies R.

- Now we consider Einstein's equations, $G_{\mu\nu} = 8\pi G T_{\mu\nu}$. First, taking the trace-free part of the spatial components, we have

$$\partial_{\langle i}\partial_{j\rangle}(\Phi - \Psi) = 0$$

  since we have assumed there is no anisotropic stress. Assuming the perturbations vanish at infinity, this implies $\Phi = \Psi$, which dramatically simplifies the equations above.

- The 00 component is

$$\mathcal{H}^2 = \frac{8\pi G}{3}a^2\overline{\rho}$$

  at zeroth order, which is just the Friedmann equation. At first order, we have

$$\nabla^2\Phi = 4\pi G a^2\overline{\rho}\delta + 3\mathcal{H}(\Phi' + \mathcal{H}\Phi)$$

  where we simplified using the Friedmann equation.

- Finally, the $0i$ component is
$$\partial_i(\Phi' + \mathcal{H}\Phi) = -4\pi G a^2 q_i.$$
Since the left-hand side is curl-free, this is only consistent if q_i has only a scalar part, which is as expected; the vector part of q_i would source vector perturbations. **(right?)** We hence may write $q_i = (\overline{\rho} + \overline{P})\partial_i v$. Assuming the perturbations decay at infinity, integrating both sides gives
$$\boxed{\Phi' + \mathcal{H}\Phi = -4\pi G a^2 (\overline{\rho} + \overline{P})v.}$$

- Substituting this into the 00 Einstein equation simplifies it to
$$\boxed{\nabla^2\Phi = 4\pi G a^2 \overline{\rho}\Delta,} \quad \overline{\rho}\Delta \equiv \overline{\rho}\delta - 3\mathcal{H}(\overline{\rho} + \overline{P})v$$
where Δ is defined in the same way as above, but specialized to Newtonian gauge.

- Now consider the spatial trace part, $G^i{}_i = 8\pi G T^i{}_i$. Note that the indices on $G^i{}_i$ are contracted with the full metric, so there is an additional term from the metric perturbation, and
$$G^i{}_i = -3a^{-2}\left(-(2\mathcal{H}' + \mathcal{H}^2) + 2(\Phi'' + 3\mathcal{H}\Phi' + (2\mathcal{H}' + \mathcal{H}^2)\Phi\right).$$
The trace of the energy-momentum tensor is $T^i_{\ i} = -3(\overline{P} + \delta P)$. At zeroth order,
$$2\mathcal{H}' + \mathcal{H}^2 = -8\pi G a^2 \overline{P}$$
which is just the second Friedmann equation. At first order, we get
$$\boxed{\Phi'' + 3\mathcal{H}\Phi' + (2\mathcal{H}' + \mathcal{H}^2)\Phi = 4\pi G a^2 \delta P.}$$

- Of course, the equations of motion we have derived are redundant due to the Bianchi identity. For example, we know the zeroth-order second Friedmann equation can be derived from the first Friedmann equation and the energy continuity equation; the first-order second Friedmann equation can be derived in a similar way.

One useful conserved quantity is the comoving curvature perturbation.

- In an arbitrary gauge, the induced metric for hypersurfaces of constant time is
$$\gamma_{ij} = a^2((1+2C)\delta_{ij} + 2E_{ij}), \quad E_{ij} = \partial_{\langle i}\partial_{j\rangle}E$$
for scalar perturbations. A tedious but straightforward computation shows that the three-dimensional Ricci scalar for the hypersurfaces satisfies
$$a^2 R_{(3)} = -4\nabla^2\left(C - \frac{1}{3}\nabla^2 E\right).$$

- We define the comoving curvature perturbation as
$$\mathcal{R} = C - \frac{1}{3}\nabla^2 E + \mathcal{H}(B + v).$$
The point of this quantity is that it is gauge invariant, as can be seen by plugging in the gauge transformations, and in comoving gauge ($B = v = 0$) it reduces to $C - (1/3)\nabla^2 E$, which appears in the expression for $R_{(3)}$.

- In Newtonian gauge, we have $B = E = 0$ and $C \equiv -\Phi$, so

$$\mathcal{R} = -\Phi + \mathcal{H}v.$$

We can use the $0i$ Einstein equation to eliminate the peculiar velocity v, for

$$\mathcal{R} = -\Phi - \frac{\mathcal{H}(\Phi' + \mathcal{H}\Phi)}{4\pi G a^2(\overline{\rho} + \overline{P})}.$$

- To investigate the time-dependence of $\mathcal{R}$, we take the time derivative of both sides and simplify using the Friedmann equation and Poisson equation, giving

$$\boxed{-4\pi G a^2(\overline{\rho} + \overline{P})\mathcal{R}' = 4\pi G a^2 \mathcal{H}\,\delta P_{\text{nad}} + \mathcal{H}\frac{\overline{P}'}{\overline{\rho}'}\nabla^2\Phi.}$$

Here we have defined the non-adiabatic pressure perturbation

$$\delta P_{\text{nad}} = \delta P - \frac{\overline{P}'}{\overline{\rho}'}\delta\rho$$

which vanishes for adiabatic fluctuations.

- Setting $\delta P_{\text{nad}} = 0$, the right-hand side scales like $\mathcal{H}k^2\Phi \sim \mathcal{H}k^2\mathcal{R}$, while the left-hand side is like $\mathcal{H}^2\mathcal{R}'$ by the Friedmann equation. Rearranging, we have

$$\frac{d\log\mathcal{R}}{d\log a} \sim \left(\frac{k}{\mathcal{H}}\right)^2$$

which means that super-Hubble Fourier modes of $\mathcal{R}$ evolve slowly.

- The quantity $\mathcal{H}$ decreases during inflation, then increases once inflation ends. Quantum fluctuations during inflation determine the value of Fourier modes of $\mathcal{R}$, which are frozen in once their wavelengths become larger than $1/\mathcal{H}$. Once their wavelengths become smaller again, they may evolve again, growing to give rise to the structure observed in our universe.

4.4 Structure Formation

Now we investigate relativistic structure formation. This is usually done numerically, but we will make approximations to get simple analytic results. We will start by considering $\mathcal{R}$, which is useful because it is conserved and sourced by inflation; however, to get physical results we will determine the evolution of Φ, which in turn determines the evolution of the density contrasts δ_i.

- First, note that by using the Friedmann equation, we have

$$\mathcal{R} = -\Phi - \frac{2}{3(1+w)}\left(\frac{\Phi'}{\mathcal{H}} + \Phi\right)$$

in Newtonian gauge, where we have assumed the background is dominated by a single component with equation of state parameter w. Sound waves in this component have $c_s^2 \approx w$.

- For adiabatic perturbations, Einstein's equations imply the gravitational potential evolves as

$$\Phi'' + 3(1+w)\mathcal{H}\Phi' + wk^2\Phi = 0.$$

However, this equation only applies if w is constant; we must revert to the original Einstein equation when w changes.

- On superhorizon scales, $k \ll \mathcal{H}$, the last term above is negligible, and the growing mode has Φ constant, regardless of the value of w. The decaying mode falls exponentially over a conformal timescale $1/\mathcal{H}$.

- In Newtonian gauge, the Poisson equation reads

$$\delta = -\frac{2}{3}\frac{k^2}{\mathcal{H}^2}\Phi - \frac{2}{\mathcal{H}}\Phi' - 2\Phi$$

where δ is the total density contrast. On superhorizon scales, both the first two terms are negligible compared to the third, for both the growing and decaying modes, so

$$\delta \approx -2\Phi = \text{const}.$$

- During radiation domination, we have $\delta_r \approx \delta$, and hence for adiabatic perturbations

$$\delta_m = \frac{3}{4}\delta_r \approx -\frac{3}{2}\Phi_{\text{RD}}.$$

However, Φ changes when we transition to matter domination, even for $k \gg \mathcal{H}$. To track its change, we use the conservation of $\mathcal{R}$. In the superhorizon limit we have

$$\boxed{\mathcal{R} = -\frac{5+3w}{3+3w}\Phi}$$

by approximating the expression for $\mathcal{R}$ above. By conserving $\mathcal{R}$, we have

$$\Phi_{\text{MD}} = \frac{9}{10}\Phi_{\text{RD}}$$

again on superhorizon scales.

Now we consider the evolution of Φ when modes enter the horizon.

- During radiation domination, $w = 1/3$, so the evolution equation is

$$\Phi'' + \frac{4}{\tau}\Phi' + \frac{k^2}{3}\Phi = 0.$$

The general solution is given in terms of spherical Bessel and Neumann functions,

$$\Phi_{\mathbf{k}}(\tau) = A_{\mathbf{k}}\frac{j_1(x)}{x} + B_{\mathbf{k}}\frac{n_1(x)}{x}, \quad x = \frac{k\tau}{\sqrt{3}}.$$

Explicitly, the special functions above are

$$j_1(x) = \frac{\sin x}{x^2} - \frac{\cos x}{x}, \quad n_1(x) = -\frac{\cos x}{x^2} - \frac{\sin x}{x}.$$

- Note that $n_1(x)$ blows up at early times (small x), so we reject it as a solution, $B_{\mathbf{k}} = 0$. We match the value $A_{\mathbf{k}}$ to the primordial value of the potential, $\Phi_{\mathbf{k}}(0) = (-2/3)\mathcal{R}_{\mathbf{k}}(0)$, giving

$$\Phi_{\mathbf{k}}(\tau) = -2\mathcal{R}_{\mathbf{k}}(0)\,\frac{\sin x - x\cos x}{x^3}.$$

  The quantity $\mathcal{R}_{\mathbf{k}}(0)$ will be determined statistically by inflation.

- The mode enters the horizon when $x \sim 1$. For $x \gg 1$ we have

$$\Phi_{\mathbf{k}}(\tau) \approx -6\mathcal{R}_{\mathbf{k}}(0)\frac{\cos(k\tau/\sqrt{3})}{(k\tau)^2}.$$

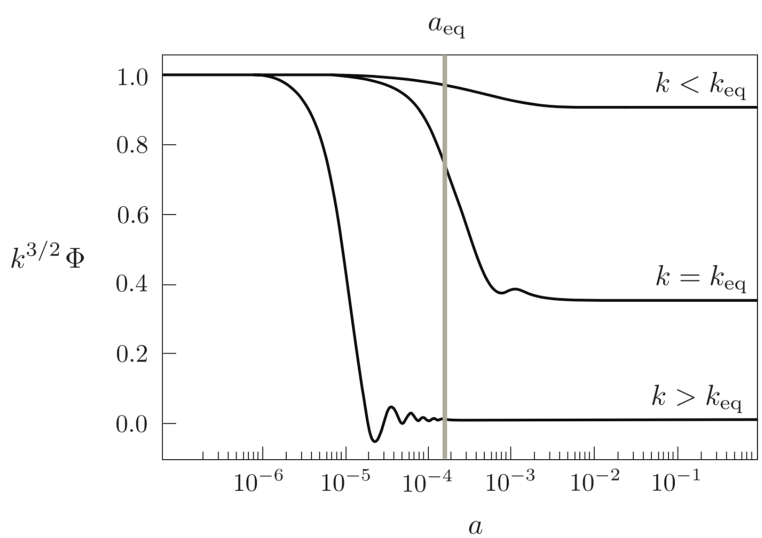
  Hence during the radiation era, subhorizon modes of Φ oscillate with frequency $k/\sqrt{3}$ and decay as $1/\tau^2 \propto 1/a^2$.

- In the matter era, $w = 0$, so the evolution equation is

$$\Phi'' + \frac{6}{\tau}\Phi' = 0.$$

  The growing mode is constant, so the gravitational potential is frozen on all scales during matter domination.

- The results are shown in the numeric plot below. Let k_{eq} be the value of the horizon scale at matter-radiation equality.



  For $k < k_{\text{eq}}$, the mode is frozen completely, except for the 10% decrease during the transition to matter domination. For $k \sim k_{\text{eq}}$, the mode is somewhat suppressed during the radiation era, while for $k \gg k_{\text{eq}}$, the mode is strongly suppressed. We will see that inflation predicts $|\mathcal{R}_{\mathbf{k}}| \sim k^{-3/2}$, explaining the choice of y-axis.

- For tensor perturbations, it can be shown that during matter domination and zero anisotropic stress,

$$\hat{E}_{ij,\mathbf{k}}(\tau) \approx \frac{k\tau\cos(k\tau) - \sin(k\tau)}{(k\tau)^3}.$$

This is constant on superhorizon scales, which is a general result independent of the matter content. It oscillates and decays on subhorizon scales; the same behavior also appears for radiation domination.

Next, we consider the evolution of perturbations in the radiation density.

- These perturbations dominate during the radiation era and hence determine the value of Φ, so the equation of motion is the Poisson equation,

$$\delta_r = -\frac{2}{3}(k\tau)^2\Phi - 2\tau\Phi' - 2\Phi, \quad \Delta_r = -\frac{2}{3}(k\tau)^2\Phi.$$

  As a result, for $x \ll 1$, δ_r is constant while $\Delta_r \propto \tau^2 \propto a^2$. Inside the horizon,

$$\delta_r \approx \Delta_r = 4\mathcal{R}(0)\cos(k\tau/\sqrt{3}).$$

  We thus see that δ_r oscillates inside the horizon with constant amplitude; it is the solution to

$$\delta_r'' - \frac{1}{3}\nabla^2\delta_r = 0.$$

- The result $\delta_r \approx \Delta_r$ above is essentially because they are equal in some gauge, and there are no gauge ambiguities for subhorizon modes. This is because the only gauge transformations that are relevant to our perturbation theory are those that preserve the approximately FLRW background, and which hence vary slowly with respect to the horizon scale. **(correct?)**
- During the matter era, radiation perturbations are subdominant, so their evolution is instead determined by conservation equations. On subhorizon scales, the energy continuity and Euler equations are

$$\delta_r' = -\frac{4}{3}\nabla\cdot\mathbf{v}_r, \quad \mathbf{v}_r' = -\frac{1}{4}\nabla\delta_r - \nabla\Phi$$

  which combines to give

$$\delta_r'' - \frac{1}{3}\nabla^2\delta_r = \frac{4}{3}\nabla^2\Phi = \text{const}.$$

  Then δ_r oscillates on subhorizon scales with mean $\delta_r = -4\Phi_{\text{MD}}(k)$.
- The acoustic oscillations in the perturbed radiation density give rise to peaks in the spectrum of CMB anisotropies, as we will see below.

Now, we consider the evolution of perturbations in the dark matter density.

- During radiation and matter domination, the Hubble parameter is

$$\mathcal{H}^2 = \frac{\mathcal{H}_0^2\Omega_m^2}{\Omega_r}\left(\frac{1}{y} + \frac{1}{y^2}\right), \quad y = \frac{a}{a_{\text{eq}}}.$$

  The energy continuity and Euler equations for matter are

$$\delta_m' = -\nabla\cdot\mathbf{v}_m, \quad \mathbf{v}_m' = -\mathcal{H}\mathbf{v}_m - \nabla\Phi.$$

  These combine to the equation of motion

$$\delta_m'' + \mathcal{H}\delta_m' = \nabla^2\Phi.$$

- Note that Φ is sourced by both matter and radiation. From our work above, we know that subhorizon radiation density perturbations oscillate on a timescale $\tau \sim 1/k$ during both radiation and matter domination. By contrast, the damping term above ensures that δ_m varies on a timescale $\tau \sim 1/\mathcal{H}$. Hence for subhorizon modes, the δ_r vary rapidly and their effect averages out to zero.

- This means that we may replace the right-hand side above with

$$\nabla^2 \Phi_m = 4\pi G a^2 (\overline{\rho}_m \delta_m - 3\mathcal{H}\overline{\rho}_m v_m)$$

by the Poisson equation. By the continuity equation, the second term is smaller by a factor of $\mathcal{H}^2/k^2$ and hence negligible.

- The equation of motion can be rewritten in terms of y, giving

$$\frac{d^2\delta_m}{dy^2} + \frac{2+3y}{2y(1+y)}\frac{d\delta_m}{dy} - \frac{3}{2y(1+y)}\delta_m = 0$$

which is the Meszaros equations. The solutions take the form

$$\delta_m \propto \begin{cases} 2+3y, \\ (2+3y)\log\dfrac{\sqrt{1+y}+1}{\sqrt{1+y}-1} - 6\sqrt{1+y}. \end{cases}$$

During radiation domination, $y \ll 1$, the second solution is the growing mode, with $\delta_m \propto \log a$. During matter domination, $y \gg 1$, the first solution is the growing mode, with $\delta_m \propto a$. These are precisely the same results we saw earlier with Newtonian perturbation theory.

- Alternatively, during matter domination, we can solve for δ_m given our expression for Φ during matter domination and the Poisson equation. This gives

$$\Delta_m = \frac{\nabla^2\Phi}{4\pi G a^2 \overline{\rho}} \propto \begin{cases} a \\ a^{-3/2} \end{cases}$$

which is just what we saw in the Newtonian treatment, as $\Delta_m \approx \delta_m$ for subhorizon scales. On superhorizon scales, δ_m is constant, but evidently $\Delta_m \propto a$.

- At late times, we need only account for matter and dark energy. However, only matter appears in the Poisson equation since dark energy doesn't have fluctuations,

$$\nabla^2\Phi = 4\pi G a^2 \overline{\rho}_m \Delta_m.$$

The trace part of the Einstein field equation is

$$\Phi'' + 3\mathcal{H}\Phi' + (2\mathcal{H}' + \mathcal{H}^2)\Phi = 4\pi G a^2 \delta P \approx 0$$

since pressure fluctuations are negligible.

- To convert this to an evolution equation for Δ_m, note that the Poisson equation implies $\Phi \propto a^2 \Delta_m \overline{\rho}_m \propto \Delta_m / a$. Plugging this in and simplifying with the Friedmann equations gives

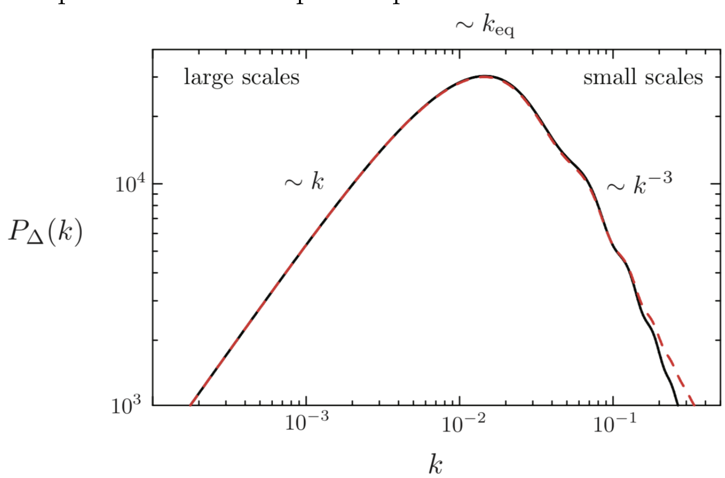
$$\Delta_m'' + \mathcal{H}\Delta_m' - 4\pi G a^2 \overline{\rho}_m \Delta_m = 0.$$

We found a very similar result using Newtonian perturbation theory.

- The simplest way to solve this equation is to work in terms of $u = \Delta_m/H$, in which case
$$\frac{d^2u}{da^2} + 3\frac{d\log(Ha)}{da}\frac{du}{da} = 0.$$
Then the decaying and growing modes are
$$\Delta_m \propto H, \quad \Delta_m \propto H \int \frac{da}{(aH)^3}$$
respectively. During matter domination, we have $\Delta_m \propto a$ again, while Δ_m approaches a constant during dark energy domination. These are consistent with our earlier results, but now also valid for superhorizon scales.

Note. The net effect of all the evolution above is that the primordial perturbations are "post-processed". We can express this in terms of a transfer function,
$$\Delta_{m,\mathbf{k}}(z) = T(k,z)\mathcal{R}_{\mathbf{k}}$$
for the matter fluctuations, where z is the redshift. The observed matter power spectrum is
$$P_\Delta(k,z) = |\Delta_{m,\mathbf{k}}(z)|^2 \sim \begin{cases} k & k \ll k_{\mathrm{eq}}, \\ k^{-3} & k \gg k_{\mathrm{eq}}. \end{cases}$$
To understand the asymptotic behavior, note that the primordial perturbations satisfy $|\mathcal{R}_{\mathbf{k}}|^2 \propto 1/k^3$, as we'll see below. Additionally, as k increases up to k_{eq}, the modes reenter the horizon earlier during matter domination, and hence have more time to grow. **(don't superhorizon modes of Δ_m also grow linearly? are we really talking about δ_m?)** Since subhorizon modes grow as $\Delta_m \propto a$, and $\mathcal{H} \propto 1/\sqrt{a}$, this means the transfer function grows as k^2 in this regime.

Note. A numerical plot of the matter power spectrum is shown below.



The dashed line shows nonlinear corrections. On small scales, we can see baryon acoustic oscillations. The story behind these is as follows. We have seen that dark matter perturbations grow once matter domination begins. However, baryon perturbations cannot grow until decoupling, as until this point they are strongly coupled with the photons via Compton scattering. After decoupling, the baryons intuitively fall into the potential well generated by the dark matter, with δ_b rapidly rising to match the dark matter density contrast δ_c. This results in some oscillatory behavior as a function of k.

5 Initial Conditions From Inflation

5.1 Quantum Fluctuations

Note. **(say something about inhomogeneous initial conditions, these are stretched out and hence effectively negligible in physical coords but not in comoving ones? this is confusing.)**

Now we explain how primordial fluctuations are sourced by the quantum fluctuations of the inflaton. This requires background on quantum field theory in curved spacetime, provided in the notes on General Relativity.

- The key intuition is that the inflaton field ϕ acts as a local "clock" reading off the amount of inflationary expansion still to occur. However, since ϕ is a quantum field, it necessarily has spatially varying fluctuations, which change when inflation ends, causing adiabatic perturbations.
- During inflation, fluctuations are stretched in physical size while the Hubble radius stays the same; equivalently, in terms of comoving coordinates, fluctuations have constant wavelengths, but the comoving Hubble radius shrinks. Hence fluctuations will generally exit the horizon at some point during inflation, and reenter it much later.
- At horizon exit ($k = aH$, for comoving wavenumber k), we match the quantum fluctuation $\langle|\delta\phi_k|^2\rangle$ to a classical **(why classical?)** stochastic field $\langle|\mathcal{R}_k|^2\rangle$, which is then conserved. This conservation law allows us to connect quantities during and long after inflation, despite the large amount of unknown physics at play in between.
- Upon horizon re-entry, we can feed $\langle|\mathcal{R}_k|^2\rangle$ into cosmological perturbation theory, as we did in the previous section. These perturbations are the seeds of large-scale structure, and become imprinted as temperature fluctuations in the CMB, which are measured today.
- The matching at horizon exit is simplest in spatially flat gauge, where

$$\mathcal{R} = -\frac{\mathcal{H}}{\bar{\phi}'}\delta\phi.$$

  The variance of curvature perturbations is therefore

$$\langle|\mathcal{R}_k|^2\rangle = \left(\frac{\mathcal{H}}{\bar{\phi}'}\right)^2 \langle|\delta\phi_k|^2\rangle.$$

  The gauge invariance of $\mathcal{R}$ becomes extremely useful here, as we can switch back to Newtonian gauge to handle horizon re-entry.

To understand the fluctuations $\delta\phi_k$ further, we need to investigate the dynamics of the inflaton field in an FRW background. This is quite similar to what we did in the notes on General Relativity, but in that case we considered a massive field with no vev.

- We start from the inflaton action

$$S = \int d\tau d\mathbf{x}\,\sqrt{-g}\left(\frac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - V(\phi)\right)$$

where the $\mathbf{x}$ are comoving coordinates. We write the perturbed inflaton field as

$$\phi(\tau, \mathbf{x}) = \overline{\phi}(\tau) + \frac{f(\tau, \mathbf{x})}{a(\tau)}.$$

- In order to derive the linearized equation of motion for $f(\tau, \mathbf{x})$, we need to expand the action to second order in it. One complication is that ϕ also affects the geometry. However, in spatially flat gauge, the metric perturbations δg_{00} and δg_{0i} are suppressed relative to the inflaton fluctuations by a factor of the slow-roll parameter ϵ. Hence at leading order in the slow-roll expansion, we may take the metric to be the unperturbed FRW metric, giving

$$S = \int d\tau d\mathbf{x} \left(\frac{1}{2} a^2 ((\phi')^2 - (\nabla \phi)^2) - a^4 V(\phi) \right) \equiv \int d\tau d\mathbf{x}\, \mathcal{L}.$$

- At first order in f, we have

$$\mathcal{L}^{(1)} = a\overline{\phi}' f - a'\overline{\phi}' f - a^3 \partial_\phi V\, f.$$

Integrating the first term by parts, we have

$$\mathcal{L}^{(1)} = -\left(\partial_\tau (a\overline{\phi}') + a'\overline{\phi}' + a^3 \partial_\phi V \right) f.$$

Setting this to zero gives the zeroth order equation of motion for f, i.e. the equation of motion for the background field,

$$\overline{\phi}'' + 2\mathcal{H}\overline{\phi}' + a^2 \partial_\phi V = 0.$$

This is a familiar result, now written in comoving coordinates.

- The terms quadratic in f are

$$\mathcal{L}^{(2)} = \frac{1}{2} \left((f')^2 - (\nabla f)^2 - 2\mathcal{H} f f' + (\mathcal{H}^2 - a^2 \partial_\phi^2 V) f^2 \right).$$

Integrating the $ff' = (f^2)'/2$ term by parts gives

$$\mathcal{L}^{(2)} = \frac{1}{2} \left((f')^2 - (\nabla f)^2 + \left(\frac{a''}{a} - a^2 \partial_\phi^2 V \right) f^2 \right)$$

where we used $\mathcal{H}' = a''/a - \mathcal{H}^2$.

- During slow-roll inflation, we have

$$\frac{\partial_\phi^2 V}{H^2} \approx \frac{3M_{\text{pl}}^2 \partial_\phi^2 V}{V} = 3\eta_v \ll 1.$$

Since $a' = a^2 H$ with H approximately constant,

$$\frac{a''}{a} \approx 2a' H = 2a^2 H^2.$$

- Hence the $\partial_\phi^2 V$ term in $\mathcal{L}^{(2)}$ is negligible, giving

$$\mathcal{L}^{(2)} = \frac{1}{2}\left((f')^2 - (\nabla f)^2 + \frac{a''}{a} f^2\right).$$

The equation of motion for the Fourier modes,

$$f''_{\mathbf{k}} + \left(k^2 - \frac{a''}{a}\right) f_{\mathbf{k}} = 0$$

is called the Mukhanov-Sasaki equation, and is simply the Klein-Gordan equation with a time-varying mass. As expected, on subhorizon scales the effective mass term is negligible.

- If we account for the metric fluctuations, we get the more accurate result

$$f''_{\mathbf{k}} + \left(k^2 - \frac{z''}{z}\right) f_{\mathbf{k}} = 0, \quad z^2 = 2a^2\epsilon$$

which has some dependence on ϵ and its derivatives.

Next, we quantize the field f using the standard techniques.

- The conjugate momentum is $\pi = f'$, and the canonical commutators are

$$[\hat{f}(\tau, \mathbf{x}), \hat{\pi}(\tau, \mathbf{x}')] = i\delta(\mathbf{x} - \mathbf{x}').$$

We use the symmetric Fourier transform convention

$$\hat{f}(\tau, \mathbf{x}) = \int \frac{d\mathbf{k}}{(2\pi)^{3/2}} \hat{f}_{\mathbf{k}}(\tau) e^{i\mathbf{k}\cdot\mathbf{x}}.$$

As a result, we have

$$[\hat{f}_{\mathbf{k}}(\tau), \hat{\pi}_{\mathbf{k}'}(\tau)] = \int \frac{d\mathbf{x}}{(2\pi)^{3/2}} \int \frac{d\mathbf{x}'}{(2\pi)^{3/2}} [\hat{f}(\tau, \mathbf{x}), \hat{\pi}(\tau, \mathbf{x})] e^{-i\mathbf{k}\cdot\mathbf{x}} e^{-i\mathbf{k}'\cdot\mathbf{x}'} = i\delta(\mathbf{k} + \mathbf{k}').$$

- We take the mode expansion

$$\hat{f}_{\mathbf{k}}(\tau) = f_k(\tau)\hat{a}_{\mathbf{k}} + f_k^*(\tau) a_{\mathbf{k}}^\dagger$$

where the functions $f_k(\tau)$ and $f_k^*(\tau)$ satisfy the Mukhanov-Sasaki equation.

- As usual, if we normalize the Wronskian of the mode functions to

$$W[f_k, f_k^*] = -i(f_k \partial_\tau f_k^* - (\partial_\tau f_k) f_k^*) = 1$$

then the creation and annihilation operators satisfy the usual commutation relations,

$$[\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{k}'}^\dagger] = \delta(\mathbf{k} + \mathbf{k}').$$

Note that the conventions here differ slightly from the notes on General Relativity.

- As usual for quantum field theory in curved spacetime, the vacuum state is ambiguous. However, in the case of inflation, where the background is approximately de Sitter, there is a preferred vacuum state called the Bunch-Davies vacuum.

- We note that every mode is a subhorizon mode in the distant past, $\tau \to -\infty$. Furthermore, subhorizon modes do not "feel the curvature", so we may treat them as if they are in Minkowski space. **(what justifies this?)** In this case there is a distinguished set of modes, which are just complex exponentials. We hence define the mode functions in the Bunch-Davies vacuum to satisfy
$$\lim_{\tau\to-\infty} f_k(\tau) = \frac{e^{-ik\tau}}{\sqrt{2k}}$$
where the normalization factor is from the Wronskian.

- In de Sitter space, the Mukhanov-Sasaki equation is
$$f_k'' + \left(k^2 - \frac{2}{\tau^2}\right) f_k = 0$$
which has the exact solution
$$f_k(\tau) = \alpha \frac{e^{-ik\tau}}{\sqrt{2k}}\left(1 - \frac{i}{k\tau}\right) + \beta \frac{e^{ik\tau}}{\sqrt{2k}}\left(1 + \frac{i}{k\tau}\right).$$
For the Bunch-Davies vacuum, we hence have $\alpha = 1$ and $\beta = 0$, giving
$$f_k(\tau) = \frac{e^{-ik\tau}}{\sqrt{2k}}\left(1 - \frac{i}{k\tau}\right).$$

- Finally, we can easily compute the fluctuations of $\hat{f}$ from its mode expansion,
$$\hat{f}(\tau, \mathbf{x}) = \int \frac{d\mathbf{k}}{(2\pi)^{3/2}} (f_k(\tau)\hat{a}_{\mathbf{k}} + f_k^*(\tau) a_{\mathbf{k}}^\dagger) e^{i\mathbf{k}\cdot\mathbf{x}}.$$
Plugging this in, we have
$$\langle|\hat{f}|^2\rangle \equiv \langle 0|\hat{f}^\dagger(\tau, \mathbf{0})\hat{f}(\tau, \mathbf{0})|0\rangle = \int \frac{d\mathbf{k}}{(2\pi)^3} f_k(\tau) f_{k'}^*(\tau) \langle 0|[\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{k}'}^\dagger|0\rangle.$$
Using the commutation relations gives
$$\langle|\hat{f}|^2\rangle = \int \frac{d\mathbf{k}}{(2\pi)^3} |f_k(\tau)|^2 = \int d\log k\, \frac{k^3}{2\pi^2} |f_k(\tau)|^2.$$
Accordingly, we define the dimensionless power spectrum as
$$\Delta_f^2(k, \tau) = \frac{k^3}{2\pi^2} |f_k(\tau)|^2.$$
All other dimensionless power spectra are defined in terms of the Fourier transforms of fluctuations in the same way.

- Scaling back to the original field, using the mode functions, and $\tau = -1/aH$, we have
$$\Delta_{\delta\phi}^2(k, \tau) = a^{-2}\Delta_f^2(k, \tau) = \left(\frac{H}{2\pi}\right)^2 \left(1 + \left(\frac{k}{aH}\right)^2\right).$$

In de Sitter space, this spectrum is time-independent, but in the inflationary universe the Hubble constant slowly changes. We hence evaluate $\Delta^2_{\delta\phi}$ for each mode k at the moment it leaves the horizon, giving

$$\Delta^2_{\delta\phi}(k) \approx \left(\frac{H}{2\pi}\right)^2 \bigg|_{k=aH}.$$

The distribution is approximately Gaussian, with modes of different $\mathbf{k}$ uncorrelated. **(why not a factor of 2?)**

5.2 Primordial Perturbations

We now relate the fluctuations found above to measurable parameters.

- Relating the inflaton fluctuations to the conserved curvature perturbation as described above,

$$\Delta^2_{\mathcal{R}}(k) = \left(\frac{\mathcal{H}}{\bar{\phi}'}\right)^2 \Delta^2_{\delta\phi}(k) = \left(\frac{H}{\dot{\bar{\phi}}}\right)^2 \Delta^2_{\delta\phi}(k) = \frac{1}{2\epsilon}\frac{\Delta^2_{\delta\phi}}{M_{\text{pl}}^2}$$

  where we used $\epsilon = (\dot{\bar{\phi}}^2/2)/M_{\text{pl}}^2 H^2$. Using our previous result and the slow roll approximation,

$$\Delta^2_{\mathcal{R}}(k) = \frac{1}{8\pi^2}\frac{1}{\epsilon}\frac{H^2}{M_{\text{pl}}^2}\bigg|_{k=aH} = \frac{1}{12\pi^2}\frac{V^3}{M_{\text{pl}}^6 (V')^2}$$

  where V is evaluated at the value that $\bar{\phi}$ takes when the mode k leaves the horizon. Since we are only dealing with scalar perturbations, the left-hand side is also called Δ^2_s.

- We hence expect that $\Delta^2_{\mathcal{R}}(k)$ will be roughly independent of k, and hence scale-invariant. There will be small deviations from scale invariance due to the rolling of the inflaton, and hence we expect that near a reference scale k_*,

$$\Delta^2_{\mathcal{R}}(k) \approx A_s \left(\frac{k}{k_*}\right)^{n_s - 1}, \quad n_s - 1 = \frac{d\log\Delta^2_{\mathcal{R}}}{d\log k}.$$

- We may compute $n_s - 1$ in terms of the slow roll parameters by noting that

$$\frac{d\log\Delta^2_{\mathcal{R}}}{d\log k} = \frac{d\log\Delta^2_{\mathcal{R}}}{dN}\frac{dN}{d\log k} = \left(2\frac{d\log H}{dN} - \frac{d\log\epsilon}{dN}\right)\frac{dN}{d\log k}.$$

  The quantity in brackets is simply $-2\epsilon - \eta$. For the second term, note that the horizon crossing condition $k = aH$ gives

$$\log k = N + \log H$$

  so that

$$\frac{dN}{d\log k} = \left(\frac{d\log k}{dN}\right)^{-1} = \left(1 + \frac{d\log H}{dN}\right)^{-1} = (1-\epsilon)^{-1}.$$

- Expanding to first order in the slow roll parameters, we hence have

$$n_s - 1 \approx -2\epsilon - \eta.$$

  Alternative, we may write this in terms of the inflaton potential in the slow roll approximation,

$$n_s - 1 \approx -3M_{\text{pl}}^2\left(\frac{V'}{V}\right)^2 + 2M_{\text{pl}}^2\frac{V''}{V}.$$

- We can also consider tensor perturbations to the metric,

$$ds^2 = a^2(\tau)\left(d\tau^2 - (\delta_{ij} + 2\hat{E}_{ij})dx^i dx^j\right).$$

This gives a second-order variation of

$$\mathcal{L}^{(2)} = \frac{M_{\text{pl}}^2}{8} a^2 \left((\hat{E}'_{ij})^2 - (\nabla \hat{E}_{ij})^2\right).$$

- For concreteness, consider fluctuations with $\mathbf{k} \propto \hat{\mathbf{z}}$. Then $\hat{E}_{ij}$ can be expanded as

$$\frac{M_{\text{pl}}}{2} a\hat{E}_{ij} = \frac{1}{\sqrt{2}} \begin{pmatrix} f_+ & f_\times & 0 \\ f_\times & -f_+ & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Then the second-order Lagrangian is

$$\mathcal{L}^{(2)} = \frac{1}{2} \sum_{I=+,\times} (f_I')^2 - (\nabla f_I)^2 + \frac{a''}{a} f_I^2.$$

This is simply two copies of the Lagrangian found for $f = a\delta\phi$ above.

- Therefore, we can simply reuse the result to compute the power spectrum of tensor modes,

$$\Delta_t^2 \equiv 2\Delta_{\hat{E}}^2 = 2\left(\frac{2}{aM_{\text{pl}}}\right)^2 \Delta_f^2 = \frac{2}{\pi^2} \frac{H^2}{M_{\text{pl}}^2}\bigg|_{k=aH}.$$

Note that unlike the scalar modes, we didn't have to convert to $\mathcal{R}$ and convert back, because $\hat{E}_{ij}$ is already in the form of a metric perturbation, which is conserved on superhorizon scales. **(right?)** Tensor modes are a very generic prediction of inflation, and measuring their amplitude would give direct information about the inflationary energy scale H.

- We define the scale-dependence of the tensor spectrum by

$$\Delta_t^2(k) = A_t \left(\frac{k}{k_*}\right)^{n_t}$$

where scale-invariance corresponds to $n_t = 0$, and define the tensor to scalar ratio $r = A_t/A_s$. In the slow roll approximation, we have

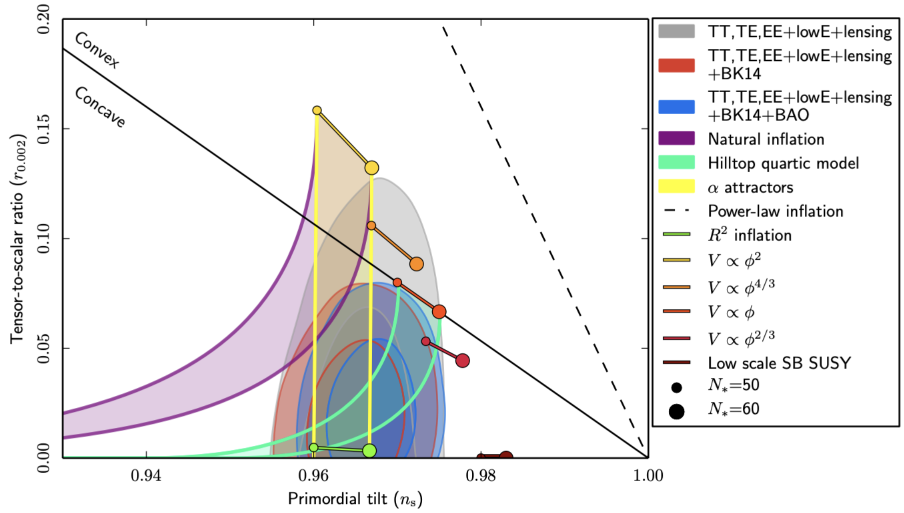
$$r = 16\epsilon, \quad n_t = -2\epsilon.$$

In particular, the ratio $r/n_t = -8$ is independent of the slow roll parameters, so its value when measured can provide a consistency check.

- The latest observations of the CMB power spectrum from the Planck satellite indicate

$$A_s = (2.196 \pm 0.060) \times 10^{-9}, k_* = 0.05\,\text{Mpc}^{-1}, \quad n_s = 0.9649 \pm 0.0042.$$

The results are shown graphically below.



The observed magnitude of the scalar perturbations points to an inflationary energy scale of $H \sim \epsilon^{1/4} 10^{16}\,\text{GeV}$.

- Tensor perturbations are detected in the CMB by the pattern of light polarization. Such a pattern can be decomposed into "E-modes" with vanishing curl, and "B-modes" with curl, named in analogy to electrostatic and magnetostatic fields; tensor perturbations result in B-modes. These have not been detected; instead the tensor-to-scalar ratio is bounded by

$$r \lesssim 0.07.$$

We can also check the spectrum of scalar perturbations by evolving them in time with cosmological perturbation theory, and measuring the matter power spectrum.

Note. A collection of identities to perform the calculations above. The equations of motion are

$$H^2 = \frac{V}{3M_{\text{pl}}^2}, \quad 3H\dot{\phi} \approx -V'.$$

The slow-roll parameters are

$$\epsilon = -\frac{\dot{H}}{H^2} = -\frac{d\log H}{dN} = \frac{\dot{\phi}^2/2}{M_{\text{pl}}^2 H^2}, \quad \eta = \frac{d\log\epsilon}{dN} = \frac{\dot{\epsilon}}{H\epsilon}, \quad \delta = -\frac{\ddot{\phi}}{H\dot{\phi}}, \quad \eta = 2(\epsilon - \delta).$$

There are also parameters defined in terms of the potential,

$$\epsilon_v = \frac{M_{\text{pl}}^2}{2}\left(\frac{V'}{V}\right)^2, \quad \eta_v = M_{\text{pl}}^2 \frac{V''}{V}$$

which are related by

$$\epsilon_v \approx \epsilon, \quad \eta_v \approx \delta + \epsilon = 2\epsilon - \frac{\eta}{2}.$$

In order to convert to conformal time, we have

$$d\tau = \frac{dt}{a}, \quad \mathcal{H} = aH, \quad a' = a^2 H.$$

Second derivatives are handled by

$$\dot{H} = \frac{\ddot{a}}{a} - H^2, \quad \mathcal{H}' = \frac{a''}{a} - \mathcal{H}^2.$$

For a de Sitter universe we have constant H, so

$$a(t) = e^{Ht}, \quad \tau = -\frac{1}{Ha} = -\frac{1}{\mathcal{H}}.$$

Expressed in terms of slow-roll parameters, we have

$$\epsilon = 1 - \frac{\mathcal{H}'}{\mathcal{H}^2}, \quad \eta = \frac{\epsilon'}{\mathcal{H}\epsilon}.$$

6 Observational Data

6.1 Dark Matter

We now consider the evidence for dark matter. One of the most straightforward pieces of evidence comes from galaxy rotation curves.

- In typical galaxies, we infer from the galactic luminosity distribution that the mass is strongly concentrated at the center. Therefore, away from the center, stellar velocities should fall off as $v \propto 1/\sqrt{r}$.
- We can measure the function $v(r)$ by measuring redshifts in galaxies tilted relative to us. The result is that $v(r)$ is flat at high distances, or can even increase. Historically, such observations were made in 1939 for the Andromeda galaxy and extended to larger radii in the 1970s by Rubin and Ford. Later, similar measurements were done at radii well beyond the visible disk by measuring 21 cm line emission.
- The curves can be fit by assuming a spherically symmetric dark matter halo whose density falls off much slower than the baryonic mass density. The density functions $\rho(r)$ can be computed by numerical simulations; at the simplest level, they simply solve the collisionless Boltzmann equation. These simulations favor a "Buckert" profile
$$\rho(r) = \frac{\rho_0}{(1+\tilde{r})(1+\tilde{r}^2)}, \quad \tilde{r} = \frac{r}{r_s}.$$
There are many other density profiles used for modeling, such as the Navarro-Frenk-White, Einasto, and Hernquist profiles.
- We can also infer our local dark matter density by measuring the local vertical distribution of stars; however, this method is quite noisy and gives a value that could be consistent with zero.
- There are various problems found when modeling galaxies numerically; for example, the halo density profile is generically predicted to be "cuspy", but observations indicate "cored" density profiles. This could be explained by the "feedback effect" of baryons, or by including dark matter self-interaction; such work is currently in progress.

Note. MOND is an alternative hypothesis that explains galaxy rotation curves well. In MOND, the gravitational acceleration due to a mass takes the limiting value
$$\lim_{r\to\infty} g = \sqrt{\frac{GMa_0}{r^2}}$$
for a constant a_0. This explains flat rotation curves, because we expect
$$\frac{v^2}{r} = \sqrt{\frac{GMa_0}{r^2}}$$
at large distances, and the factors of r cancel. It also provides a direct explanation of the Tully-Fisher relation $L \propto v^4$, assuming that $M \propto L$. There are several options for g at intermediate values of r which can be adjusted to fit observations at intermediate radii. In general, MOND seems to do a better job of fitting rotation curves, requiring fewer parameters than dark matter, but completely fails to match evidence from galaxy clusters and the CMB. Some hybrid models containing both MOND and dark matter have been proposed to fit all such observations.

Evidence for dark matter also comes from observing clusters of galaxies.

- Historically, this was the first evidence for dark matter, proposed in 1933 by Fritz Zwicky to explain observations of the Coma cluster.
- Regarding each galaxy as a point mass, we can show the 'virial theorem'

$$\ddot{I} = 4T + 2U$$

  where I is the moment of inertia, T is the total kinetic energy, and U is the total potential energy. This is shown by directly expanding the expression for $\ddot{I}$. In the steady state, we have

$$T = -\frac{U}{2}$$

  which allows us to estimate the mass of a cluster of galaxies from the observed $\langle v^2 \rangle$.
- In practice, we can estimate $\langle v_\perp^2 \rangle$ from redshifts, and multiple by 3 assuming spherical symmetry. We also assume $\rho(r)$ is proportional to the luminosity $L(r)$, allowing us to estimate the potential energy. Combining these gives a total mass much higher than the stellar mass.
- Dark matter is also responsible for presence of the hot intracluster gas, which emits X-rays; if dark matter were not present, most of the gas would be gone. Assuming the gas is in hydrostatic equilibrium and obeys the ideal gas law,

$$\frac{dP}{dr} = -\frac{GM(r)\rho(r)}{r^2}, \quad P = \frac{\rho kT}{m}$$

  we find that

$$M(r) = \frac{kT(r)r}{Gm}\left(-\frac{d\log(\rho T)}{d\log r}\right)$$

  so that $M(r)$ can be inferred from the functions $T(r)$ and $\rho(r)$, which are measured by X-ray emission. This was done in the 1990s and gave a dark matter amount consistent with the above.

We can also see evidence for dark matter through gravitational lensing.

- Baryonic dark matter can come in the form of MACHOs (massive compact halo objects), i.e. stellar remnants and brown dwarfs. When a MACHO passes exactly between the Earth and a distant star, we see a bright ring around the star. If the object is slightly off center, we instead get a bright arc.
- In practice, these arcs and rings are small enough that we can only detect a momentary brightening of the star. Surveying the sky for these brightening events gives us an estimate of the number of MACHOs in the galaxy. The result is a small fraction of the dark matter density.
- Galaxy clusters also lens light; since the angle is proportional to mass, we can measure the cluster masses, getting results in agreement with the observations above.
- Gravitational lensing was also used to analyze the Bullet cluster, which was formed by the collision of two large galaxy clusters. It indicated that the visible matter, as measured by X-ray spectroscopy, lagged behind most of the gravitational mass, indicating the presence of weakly-interacting dark matter. This is regarded by some as the "nail in the coffin" for MOND.

Finally, cosmological observations give us a "direct count" of matter.

- We can infer the mass density of stars by measuring luminosities, and assuming that a large population of stars has the same mass-to-luminosity ratio as a typical population near us. Additional mass comes from nonluminous stellar remnants, such as white dwarfs, neutron stars, and black holes, as well as brown dwarfs.
- Galaxies also contain significant amounts of interstellar gas, which accounts for 10% of the mass of the Milky Way. For rich clusters of galaxies, the mass of the intergalactic gas can be larger than the mass in the galaxies. This gas is extremely hot and is detected by X-rays using the Sunyaev–Zel'dovich effect.
- By combining all these measurements, a total baryonic density of $\Omega_b = 0.05$ is found, while redshift measurements in the concordance model indicate Ω_m is much larger; in the model the difference is accounted for by dark matter.
- Stringent independent constraints on Ω_b come from the acoustic peaks in the CMB, and from the results of BBN, which both yield the same number.
- The dark matter density also plays an important role in structure formation in the early universe, as we've seen above. Measurements of the matter power spectrum yield another independent, consistent value for Ω_m.
- Various modified gravity theories which reduce to MOND in the nonrelativistic limit, such as TeVeS, have also been applied to cosmological scales. However, they currently do a much worse job of fitting the observations than the standard cosmological model.

We now summarize the little that is known about dark matter.

- Dark matter candidates can be divided into "cold" and "hot", depending on their mean velocities. We know most, if not all dark matter must be cold, as an excess of hot dark matter would tend to smooth out small mass fluctuations, suppressing structure formation.
- If it's assumed that dark matter is in thermal equilibrium with the SM thermal bath, then whether dark matter is cold or hot simply depends on its mass; lighter particles would be hot.
- The interactions of dark matter are so far consistent with being purely gravitational. Even dark matter self-interactions must be weak, as otherwise they would ruin the stability of the halo. A small amount of decay or annihilation is permitted, which could be used to explain some experimental anomalies. Interaction effects may also increase the accuracy of numerical simulations of galaxies.
- Dark matter is almost certainly electrically neutral. Even "millicharged" DM is strongly constrained by cosmology.

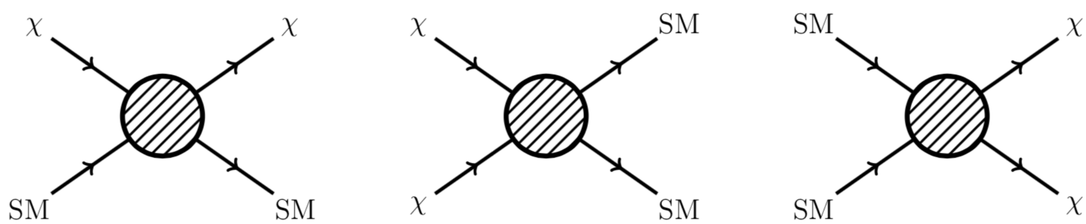
We now briefly consider some dark matter candidates.

- The lightest stable supersymmetric particle, typically a neutralino, could be cold dark matter; it is the canonical example of a WIMP. Many experiments have tried to directly detect WIMPs recoiling on nuclei by the weak force, and have now ruled out most of the relevant parameter space.

- Neutrinos could provide the correct density if their mass was several eV and they come into thermal equilibrium, but they would have to be hot dark matter and hence supplanted with an additional cold component. While we don't know the masses of the three known neutrinos, they are certainly much less than an eV. However, there could be sterile neutrinos at this mass.
- Sterile neutrinos that do not come into thermal equilibrium could account for the dark matter density if their mass is on the keV scale. Resonant scattering off particles of this mass scale has been used as an explanation for a deficit of 3.5 keV photons emitted from galaxy centers.
- Axions are extremely light, but also count as cold dark matter because their production mechanism is different. It is essentially in a BEC-like state, with very low velocity dispersion.

Next we consider dark matter detection.

- The three main paradigms for dark matter detection are shown below.



(a) Direct detection (b) Indirect detection (c) Collider production

The amplitudes for all three are clearly related by crossing symmetry.

- In a direct detection experiment, one watches for recoils of SM particles (e.g. atomic nuclei) off dark matter particles. Such experiments are placed deep underground to shield cosmic rays.
- Many such experiments have been done, placing stringent bounds on WIMP cross sections. At this point, the generic 'WIMP miracle' has been ruled out. While TeV scale dark matter can still exist, there must be some tuning to avoid the direct detection bounds.
- A single experiment, the DAMA collaboration, has seen a significant annual modulation in scattering events, which could be due to the seasonal variation of the Earth's velocity through the dark matter halo. However, the results are not taken seriously due to possible systematic effects. Since DAMA uses a different material than other experiments, it can be consistent with other experiments if the dark matter couping is strongly spin-dependent.
- Astrophysicists use indirect detection methods, searching for the produces of annihilating dark matter. Such signatures include an excess of antimatter, mono-energetic gamma rays, and high-energy neutrinos.
- Historically many indirect detection experiments have reported anomalies which could be explained by dark matter. However, such experiments are often 'direction blind', so the excesses can also come from unknown astronomical objects in a very large volume, such as an undiscovered population of pulsars.
- TeV scale dark matter could also be produced at the LHC, with the classic signal being 'missing' energy or momentum.

6.2 The CMB

We now consider the CMB more closely.

- It is difficult to measure the full CMB spectrum on the ground due to absorption by water molecules in the atmosphere. The first accurate measurement over all wavelengths was performed by the COBE satellite in 1989, which found that the CMB's spectrum in all directions was very close to that of an ideal blackbody.
- The CMB's temperature was found to be anisotropic due to the Earth's peculiar velocity. By subtracting off Earth's velocity about the Sun, the Sun's velocity in the galaxy, and the galaxy's motion velocity relative to the Local Group, we can measure the peculiar velocity of the Local Group; it is moving quickly towards the nearest supercluster.
- Subtracting off this dipole distortion gives much smaller fluctuations, with

  $$\langle T \rangle = 2.725\,\mathrm{K}, \quad (\delta T/T)_{\mathrm{rms}} = 1.1 \times 10^{-5}.$$

  Thus the CMB is very nearly isotropic, providing strong evidence for the Big Bang.