Lecture Notes on

Cosmology

Kevin Zhou

kzhou7@gmail.com

These notes cover introductory cosmology, along with a brief overview of dark matter. The primary sources were:

- Daniel Baumann's Cosmology lecture notes. An exceptionally clean and clear set of notes, used at both Cambridge and Oxford. Has a theoretical bias, with little contact with experimentally measured quantities. These lecture notes closely follow Baumann's.
- David Tong's Cosmology lecture notes. Another clear set of notes, at a more basic level. At Cambridge, Baumann's notes are used for Part III while Tong's are used for Part II.
- Subir Sarkar's Astroparticle Physics lecture notes. A broad, up-to-date overview of many current observational issues in astrophysics and cosmology.
- Peter Graham's Physics 362 course. There are no public lecture notes, but some of the wisdom of the course, especially for alternative scenarios and quick estimates, is baked into these notes.
- Ryden, *Introduction to Cosmology*. A well-written undergraduate-level introduction to cosmology, assuming very little background; a bit vague, but good for a first pass to get 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, with a deft hand.
- Kolb and Turner, *The Early Universe*. A classic and clear textbook on cosmology and astroparticle physics, aimed at particle physicists. Contains extensive discussion of topics like GUT baryogenesis and 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 However, many observational statements are out of date.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,

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

  where today h has been measured from CMB observations to be

$$h = 0.67 \pm 0.01.$$

  However, there is an outstanding 3σ discrepancy with more nearby supernova measurements, which give $h = 0.73 \pm 0.02$.

- In the nonrelativistic limit $z \ll 1$, Hubble's law 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.

- Hubble's law can also be explained by a "steady state" model, once advocated by Bondi and Gold. This model assumes the "perfect" cosmological principle, which imposes 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.

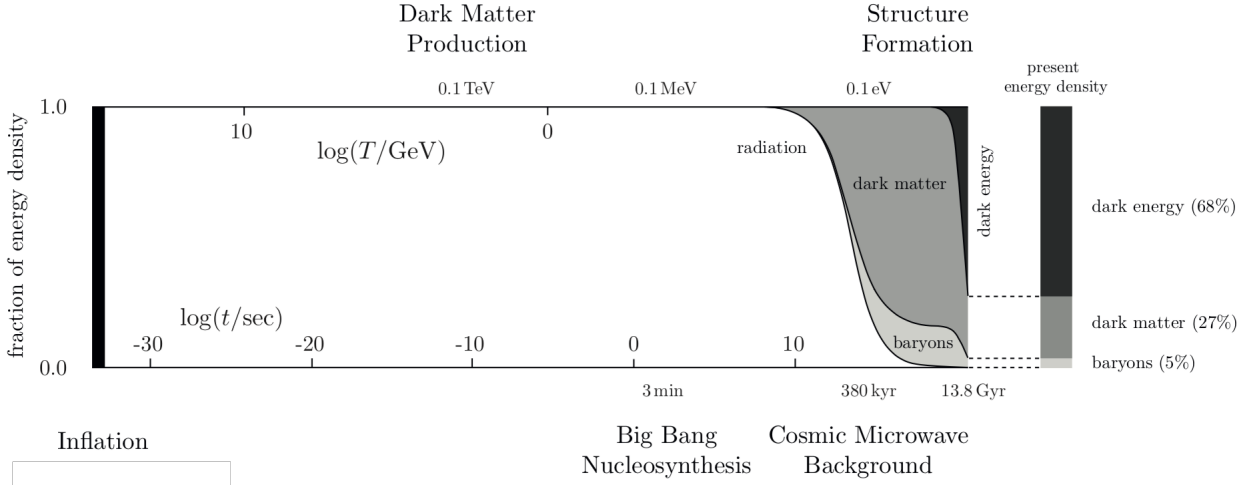
A slick method is to consider consecutive wavecrests of light. Consider two wavecrests emitted at times t_1 and $t_1 + \delta t_1$, and absorbed at times t_0 and $t_0 + \delta t_0$. Since null paths in the radial direction obey $dr/dt = 1/a$,

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

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.

Note. Is the universe infinite in spatial extent? The naive answer is that it is if $k = 0$ or $k = -1$. The correct answer is that we have no idea. For example, an infinite $k = 0$ spatial universe is observationally equivalent to a finite $k = 0$ universe with nontrivial topology such as a torus, as long as the torus is very large. It is possible to test for finiteness by looking at the sky in opposite directions and trying to see repeated structures in the CMB, but it's not possible to rule out finiteness. Since an infinite FRW universe fits all the observations, and is mathematically simpler than having, e.g. a torus structure (which would break isotropy), we'll just use it as a default.

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 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.

- We will use the slick form of the geodesic equation,

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

  which will allow us to formally handle the massless and massive cases simultaneously. Here, the partial derivative on the left only makes sense if we imagine p^μ (which is originally defined only on the geodesic) is extended to a full vector field. It turns out this is always possible, and moreover that the result does not depend on the extension. Now, since the FRW spacetime is homogeneous, it is possible to choose the vector field so that $\partial_i p^\mu = 0$, giving

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

  This trick allows us to sidestep annoying issues with, e.g. parametrization of massless geodesics.

- For massive particles at rest in the comoving frame, $p^i = 0$, we have $dp^\mu/dt = 0$. For moving particles, consider the $\mu = 0$ component,

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

  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, $T \propto \sigma_v^2 \propto 1/a^2$.

Note. Introductory textbooks often go out of their way to stress that the cosmological redshift is *not* a Doppler shift. But the fact is that there's no real distinction between the two. In general, you can always use the geodesic equation to relate the frequencies of a photon as measured by the emitter and absorber, and this reproduces the Doppler shift, the cosmological redshift, and gravitational redshift in appropriate special cases. The reason textbooks make a distinction is because "Doppler shift" usually implies working in flat spacetime, which could get people confused. But cosmological redshift can also be calculated by combining the infinitesimal Doppler shifts as light travels through a set of overlapping patches, each of which are essentially flat. Sometimes students wonder if you have to *add* the effects of Doppler shift and cosmological redshift, but that doesn't make sense; there's only one redshift effect.

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 relation 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,

$$T^\mu{}_\nu = \text{diag}(\rho, -p, -p, -p).$$

- 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.

- An additional component with $w < -1/3$, known as dark energy, is required to account for the accelerating expansion of the universe; measurements indicate that $w = -1.03 \pm 0.03$. A cosmological constant has $w = -1$, and hence a constant energy density $\rho \propto a^0$, and it may arise in QFT from vacuum energy density, which is not diluted by expansion. Other models of dark energy, such as slowly rolling scalar fields, can have $w \approx -1$, but for concreteness we'll take $w = -1$ exactly.

- 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

$$\gamma_{ij} = \delta_{ij} + \frac{kx_ix_j}{1 - k(x_kx^k)} = \delta_{ij} + kx_ix_j + O(x^4)$$

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 the context of the Raychaudhuri equation in general relativity, where energy density and pressure cause geodesics to contract. The first Friedmann equation is also called *the* Friedmann equation.

Note. A Newtonian 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.

How could this approach work, when the expansion of space is a relativistic effect? In patches over which the curvature is negligible, Newtonian gravity gives the same results as relativity. And because the expansion of the universe is homogeneous, we can take a patch of any size, so we simply work in one sufficiently small for Newtonian gravity to hold. Of course, the derivation only works for matter-domination; radiation and dark energy have no natural Newtonian analogue.

In the 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.) Roughly speaking, the density determines the amount of potential energy, the current Hubble constant determines the amount of kinetic energy, and the critical density is where the total energy is zero.

Note. Why does the expansion of the universe not expand ordinary matter, such as people, planets, or galaxy clusters? The usual answer is that the expansion of the universe is a force pulling everything apart, but objects can resist this force by attraction. However, this is misleading.

To see why, consider the case of a matter-dominated universe. In the FRW metric, it doesn't make sense to say the matter particles are being pulled apart, because they simply follow geodesics, which experience no force by definition. Thus, any statement about forces must implicitly be in the Newtonian picture. In the Newtonian picture explained above, which is valid on patches much smaller than the curvature scale, the expansion of the universe is simply due to an initial outward *velocity*, and does not correspond to a force. If you hollow out a sphere in an expanding Newtonian universe, and put two objects inside it at rest with respect to each other, they'll stay that way forever. (Of course, the same result can be obtained in a fully relativistic calculation: hollowing out a sphere changes the metric, which changes the geodesics. It's just that showing this explicitly would be totally intractable.)

However, the usual answer is correct when dark energy is present. In the Newtonian picture, a positive dark energy density has the same effect as a uniform, omnipresent *negative* matter density. This creates a gravitational field that directly repels matter particles away from each other, and which must be balanced by an attractive force.

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

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

  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. The matter-dominated case is known as the Einstein–de Sitter universe.
- 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 χ_{ph} is dominated by its upper bound. Explicitly, we have

$$\chi_{\text{ph}}(a) = \frac{2H_0^{-1}}{1+3w}\left(a^{(1+3w)/2} - a_i^{(1+3w)/w}\right) \equiv \tau - \tau_i$$

  where $\tau_i = 0$ since $a_i = 0$, giving

$$\chi_{\text{ph}} \sim (aH)^{-1}$$

  up to $O(1)$ prefactors; the latter is sometimes called the "Hubble horizon".

- This leads to a problem, because then χ_{ph} is too small for different regions of the CMB to have ever been in causal contact. To estimate the number of causal patches, note that

$$\frac{\chi_{\text{ph}}(a_{\text{CMB}})}{\chi_{\text{ph}}(a_0)} \approx \sqrt{\frac{a_{\text{CMB}}}{a_0}} \approx \sqrt{1100}$$

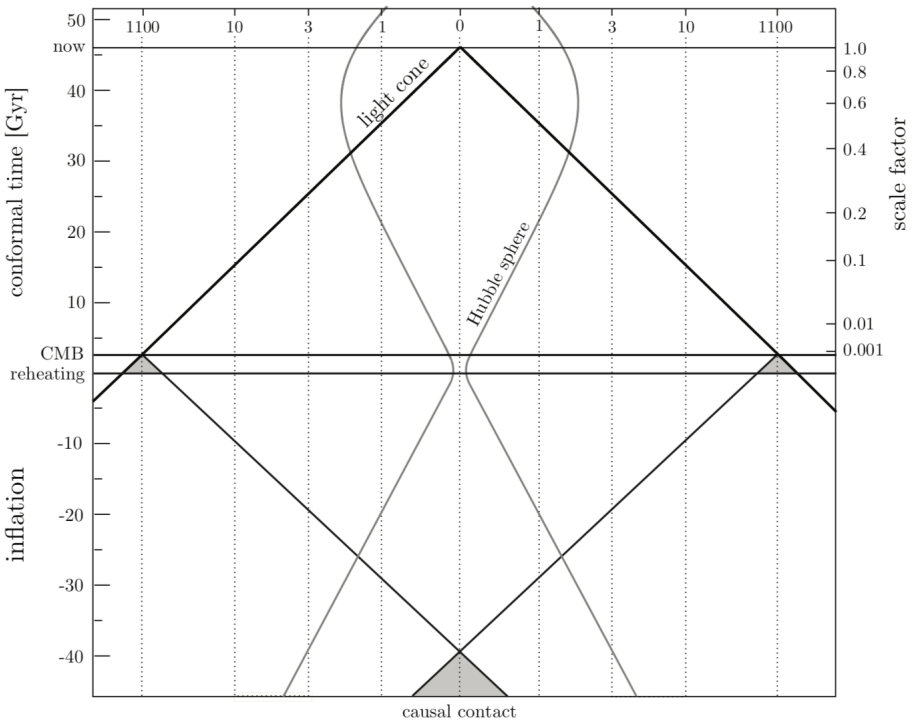
  where we used the fact that the universe has been matter dominated since CMB formation. Then the angular extent of a patch is $1/\sqrt{1100} \approx 1.7°$, giving about 10^4 causally disconnected patches. The horizon problem asks why they "know" to be at the same temperature. This problem is made even worse by modern CMB observations, which indicate correlations between points in the CMB at scales much larger than 1.7°.

These problems can be addressed by a shrinking Hubble sphere.

- The essence of the problem is that we want to set $\chi_{\text{ph}} \gg (aH)^{-1}$, 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 slightly weaker than $(a_I H_I)^{-1} > \chi_{\rm ph}$, 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 assumed that after inflation ends, we reheat to a high (e.g. GUT) scale,

$$T_E \sim 10^{15}\,\text{GeV}, \quad T_0 = 10^{-3}\,\text{eV} = 2.7\,\text{K}.$$

- 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

$$N_{\rm tot} > \log 10^{28} = 64$$

e-folds during inflation. A more accurate accurate comes from the largest scales observed in the CMB, which have to be created $N_{\rm cmb} = 60$ e-folds before the end of inflation.

- This leads to the standard statement that inflation must last for at least 60 e-folds. However, many assumptions go into this result. We could get less strict bounds by assuming we reheat to substantially lower than T_E, or, e.g. including a period of early matter domination. But 60 is a standard benchmark because the simplest models of inflation end up satisfying the standard assumptions.
- Inflation can last for much longer than 60 e-folds, and indeed does in many simple models, but it is the last 60 e-folds which create the cosmological perturbations we see today. (More precisely, the scales we can observe in the CMB correspond to perturbations created around 50 to 60 e-folds before the end of inflation.)

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.

Note. Inflation doesn't solve the horizon problem by letting the thermal plasma reach equilibrium during it. In most models of inflation, there isn't a thermal plasma present at all during inflation; there's just the energy in the inflaton field. (However, "warm inflation" does have such a plasma, which can significantly change the dynamics of the inflaton.) Inflation solves the horizon problem by making the inflaton field itself homogeneous, which then automatically produces a homogeneous plasma during reheating.

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. In most inflationary models we have $w \approx -1$, which is needed to recover the observed primordial perturbations.

- 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. (If kinetic energy dominates, we have $w = 1$, and an energy density redshifting as a^{-6}. In homage to inflation, a period of kinetic energy domination is called "kination".)

- Substituting this into the Friedmann equations and defining the (reduced) Planck mass

$$M_{\rm pl} = \sqrt{\hbar c/8\pi G} = 2.4 \times 10^{18}\,{\rm GeV/c^2},$$

and working in units where $\hbar = c = 1$, we have

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

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

- Next, differentiating the first equation gives

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

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

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

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

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

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

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

The initial $\dot{\phi}$ might not obey this, but it approximately will within $O(1)$ Hubble time.

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

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

Differentiating the equation of motion gives

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

which implies that

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

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

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

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

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

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

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

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

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

so inflation in this model requires super-Planckian values for the inflaton; one can show this also holds for power law potentials $V \propto \phi^p$. The number of e-folds is

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

and the fluctuations observed in the CMB are created at $\phi_{\rm CMB} \approx 2\sqrt{N_{\rm cmb}}M_{\rm pl} \sim 15M_{\rm pl}$.

We require the energy density to not be near the Planck scale, or else unknown quantum gravity effects will become relevant; this yields the constraint $m \ll M_{\rm pl}/15$. On the other hand, we always get super-Planckian field values, which can also be argued to be problematic from a UV perspective. For example, in the paradigm of effective field theory, one would expect additional Planck-suppressed operators to appear, which would significantly change the potential, e.g. increasing the inflaton's mass. This is called the "eta problem", and is analogous to the hierarchy problem for the Higgs. Another related problem occurs for ultralight axions, which receive quantum gravitational corrections to their potential.

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 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 strong 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, having to deal with initial conditions. The flatness problem is closely related to the cosmological constant problem and has a similar anthropic solution. If one doesn't like naturalness problems, then the most compelling argument for inflation today is that it predicts the primordial perturbations observed in our universe, as we will show later.

Note. We've been careful to avoid conflating dark energy, vacuum energy, and the inflaton. Dark energy is a hypothetical substance with $w \approx -1$ that drives the current accelerating expansion of the universe. Vacuum energy is a natural candidate for dark energy, and has $w = -1$ exactly. However, there are also dynamical models for dark energy (such as 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, w would deviate slightly from -1. Currently, 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. The field may tunnel through the barrier, leading to the growth of "bubbles" of true vacuum. If bubble nucleation is too rare, then inflation is eternal, because the space in between the bubbles continues to grow exponentially. If the nucleation rate is high enough, the bubbles eventually percolate through the universe, in analogy to how ice crystals grow through a freezing liquid. However, when the bubbles collide, they produce strong inhomogeneities that undo the solution to the horizon problem. This is the "graceful exit problem".
- Another way of saying this is that to achieve graceful exit, one needs a "clock" measuring the "amount of inflation left" throughout the universe. Yet another way is that, without a clock, one has perfect de Sitter space – and as we saw earlier, perfect de Sitter space has foliations where it is shrinking, not expanding! Without a clock, one can't explain expansion at all.

- 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 near $\varphi = 0$. The uniform field value functions as the clock.

- 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 postulating "chaotic" initial conditions with Planckian energy densities. For example, we assume

$$V(\phi) \sim M_{\rm pl}^4, \quad \dot{\varphi}^2 \sim (\nabla\varphi)^2 \sim M_{\rm pl}^4, \quad R \sim M_{\rm pl}^2$$

  where R is the Ricci scalar. This is the default way of thinking in inflation today.

- Depending on who one asks, there might be an "initial patch problem", i.e. one needs a Planckian patch which is already sufficiently homogeneous, so that it can get started inflating before collapsing. But assuming this process can begin, inflation could then the patch more homogeneous, extracting a set of homogeneous universes from the chaotic initial conditions. The dynamics of this murky initial period were divined by "quantum cosmologists", a secretive sect which contemplated deep metaphysical notions such as "the wavefunction of the universe".

- 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.

- A more recent distinction is between "small-field" and "large-field" inflation, depending on whether the field displacement of the instanton is less than or more than the Planck mass. New inflation and chaotic inflation are small-field and large-field respectively. Large-field inflation predicts larger, possibly soon observable tensor perturbations, but the inflaton potentials could be unstable against quantum gravitational corrections.

- One elegant family of inflation models is natural inflation, where the potential takes the form

$$V(\phi) = V_0 \left(1 + \cos\frac{\phi}{f}\right).$$

  This often arises if the inflaton is taken to be an axion. This is useful in large-field inflation because the shift symmetry of the axion can be used to protect the potential from corrections.

- Small-field inflation is a bit trickier, since for a single inflaton field, one needs to reliably produce a flat part of the potential. However, it is straightforward to realize small-field inflation using multiple fields. For example, in "hybrid inflation", there is an inflaton ϕ and a "waterfall" field σ, and the potential is such that σ is still while ϕ rolls, until it suddenly begins to roll rapidly.

- 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,

$$\Delta\phi \sim \frac{\dot{\phi}}{H} \sim \frac{V'}{H^2} \sim M_{\rm pl}^2 \frac{V'}{V} \sim \frac{M_{\rm pl}^2}{\phi}.$$

  On the other hand, the scale of quantum fluctuations is

$$|\delta\phi| \sim \frac{H}{2\pi} \sim \frac{m\phi}{M_{\rm pl}}.$$

  Therefore, the quantum fluctuations dominate when

$$\phi > M_{\rm pl}\sqrt{\frac{M_{\rm pl}}{m}}.$$

  Or, phrased in terms of the potential and H, this occurs when $V' \lesssim H^3$.

- We see that we can avoid eternal inflation if we simply take m small and postulate an appropriate initial condition. However, if we take chaotic initial conditions with Planckian energy densities, the inequality above is automatically satisfied, leading to eternal inflation. Depending on taste, one can ensure eternal inflation to facilitate anthropics or forbid it.

- One could say that eternal inflation isn't a problem because observable quantities depend on only the last 60 e-folds of inflation. However, the "measure problem" creates issues in calculating probabilities, since all denominators are infinite. The standard way around this in science is to declare a cutoff and compute probabilities within it, and then take the limit where the cutoff is removed, but this procedure is ambiguous. 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, so predictions are 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 $n_s \approx 1$, where r varies strongly between models.

- 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! However, these models are generally quite complicated; we'll focus on the simpler models.

- 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 alternatives to inflation, such as cyclic cosmology, which involve a "big bounce". All such theories run into the "singularity problem". They cannot calculate what will happen at a cosmological singularity, where nonperturbative quantum gravity can play a role; instead they must use conjecture. One can have as many cosmological theories as one has conjectures about Planck-scale physics. To be fair, of course, one can also apply this criticism to chaotic/eternal inflation. Quantum gravity is a contentious subject.

3 Thermal History

3.1 The Hot Big Bang

We begin with a basic overview of the first three minutes of the universe. First, we supply some useful general ideas.

- 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. Note that almost all cross sections in cosmology are thought of in terms of σv, since they always appear together in this rate.

- An important point is that the timescale t_c doesn't depend much on the current conditions. For example, consider a process $\chi + \chi \leftrightarrow$ (particles in equilibrium). If the rate of this reaction forward and backward in equilibrium is Γ, then if χ initially is overabundant by a factor of N, then $t_c \sim (\log N)/\Gamma$, because when N is high, the forward process occurs faster. This is why particles have a chance of remaining in equilibrium even as their equilibrium abundance begins to fall exponentially.

- We focus on the case $T \gtrsim 100\,\text{GeV}$, 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

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

  where we used $\alpha \gtrsim 0.01$, valid for charged or strongly interacting particles. Hence when $100\,\text{GeV} \ll T \ll 10^{16}\,\text{GeV}$, 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.

- To convert between times and temperatures during radiation domination, note that $\rho \sim T^4$ and $\rho \sim H^2 M_{\rm pl}^2$. Therefore, we have $t \sim 1/H \sim M_{\rm pl}/T^2$.
- It is also useful to estimate decay rates for particles. For a particle A which decays into two much less massive, distinct particles, through a vertex factor of g,

$$\Gamma \sim \frac{g^2 m_A}{8\pi}$$

  where the m_A appears on dimensional groups, g^2 appears because Γ involves a squared matrix element, and $1/8\pi$ is the generic phase space factor. Because $M_{\rm pl}/T \gg 1$, this means that any particles that can decay in this way vanish almost immediately once $T \lesssim m_A$, i.e. once they stop being thermally produced.
- Exceptions to this general rule can occur for decays that are substantially slower. For example, for neutron decay the amplitude picks up a factor of g_W^2/m_W^2 from the intermediate W boson,

$$\Gamma \sim \frac{g_W^4}{128\pi^3} \frac{(m_n - m_p)^5}{m_W^4} \sim \frac{1}{15\,\text{min}}$$

  where $1/128\pi^3$ is the typical numeric factor of three-body decays, and the dimensional phase space must be taken to be $(m_n - m_p)^5$ because $m_n \approx m_p$. This slow neutron decay will be important when studying nucleosynthesis, as they only decay when $H \sim \Gamma$, or $T \sim 100\,\text{keV}$.
- If equilibrium persisted forever, all massive particles species would eventually be exponentially suppressed. However, consider a species of massive particle with an equal number of antiparticles, with a conserved charge. These particles must be removed by annihilation with their antiparticles, and 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

$$\frac{\Gamma}{H} \sim G_F^2 M_{\rm pl} T^3 \sim \left(\frac{T}{1\,\text{MeV}}\right)^3.$$

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

Example. Consider a collection of photons with energy $\omega \ll m_e$, which thermalize by light-by-light scattering. The leading light-by-light interaction is $2\gamma \to 2\gamma$ through an electron loop. Estimating

the diagram is subtle, because the electron propagators can have powers of m_e. An easier route is to note that upon integrating out the electron, the leading term in the effective Lagrangian is

$$\mathcal{L} \supset \frac{\alpha^2}{16\pi^2 m_e^4} F^4$$

where the powers of m_e follow by dimensional analysis. This leads to a cross section

$$\sigma \sim \frac{\alpha^4 \omega^6}{m_e^8}$$

where the powers of ω follow by dimensional analysis. However, this process can't thermalize the photons by itself, because the scattering is elastic. The leading nonelastic process is $2\gamma \to 4\gamma$, where

$$\mathcal{L} \supset \frac{\alpha^3}{16\pi^2 m_e^8} F^6, \quad \sigma \sim \frac{\alpha^6 \omega^{14}}{m_e^{16}}.$$

Note there are no odd powers of F in the effective Lagrangian, as they vanish by F's antisymmetry. Also note that gauge invariance forces the appearance of F, which pulls out factors of external momenta, making the loop diagrams converge; this is why there is no large logarithm.

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. Questions marks indicate that we have no idea when it happened: it could be as late as the start of Big Bang nucleosynthesis, or as early as the end of inflation. Similarly, we don't know precisely when inflation ended.

- 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 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. Incidentally, a useful mnemonic during radiation domination is $t/1\,\mathrm{s} \sim (1\,\mathrm{MeV}/T)^2$.

- 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 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. Stars form at $z \sim 30$, which causes most hydrogen gas to reionization at $z \sim 10$. The ionization of the hydrogen gas over time can be measured through the 21 cm line, which only exists for neutral hydrogen.

It should be noted that many elements of the story above rely on extrapolation; the earliest element with good direct support is BBN, which strongly constrains many alternative models. Any events before BBN might never have happened.

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 by definition,

$$n = \frac{g}{(2\pi)^3}\int d\mathbf{p}\, f(p).$$

- 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 $\mathbf{v} = \mathbf{p}/E$.

- 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

$$\mu_1 + \mu_2 = \mu_3 + \mu_4.$$

  Photons can always be easily produced by, e.g. double Compton scattering

$$e^- + \gamma \leftrightarrow e^- + \gamma + \gamma$$

  which sets $\mu_\gamma = 0$. As a result, by considering the process

$$X + \overline{X} \leftrightarrow \gamma + \gamma$$

  we must have $\mu_X = -\mu_{\overline{X}}$ in just about any reasonable situation. 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

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

  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

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

  Hence we have

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

  A very similar computation yields

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

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

$$\frac{\rho}{n} \approx \begin{cases} 2.7\,T & \text{bosons,} \\ 3.2\,T & \text{fermions.} \end{cases}$$

- 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 intuitive: the prefactor $(mT)^{3/2} \sim p^3$ counts the number of accessible momentum states, while $e^{-m/T}$ is the Boltzmann factor for the rest energy.

- 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. We define the number of relativistic degrees of freedom $g_*(T)$,

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

  where the sum is over species with $m < T_i$, and T_i is the temperature of the species, which most of the time will just be equal to T. This definition is chosen so that

$$\rho \approx \frac{\pi^2}{30} g_*(T) T^4$$

  during radiation domination, or equivalently,

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

  The factor of $M_{\rm pl}$ gives $H \ll T$, in contrast to $H \sim T$ during inflation.

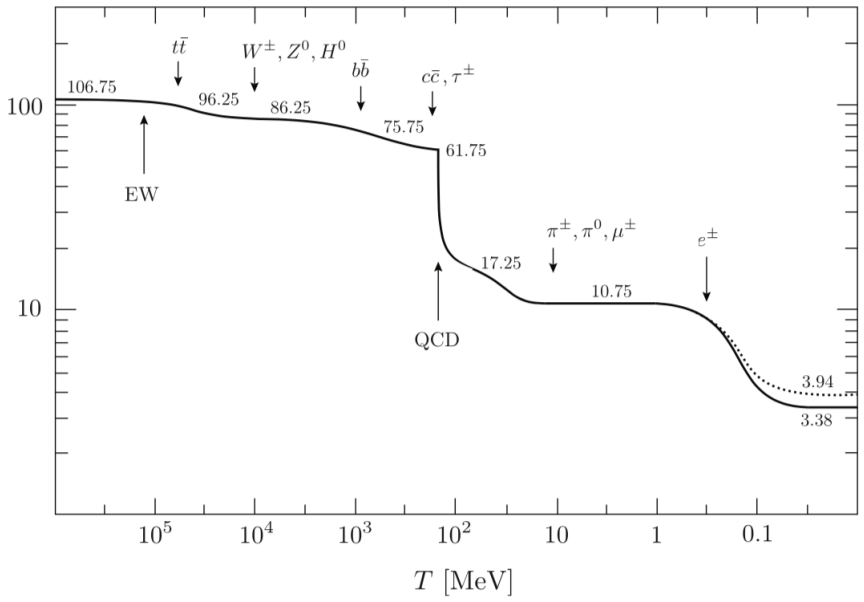
- At high temperatures, the SM degrees of freedom are counted as follows.

  – 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 $T \geq 100\,\text{GeV}$.

- 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.

Note. More about the electroweak and QCD phase transitions. There are three possibilities for each of these: a violent first order transition driven by bubble nucleation, a second order transition, and

a rapid "crossover", which looks like a second order transition but is perfectly analytic. (Crossovers hence aren't true phase transitions at all, though it's conventional to include them in.) For the purposes of cosmology, the main concern is whether these transitions are first order or not. Such violent transitions would generate relics that we could see now, such as gravitational waves.

Lattice calculations indicate that the QCD phase transition is also a crossover, and that the electroweak phase transition is a crossover for $m_H \gtrsim 80\,\mathrm{GeV}$. However, extensions to the SM such as GUTs often generate first-order phase transitions. One can also calculate whether further minima occur in the Higgs potential for higher Higgs vevs, and remarkably the measured value of the Higgs puts the SM just on the boundary between metastability and absolute stability.

We now discuss the conservation of entropy.

- First, we derive a useful identity. Consider a comoving volume, and focus on a single particle species which is in thermal equilibrium. The first law of thermodynamics gives

$$d(\rho V) = T\,dS - p\,dV + \mu\,dN.$$

- Now we switch to the intensive number and entropy,

$$n = \frac{N}{V}, \quad s = \frac{S}{V}.$$

  Plugging these results in gives

$$(Ts - p - \rho + \mu n)\,dV + (T\,ds - d\rho + \mu\,dn)V = 0.$$

  This has completely separated variation of intensive parameters and extensive parameters, so both terms must separately vanish.

- We can show this more formally. Think about the entropy contained in an imaginary box of fixed physical volume, embedded in this comoving volume. The first term above vanishes, so

$$T\,ds - d\rho + \mu\,dn = 0.$$

  But this is a relation among intensive variables, which holds regardless of the volume we are considering. Thus the second term always vanishes, so the first term always vanishes.

- This gives the useful identity

$$s = \frac{\rho + p - \mu n}{T}.$$

  For multiple species, this is true if all quantities above are taken to be for that specific species. (Often, this equation is quoted with $\mu = 0$, because it's usually used when μ is negligible.)

- Now think about the entropy due to *all* particle species, again in a comoving volume. We assume everything is in chemical equilibrium, so the $\sum_i \mu_i\,dN_i$ terms don't contribute anything, so

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

  However, this vanishes by the continuity equation,

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

  The conservation of entropy makes perfect sense, as everything within our comoving volume is in equilibrium, and there can't be any heat transfer with anything outside.

- 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 useful when species annihilate, because this process is adiabatic,

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

  which implies $T \propto g_{*S}^{-1/3} a^{-1}$. This reproduces the usual $1/a$ falloff, but when a species annihilates, the temperature has a different dependence, as the entropy of the annihilating species is transferred to other relativistic species. Specifically, the temperature continues to decrease, but more slowly, since annihilation is a gradual process that occurs as a changes by an $O(1)$ factor.

- Note that entropy is not conserved in nonequilibrium processes, such as WIMP freeze out (or more generally the later stages of any annihilation process), structure formation, and the entire history of life. However, in the standard cosmological story, this doesn't matter because relativistic species (i.e. photons) carry the vast majority of the entropy. However, it is possible to induce large changes in the entropy, e.g. in a first-order phase transition.

- 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 $T \propto 1/a$ since all neutrinos are redshifted equally, even when the "temperature" drops below the neutrino mass. The phase space distribution just shrinks towards $p = 0$. If there was a chemical potential at decoupling, then it also redshifts as $1/a$.

- A similar story holds for a species that decouples when it is nonrelativistic, with its momentum redshifting as $1/a$ and hence its kinetic energy redshifting as $1/a^2$. The phase space distribution remains thermal, with $T \propto 1/a^2$. Thus, e.g. we know that WIMP DM would have extremely low temperature by the time of structure formation.

- There is a technicality. The number density of a nonrelativistic species at temperature T is

$$n \sim T^{3/2} e^{-(m-\mu)/T}.$$

  However, the number density of a decoupled species does not exponentially decay, but rather redshifts as $1/a^3$, which is completely accounted for by the $T^{3/2}$ prefactor. Thus the exponential has to stay constant, which can be done by introducing an artificial, growing chemical potential which keeps $(m-\mu)/T$ constant.

- If a species decouples when it is semi-relativistic, its later distribution can't be taken approximately thermal at all, for any values of μ and T. In this case we have to fall back on using the phase space distributions directly.

- Putting these annoying issues aside, this picture also gives a simple, intuitive reason S is conserved, for decoupled particles. The entropy depends on the available phase space volume. Now take a given comoving volume as the system. The number of particles inside is fixed, the physical volume grows as a^3, and the accessible momentum space falls as $1/a^3$.

- 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 $1/a$. Both photons and neutrinos *always* redshift 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

$$g_* = 2 + \frac{7}{8} \times 2N_{\mathrm{eff}} \left(\frac{4}{11}\right)^{4/3} = 3.36, \quad g_{*S} = 2 + \frac{7}{8} \times 2N_{\mathrm{eff}} \left(\frac{4}{11}\right) = 3.94$$

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

$$N_{\mathrm{eff}} = \begin{cases} 2.99 \pm 0.33 & \text{experiment}, \\ 3.046 & \text{theory}. \end{cases}$$

  Refinements of the measurement of N_{eff} by the Simons Observatory and CMB-S4, which will improve the uncertainty to about 0.06, can probe new light physics.

- The number density of neutrinos is

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

  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,

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

  However, experiments indicate that neutrinos have mass, with

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

- The temperatures of the CMB and "CνB" are

$$T_0 = 2.73\,\text{K} = 0.24\,\text{meV}, \quad T_\nu = \left(\frac{4}{11}\right)^{1/3} T_0 = 1.95\,\text{K} = 0.17\,\text{meV}.$$

  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

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

  In fact, more stringent experimental tests show that

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

  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 seems likely that no neutrino species remain relativistic today. However, it's logically possible that one neutrino species is massless.

- Note that we have implicitly assumed above that there is no neutrino asymmetry, $n_\nu = n_{\overline{\nu}}$. In some models of baryogenesis, one ends up with such an asymmetry. A simple constraint on this is that too much asymmetry would lead to a measurable change in ρ_ν and hence the total energy density of the universe; however, this constraint is extremely weak.

3.3 The Boltzmann Equation

Next, we introduce a simple form of the Boltzmann equation to describe 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.

- Usually, reactions involving three or more particles are subleading, 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. (More properly, we can think of α as the *definition* of $\langle\sigma v\rangle$.) Hence we have

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

In a more rigorous treatment, we would start from a microscopic treatment and use time reversal symmetry to *derive* detailed balance here.

- It is useful to write this in terms of the number of particles per comoving volume $N_i = n_i/s$,

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

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.

- Many approximations have been made to arrive at the expression above. We have made the usual Boltzmann approximation of neglecting higher-body correlations, and we have further neglected any phase space structure of the species. Also, we have neglected the effects of quantum statistics, which is valid if the phase space occupancy is much less than 1. (Otherwise, we would have either Fermi blocking or Bose enhancement.) For a more complete treatment, see the notes on Undergraduate Physics.

Note. A more careful derivation would show that the quantity v in $\langle\sigma v\rangle$ is really the so-called Moller velocity,

$$v_M = \frac{\sqrt{(p_1\cdot p_2)^2 - (m_1 m_2)^2}}{E_1 E_2}.$$

This has the property that $v_M n_1 n_2$ is Lorentz invariant. However, for most applications it's good enough to take v to be the typical velocity.

As a first example, we consider the freeze-out production of traditional WIMP dark matter, which occurs during radiation domination.

- 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
$$\frac{dN_X}{dt} = -s\langle\sigma v\rangle \left(N_X^2 - (N_X^{\rm eq})^2\right).$$
We also assume that the interaction is strong enough so that the dark matter abundance reaches thermal equilibrium at some point in the early universe.

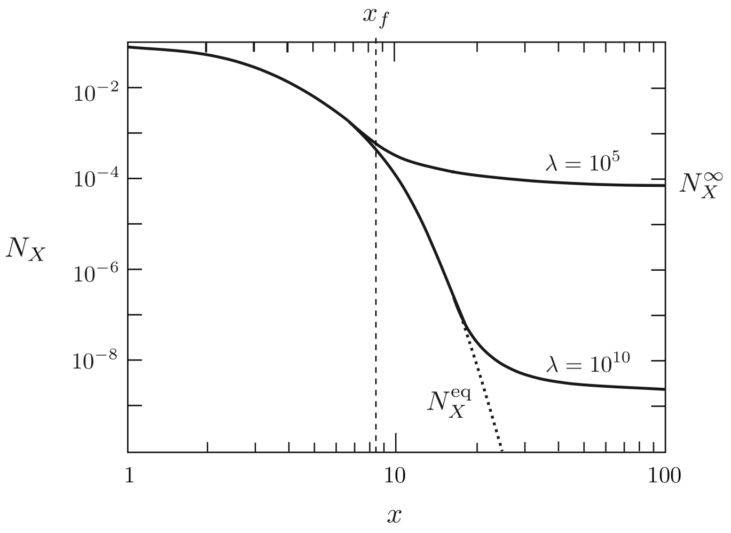
- There are many variants on this, such as having $n_X \neq n_{\overline{X}}$ or having the dominant reaction be $3X \leftrightarrow \ell + \bar{\ell}$, which lead to slightly different constraints. The case where the dark matter begins at zero density and never reaches thermal equilibrium due to extremely weak interactions is called "freeze-in", because the abundance goes up during cosmological evolution rather than down. Here we just consider classic freeze-out.

- One might ask how the Boltzmann equation changes if the dark matter can annihilate to multiple things, $\ell_i + \bar{\ell}_i$. Fortunately, nothing about the calculation changes because the dark matter production rate is fixed in terms of the annihilation rate by detailed balance. We just need $\langle\sigma v\rangle$ to count all possible annihilation processes.

- 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 the Riccati equation
$$\frac{dN_X}{dx} = -\frac{\lambda}{x^2}\left(N_X^2 - (N_X^{\rm eq})^2\right), \quad \lambda = \frac{2\pi^2}{45}g_{*S}\frac{M_X^3\langle\sigma v\rangle}{H(M_X)}$$
where λ is the dimensionless interaction strength. The pair creation/annihilation process becomes less effective as the universe expands; the opposite would be true for a decay process, since it scales differently.

- Note that λ depends on x, but it doesn't vary too much during freeze-out, so we'll treat it as constant. The reason is that usually s-wave annihilation dominates, and in this case the x-dependence of σ and v cancel; the contribution of p-wave annihilation is v^2 smaller. The results can qualitatively change if this is upset, e.g. if p-wave annihilation is the leading contribution, or if one has Sommerfeld enhancement. **(understand when *s*-wave annihilation is allowed)**

- The Riccati equation has no closed-form solution; numeric solutions are below.



When $\lambda \gg 1$, for a wide range of λ we find significant departure from equilibrium occurs near $x \sim 10$ or $x \sim 20$. For concreteness we'll take $x \sim 10$, because the point is that the dependence on λ is quite weak.

- At this point $N_X^{\rm eq}$ 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

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

Plugging this in, using $\rho_{c,0} = 3M_{\rm pl}^2 H_0^2$, 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

$$\langle \sigma v \rangle \sim \left(10^{-4}\,\mathrm{GeV}^{-1}\right)^2 \sim 0.01\,G_F \sim 3 \times 10^{-26}\,\mathrm{cm}^3/\mathrm{s}.$$

  However, for s-wave annihilation through a weak gauge boson, and assuming that the phase space for annihilation is also given by the weak scale (e.g. if the mass is $m_X \sim m_W$), then

$$\langle \sigma v \rangle \sim \frac{1}{8\pi} \frac{g_W^4}{m_W^4}\, m_W^2$$

  which matches! This is called the WIMP miracle.

- The calculation above points to a weak-scale or TeV-scale WIMP. On the other hand, if we *fix* the cross section, then the WIMP mass can vary over a wide range. It could be up to about 100 TeV before the required cross section runs into the s-wave unitarity bound,

$$\langle \sigma v \rangle \leq \frac{\pi}{m^2 v}(2j+1).$$

  This bound assumes only that we have point particles scattering (e.g. heavy composite particles would have cross sections instead set by their geometric size), and that the s-wave component of the partial wave expansion dominates; this is reasonable since higher partial waves are penalized by powers of v^2. Most of the time we cannot even saturate the s-wave unitarity bound, because couplings will be small, as will v at freeze-out.

- Another situation would be a "hot relic", where λ is small and decoupling occurs when the WIMPs are still relativistic, $x \lesssim 1$. This is how neutrinos work, and as for neutrinos, it would give a huge density of dark matter particles, on the same order of magnitude as CMB photons.

- Since the number density per comoving volume is fixed, the current DM density would be roughly proportional to the mass m. The right DM density would result for $m_{\mathrm{DM}} \sim 10\,\mathrm{eV}$. To see this, note that we want $m_{\mathrm{DM}} \sim 5T$ where T is the temperature at matter-radiation equality, so that the DM density is about 5 times higher than the baryonic mass density. Such "hot DM" would still be relativistic during structure formation, and is observationally ruled out. On the other hand, "warm DM" with $m_{\mathrm{DM}} \sim 1\,\mathrm{keV}$ is a possibility.

- Again, if we fix the cross section, the above reasoning places a weak lower bound on the WIMP mass. However, the Lee–Weinberg bound states that a WIMP cannot have a mass below about 2 GeV. This is because annihilation processes due to weak bosons get suppressed at such low energies: for $m_{\mathrm{DM}} \ll m_W$ the weak gauge boson propagator becomes $1/m_W^2$, so roughly

$$\langle \sigma v \rangle \sim \frac{g_W^4}{8\pi} \frac{m_{\mathrm{DM}}^2}{m_W^4}.$$

  The simplest way to evade this bound is if the DM annihilates through a new, non-SM interaction, which leads to the ideas of "dark sectors" and "light dark matter", covered in more detail below.

Note. Just how amazing is the WIMP miracle? Dropping all order-one constants, the WIMP freezes out at $T_f \sim m_{\mathrm{DM}}/10$, where an f subscript denotes freeze-out. To get the correct final density, it must be on par with the radiation density at the time of matter-radiation equality, so

$$T_{\mathrm{eq}}^4 \sim m_{\mathrm{DM}} n_{\mathrm{eq}} \sim m_{\mathrm{DM}} n_f \left(\frac{T_{\mathrm{eq}}^3}{T_f^3}\right)$$

from which we conclude

$$\frac{10 n_f}{T_f^2} \sim T_{\rm eq}.$$

Now the abundance at freeze-out is determined by the interaction cross-section,

$$n_f \langle \sigma v \rangle \sim H_f$$

and we also know during radiation domination that $T_f^2 \sim H_f M_{\rm pl}$, giving

$$\langle \sigma v \rangle \sim \frac{10}{M_{\rm pl} T_{\rm eq}} \sim \left(10^{-4}\,{\rm GeV}^{-1}\right)^2$$

which is really just the statement that the scale associated with the cross section is midway between the Planck scale and the atomic scale, which the weak scale indeed is. The right-hand side "could have" ranged over about 20 orders of magnitude, so the WIMP miracle is perhaps a one-in-ten piece of evidence, which isn't that strong. What really made weak-scale WIMPs so popular is the additional fact that you could get such particles automatically in many particle physics models motivated by new weak-scale physics, e.g. to solve the hierarchy problem.

WIMP direct detection experiments have been steadily improving for decades. However, we only know the annihilation cross-section of fast-moving WIMPs to any SM particles precisely, while what matters for such experiments is the scattering cross-section of slow-moving WIMPs with heavy nuclei, which is strongly model-dependent. The naive guess of Z-mediated elastic scattering was ruled out long ago, while Higgs-mediated scattering was probed more recently. However, one can easily make WIMPs that are much harder to detect by using, e.g. spin-dependence or loop suppression. What is objectively true is that we have definitively ruled out WIMPs whose spin-independent elastic WIMP-nucleon cross section naively matches their annihilation cross section.

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 $T > 1\,{\rm eV}$. We begin with equilibrium considerations. Since all particles besides the photon are nonrelativistic,

  $$n_i^{\rm eq} = g_i \left(\frac{m_i T}{2\pi}\right)^{3/2} e^{(\mu_i - m_i)/T}, \quad \mu_p + \mu_e = \mu_H.$$

- To remove the dependence on the chemical potential, we consider the ratio

  $$\left(\frac{n_H}{n_e n_p}\right)_{\rm eq} = \frac{g_H}{g_e g_p} \left(\frac{m_H}{m_e m_p} \frac{2\pi}{T}\right)^{3/2} e^{(m_p + m_e - m_H)/T}.$$

  The first factor is $4/(2 \times 2) = 1$. The exponential factor is $e^{B_H/T}$ where $B_H = 13.6\,{\rm eV}$ is the binding energy of hydrogen. Since the universe is charge neutral, we have $n_e = n_p$, giving

  $$\left(\frac{n_H}{n_e^2}\right)_{\rm eq} = \left(\frac{2\pi}{m_e T}\right)^{3/2} e^{B_H/T}$$

  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

$$\eta = \frac{n_b}{n_\gamma} = 5.5 \times 10^{-10} \left(\frac{\Omega_b h^2}{0.020} \right), \quad n_\gamma = \frac{2\zeta(3)}{\pi^2} T^3.$$

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

$$\left(\frac{1 - X_e}{X_e^2} \right)_{\text{eq}} = \frac{2\zeta(3)}{\pi^2} \eta \left(\frac{2\pi T}{m_e} \right)^{3/2} e^{B_H/T}.$$

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 $\sigma_T = (8\pi/3)(\alpha/m)^2 = 1.7 \times 10^{-3}\,\text{MeV}^{-2}$ is the Thomson scattering cross section.

- We define photon decoupling (roughly equivalent to the time of "last scattering" for a typical photon) to occur when $\Gamma_\gamma = H$. 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 "freeze out" of this system, leaving a similar relic density to the one computed more naively above.

Another complicating factor is that there is a sizable fraction of helium present at this time, due to nucleosynthesis. This doesn't qualitatively change the result, but it increases the number of reactions we have to keep track of.

3.4 Nucleosynthesis

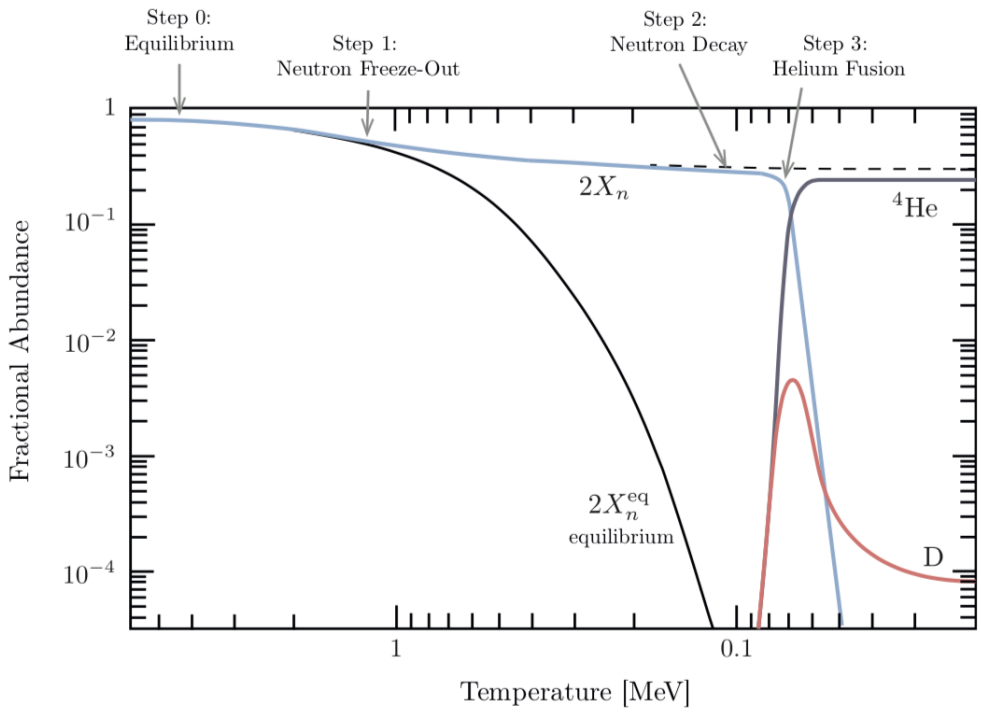
Next, we move backwards in time to cover nucleosynthesis. The successful and precise calculation of its result is one of the great triumphs of standard cosmology.

- Nucleosynthesis occurs at energy scales of $T \sim 1\,\mathrm{MeV}$. 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

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

The chemical potentials of electrons and neutrinos are negligible at this stage because they are very light (assuming the lepton asymmetries are not too large), so $\mu_n \approx \mu_p$. Then

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

We assume that protons and neutrons are brought to equilibrium in the early universe.

- Now, once we have $T < 1\,\text{MeV}$, 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 coincidentally occurs around here, which slows down the cooling a bit, but doesn't have a qualitative effect.

- For comparison, deuterium can be produced in the reaction

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

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

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

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

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

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

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

  Because of the smallness of η and B_D, $n_{\rm D}$ is negligible for $T \gtrsim 0.1\,\text{MeV}$, so we can ignore everything except for protons and neutrons before this point. The intuition for the η suppression is that there are relatively many photons available to break apart D nuclei.

- Heavier, more strongly bound nuclei such as ^{4}He and ^{12}C would already become significant at $T \sim 0.25\,\text{MeV}$ assuming equilibrium, but this does not occur because B_D is anomalously low. This is known as the "deuterium bottleneck", and appears to be yet another coincidence.

- Therefore, we will simply track the neutron fraction

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

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

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

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

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

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

  As an aside, it's a remarkable coincidence that the neutron lifetime is around the same time nucleosynthesis is occurring; if it were much shorter, as a naive dimensional estimate would give, than the universe would have essentially only protons, and hence no complex chemistry!

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

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

  and

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

  To estimate the rates of these processes, note that the electromagnetic reactions are suppressed by $\sim 10^2$, while the last three are suppressed by the relatively small amount of deuterium.

- These reactions are sufficient to ensure 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 the deuterium bottleneck.

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

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

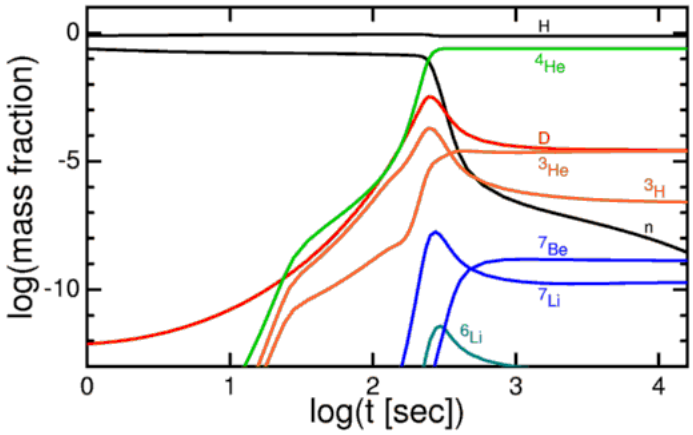
  where the time is computed using the expression for $T(t)$ in a radiation-dominated universe. At this point we account for the reactions that produce ^{4}He. Since ^{4}He has such a high binding energy, almost all the deuterium is quickly converted to it, and since each ^{4}He requires two neutrons,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

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

  or alternatively, the mass fraction of helium is about 1/4. (Real calculations solve the coupled Boltzmann equations numerically; it is intractable to do this analytically, because many reactions are happening at once, and electron-positron annihilation is happening simultaneously, changing the temperature time-dependence.)

- 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 slowed because there are no nuclei of atomic mass 5 or 8. The most common subsequent pathway is ^{4}He fusion to create ^{7}Li or ^{7}Be.

- By this time, this process is slow, as the temperature is falling to $\sim 10\,\mathrm{keV}$, which is around the temperature one needs to fuse helium in hot stellar cores. The rate falls off rapidly as temperature decreases further, because it requires tunneling through the Coulomb barrier. Thus, one ends up with a relatively small amount of ^{7}Li and ^{7}Be, as shown.



  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 lower than expected; this is called the lithium problem. Various solutions have been

proposed, ranging from systematic astrophysical or nuclear physics errors, to a new negatively charged massive particle (CHAMP) which binds to nuclei to lower the Coulomb barrier, then decays well after BBN.

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 0 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 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.

In general, it is quite hard to change any of these parameters substantially without ruining the result. Hence BBN places useful constraints on new physics. For instance, introducing new light particles in thermal equilibrium, as in some models of dark matter, would change g_*.

3.5 Models of Baryogenesis

We now say a bit about baryogenesis, one of the outstanding puzzles of cosmology. The very basics are discussed in the notes on the Standard Model.

- In the absence of a baryon asymmetry, almost all baryons and antibaryons would annihilate, leaving an equal and tiny amount of both, $(n_b + n_{\bar{b}})/n_\gamma \sim 10^{-20}$, in strong contradiction to experiment.

- Before the advent of inflation, one could simply postulate an initial baryon asymmetry. In this case, the focus was on explaining why η^{-1} was so *large*, i.e. where the large entropy per baryon came from, which could be explained by dissipative processes. However, in inflationary cosmology any initial baryon asymmetry is inflated away, in which case one has to explain how $\eta = (n_b - n_{\overline{b}})/n_\gamma$ arises after inflation ends.
- Because η changes over time, a better quantity to explain is the baryon to entropy ratio, $\eta_s = (n_b - n_{\overline{b}})/s$. These quantities are related by a factor of $g_*(T)$, which is of order 200 at the GUT scale.
- A subtlety is that electroweak sphalerons in the SM already violate baryon and lepton number, so we expect them to set $B + aL$ to zero in equilibrium. Here, a is an $O(1)$ number which depends on the right-handed fields in the SM, since the associated particles can interconvert with the ones corresponding to left-handed fields, which participate in sphalerons. A detailed calculation gives $a = 28/51$.
- In the SM, $B - L$ is conserved. Thus, if $B - L = 0$ after baryogenesis, then sphalerons set $B = L = 0$, so a baryogenesis mechanism actually has to generate $B - L$.
- Following the Sakharov conditions, a generic route for baryogenesis is to imagine some new heavy particle that decays (nonequilibrium) with net baryon number (baryon number violation, C/CP violation). Then the universe cools down, so the heavy particle can't be produced anymore, keeping the net baryon number around.
- An alternative route to baryogenesis is leptogenesis, i.e. creating an imbalance in L that is converted into nonzero B by sphalerons. One possible mechanism could be if neutrinos are Majorana and acquire mass by the seesaw mechanism. Then leptogenesis could occur by out-of-equilibrium decay of the heavy sterile Majorana neutrinos.
- Leptogenesis is popular since it works out nicely quantitatively, and it requires only minimal and well-motivated ingredients. However, it has the same problem as many other models of baryogenesis: it is difficult to test because it relies on new dynamics at very high energy scales. Because of this, baryogenesis mechanisms rarely supply direct tests, but rather motivate classes of models which allow such mechanisms to operate, which are then tested at low energies.
- A more exotic route for baryogenesis would be to break CPT symmetry, because CPT ensures the equality of the masses of baryons and antibaryons; without it, baryons could just be lighter, and that alone would lead to an asymmetry. Breaking CPT would violate cherished notions of quantum field theory, and it also has been tested to extreme accuracy in the lab. However, as we'll see, a rolling background field can effectively spontaneously break CPT in the early universe.

As a simple model for baryogenesis, we consider an $SU(5)$ GUT.

- The GUT X bosons carry baryon number and lepton number. At the GUT phase transition, the X bosons freeze out, and subsequently decay out of equilibrium. We have $X = \overline{X}$, but the decay $X \to qq$ is not balanced by the decay $X \to \overline{q}\overline{q}$, due to C and CP violation.

- More quantitatively, letting α be the GUT coupling, we have

$$\Gamma(X \to qq) \sim \alpha m_X \times \begin{cases} m_X/T & T \gg m_X, \\ 1 & T \ll m_X \end{cases}$$

where the m_X/T factor is due to time dilation, while the inverse decay rate is

$$\Gamma(qq \to X) \sim \Gamma(X \to qq) \times \begin{cases} 1 & T \gg m_X, \\ (m_X/T)^{3/2} e^{-m_X/T} & T \ll m_X. \end{cases}$$

In principle, we also need to consider the scattering process $qq \to X^* \to \overline{qq}$,

$$\Gamma(qq \to \overline{qq}) \sim n\sigma v \sim T^3 \alpha^2 \frac{T^2}{(T^2 + m_X^2)^2} \sim \alpha^2 \times \begin{cases} T & T \gg m_X, \\ T^5/m_X^4 & T \ll m_X. \end{cases}$$

For temperatures $T \gg m_X$, this process dominates.

- As mentioned in the notes on the Standard Model, turning the CP violation into an asymmetry in decay branching ratios is nontrivial, and requires the interference of a tree-level diagram with loop-level diagrams. Moreover, the particles in the loops must have the possibility of going on-shell. (The reason this makes a difference, in the interference of phases, is that it brings the $i\epsilon$ in the Feynman propagator into play.) We'll ignore these subtleties and just parametrize the asymmetry in the decay by ϵ.

- We can parametrize the decay rate in terms of the dimensionless quantity

$$K = \frac{\Gamma(X \to qq)}{2H}\bigg|_{T=m_X} \sim \frac{\alpha M_{\text{pl}}}{g_*^{1/2} m_X}.$$

Since the decay rate is effectively set by the free parameter m_X, K also parametrically determines the rates of the inverse decay and scattering processes.

- In the case $K \ll 1$, the X bosons decay slowly as $T \lesssim m_X$, and hence fall out of equilibrium. They drift along and decay much later, at which point inverse decay and scattering are negligible. This yields a baryon asymmetry of

$$\eta \sim \frac{\eta_s}{g_*(m_X)} \sim \frac{\epsilon}{g_*(m_X)}.$$

We have $\eta \sim 10^{-10}$ and $g_*(m_X) \sim 200$, so this model requires $\epsilon \sim 10^{-8}$. Note that we already have $\epsilon \lesssim 10^{-2}$ due to loop suppression.

- However, this limit corresponds to having $m_X \gg \alpha M_{\text{pl}}/g_*^{1/2} \approx M_{\text{GUT}}$. This makes it difficult for the universe to have ever been at a high enough temperature to produce the X bosons thermally; if we have $H_I \gg M_{\text{GUT}}$, then we run into bounds on CMB B-modes.

- The more realistic limit is $K \gtrsim 1$, but this requires a detailed analysis using the Boltzmann equation. In this case, we expect a smaller η because the departure from equilibrium is smaller. It turns out that as K increases, the baryon yield decreases only as $1/K$, corresponding to the smaller departure from equilibrium, but then begins to decreases exponentially for $K \gg 1$. This latter feature is because $qq \to \overline{qq}$ scattering remains effective when the X falls out of equilibrium, and erases the produced asymmetry exponentially in time. (Incidentally, this process also effectively damps out any preexisting baryon asymmetry, from before the X decay.)

- This toy model doesn't work because the $SU(5)$ GUT conserves $B-L$ through an accidental symmetry, as noted in the notes on Group Theory, so the baryon asymmetry is erased by sphalerons. However, more complicated GUTs based on, e.g. $SO(10)$ and E_6 do violate $B-L$.
- Note that even in the extreme case $K \ll 1$, the universe in this model is always close to equilibrium; the departure from equilibrium is quantified by $1/g_* \ll 1$. One can also consider "way out of equilibrium" scenarios, where the decay of the X produces most of the entropy in the universe. This could occur, for example, if X decay caused both reheating and baryogenesis simultaneously.

Next, we consider mechanisms involving rolling scalar fields.

- In spontaneous baryogenesis, we consider a real scalar field ϕ with a derivative coupling to the baryon current,
$$\mathcal{L} \supset \frac{1}{f}(\partial_\mu \phi) j_B^\mu.$$
This is quite reasonable, since this is a renormalizable interaction; the mechanism also works for generic combinations of couplings to currents, as long as one has baryon number.
- Letting ϕ be spatially constant, this becomes
$$\mathcal{L} \supset \frac{1}{f}\dot{\phi}(n_b - n_{\bar{b}})$$
so a moving field $\dot{\phi}$ produces a baryon chemical potential, $\mu_b = -\dot{\phi}/f = -\mu_{\bar{b}}$.
- Therefore, if we have efficient baryon number violating interactions, then in thermal and chemical equilibrium, we have a baryon asymmetry,
$$n_B = n_b - n_{\bar{b}} \sim -\frac{\dot{\phi} T^2}{f}, \quad \frac{n_B}{s} \sim -\frac{\dot{\phi}}{g_*(T) f T}.$$
At some point, we assume that the baryon number violating interactions become inefficient, and a baryon asymmetry is thus frozen in.
- This looks strange because it seems to violate two of the Sakharov conditions: we have thermal equilibrium, and no apparent C or CP violation. Treating ϕ as a background field, it works because CPT is spontaneously broken.
- If ϕ begins suitably large, then the final baryon abundance will be completely determined by when the baryon number violating interactions fall out of equilibrium, so the result is essentially independent of initial conditions: given the Lagrangian, there is a calculable final abundance.
- In the Affleck–Dine mechanism, we consider a complex scalar field ϕ carrying baryon number. Now, this is ambiguous because "baryon number" is only initially defined on the Standard Model fields; we could always extend its definition to ϕ so that it has zero baryon number. What we really mean is that we include terms like
$$\mathcal{L} \supset \phi \bar{q} \psi$$
where q is a quark field and ψ is a fermion with zero baryon number. Then to have a $U(1)$ symmetry that reduces to $U(1)_B$ on the SM fields, we *must* have ϕ transform under it.

- Letting ϕ carry baryon number B_ϕ, we have a contribution to baryon number of

$$n_B \supset iB_\phi(\dot{\phi}^*\phi - \phi^*\dot{\phi}) = 2B_\phi(\operatorname{Re}\phi \operatorname{Im}\dot{\phi} - \operatorname{Im}\phi \operatorname{Re}\dot{\phi}).$$

  Expanding ϕ in polar coordinates,

$$n_B \supset B_\phi r^2\dot{\theta}, \quad \phi = \frac{r}{\sqrt{2}}\, e^{i\theta}.$$

  In other words, baryon number is present if ϕ has "angular momentum" in field space.

- Now, suppose that the ϕ potential violates baryon number, e.g.

$$V(\phi) = m^2|\phi|^2 + \lambda|\phi|^4 + \lambda'(\phi^4 + \phi^{*4}).$$

  The specific form of the potential depends on the model, but this shows the general idea.

- Given initial conditions with $\dot{\phi} = 0$, the field will begin rotating in field space as it decays through Hubble friction. Meanwhile, ϕ will decay into baryons, locking in the baryon asymmetry. The Sakharov conditions are satisfied: the ϕ field is not in thermal equilibrium, and the initial angle for ϕ spontaneously breaks C and CP symmetry.

- The Affleck–Dine mechanism is particularly exceptional because it can operate at very low energy scales, as low as $T \sim 10\,\mathrm{MeV}$.

Many more options exist, many involving exotic decays. For instance, evaporating primordial black holes could also achieve baryogenesis.

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

  $$D_t \rho = -\rho \nabla \cdot \mathbf{u}, \quad D_t \mathbf{u} = -\frac{\nabla_{\mathbf{r}} P}{\rho} - \nabla_{\mathbf{r}} \Phi.$$

- 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

  $$\partial_t^2 \delta\rho - \nabla_{\mathbf{r}}^2 \delta P = 0.$$

- Defining $\delta P = c_s^2 \delta\rho$, we find the wave equation,

  $$(\partial_t^2 - c_s^2 \nabla_{\mathbf{r}}^2)\delta\rho = 0$$

  where the parameter c_s is the speed of sound; fluctuations oscillate with constant amplitude. For example, for an ideal gas with adiabatic fluctuations, we have $c_s = \sqrt{\gamma k_B T/m}$. For a relativistic fluid, we have $c_s^2 = wc^2$, though in this case our result does not technically apply.

- Next we account for the effect of gravity. To do this, we can use the "Jeans' swindle", which is to perturb about zero gravitational potential,

  $$\nabla_{\mathbf{r}}^2(\Phi + \delta\Phi) = \nabla_{\mathbf{r}}^2(\delta\Phi) = 4\pi G \delta\rho.$$

  This is not valid, because $\nabla_{\mathbf{r}}^2 \Phi = 0$ is not true for the unperturbed solution. However, as we will see below, this term is responsible for the expansion of the universe, so the Jeans' swindle gives the right result as long as we consider small scales, where the expansion is negligible.

- Accounting for gravity gives an extra term in the wave equation,

$$(\partial_t^2 - c_s^2 \nabla_{\mathbf{r}}^2)\delta\rho = 4\pi G\overline{\rho}\delta\rho.$$

  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 Jeans wavenumber,

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

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

Note. A heuristic derivation of the Jeans length. Consider a collapsing object of radius R. By dimensional analysis, the collapse occurs on the timescale $\sqrt{G\overline{\rho}}$, independent of R. (This is also parametrically the same as the Hubble time.) Pressure builds up on the timescale R/c_s, counteracting the collapse. The pressure builds up too slowly to avert collapse if this second timescale is larger.

Note. The gravitational potential Φ in an infinite, uniform Newtonian universe is notoriously tricky. One might think, as Newton did, that by translational symmetry Φ is constant, leading to no gravitational force anywhere. This is incorrect because it doesn't obey Poisson's equation. The trap is that Φ is not defined without boundary conditions, and any possible choice of boundary conditions will break the translational symmetry. As a result, a Newtonian universe always has a privileged origin towards which everything collapses. The symmetry is restored when we go to general relativity, where we instead think of space itself expanding and contracting uniformly, with no privileged point.

Now, we 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. In other words, a Newtonian expanding universe really *is* a result of particles moving away from each other, in a fixed spatial background.

- 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 $\mathbf{r}$.

- The gradients are related by $\nabla_{\mathbf{r}} = (1/a)\nabla_{\mathbf{x}}$. 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. That is, in Fourier space the gradient pulls out a factor of the comoving wavevector $\mathbf{k}$.

- At zeroth order in all perturbations (including $\mathbf{v}$), the continuity equation becomes

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

which simply recovers $\overline{\rho} \propto a^{-3}$. Similarly, the zeroth order Euler and Poisson equations are

$$\ddot{a}\mathbf{x} = -\frac{1}{a}\nabla\overline{\Phi}, \quad \nabla^2\overline{\Phi} = 4\pi G a^2 \overline{\rho}.$$

Combining these equations gives

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}G\overline{\rho}$$

which is the Newtonian version of the deceleration equation. As mentioned earlier, we can also derive the Friedmann equation in the Newtonian picture, but this derivation is even sketchier.

- 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

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

Note that if we neglect the fluctuations on the right, we recover the familiar result $\mathbf{v} \propto a^{-1}$.

- Finally, Poisson's equation at first order is

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

and there is no need for a "swindle", since the zeroth order terms have been accounted for.

- Deriving the wave equation as above gives

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

The essential differences are that we now have a Hubble friction term, and that both $c_s(t)$ and $\overline{\rho}(t)$ have time dependence. The result is that below the Jeans length, fluctuations oscillate with decreasing amplitude, while above the Jeans length, the fluctuations grow, but at a slower pace due to the expansion of the universe. Note that k_J defined above is a physical wavenumber, while here we've switched to comoving quantities, which is why a $1/a^2$ factor has appeared.

- Remarkably, this result can also be slightly modified to work for relativistic fluids. The continuity equation must be modified with factors of $1 + w$,

$$\frac{\partial \rho}{\partial t} + 3(1+w)H\rho + \frac{1}{a}(1+w)\nabla \cdot (\rho\mathbf{v}) = 0.$$

The Poisson equation receives a factor of $1 + 3w$ because pressure gravitates like energy,

$$\nabla^2\Phi = 4\pi G a^2 \rho(1+3w).$$

The resulting wave equation is

$$\ddot{\delta} + 2H\dot{\delta} - \frac{c_s^2}{a^2}(1+w)\nabla^2\delta = (1+w)(1+3w)4\pi G\overline{\rho}\delta.$$

As a concrete example, we'll consider dark matter and radiation.

- Dark matter is special because it feels gravity but doesn't interact significantly with itself or with other matter or radiation, so the pressure term above can be dropped, setting the Jeans length to zero. This means that dark matter behaves very differently from ordinary matter during structure formation, tending to clump up on smaller scales; this effect is 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 we 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.

- Next, consider radiation perturbations. Since $c_s^2 = wc^2$ with $w = 1/3$, the Jeans length and the horizon are parametrically equal, so there are only two cases. For subhorizon modes, δ_r oscillates, and hence doesn't grow. For superhorizon modes during radiation domination, where $a \propto t^{1/2}$ and hence $H = 1/2t$, we have

$$\ddot{\delta}_r + \frac{1}{t}\dot{\delta}_r - \frac{1}{t^2}\delta_r = 0$$

  which has polynomial solutions,

$$\delta_r \propto \begin{cases} t \propto a^2 \\ t^{-1} \propto a^{-2} \end{cases}.$$

  Thus, superhorizon radiation modes grow extremely quickly during radiation domination. The reason is that on superhorizon scales, the pressure of radiation doesn't provide an effective restoring force, but it does contribute to the gravitational force.

- Now consider matter perturbations during radiation domination. For subhorizon modes, the radiation perturbations oscillate rapidly, while pressureless matter only evolves over cosmological timescales. Thus the radiation perturbations can be ignored, giving

$$\ddot{\delta}_m + \frac{1}{t}\dot{\delta}_m - 4\pi G\overline{\rho}_m\delta_m \approx 0.$$

  Because we are in radiation domination, the last term is negligible. This can also be seen quantitatively by noting that $\partial_t \sim H$ and $H \sim \sqrt{\rho}/M_{\rm pl} \sim \sqrt{\rho G}$.

- The equation can then be solved straightforwardly to get, for subhorizon modes,

$$\delta_m \propto \begin{cases} \text{const} \\ \log t \propto \log a \end{cases}.$$

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.

- For superhorizon modes, the situation is more complicated because we cannot ignore δ_r, which now grows significantly. Treating this as a source term, one can show that δ_m also grows as a^2, pulled along with the radiation perturbations.

- 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

$$\ddot{\delta}_m + 2H\dot{\delta}_m - 4\pi G\overline{\rho}_m\delta_m = 0.$$

As in radiation domination, the final term is negligible. Since H is constant, the solutions are

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

Thus, matter fluctuations stop growing at this stage, so structure formation must go nonlinear before this point to be effective. This is indeed true in the standard cosmological model.

- The result is summarized in the table below.

	RD	MD	ΛD
subhorizon δ_m	$\log a$	a	1
superhorizon δ_m	a^2	a	1
subhorizon δ_r	1	1	-
superhorizon δ_r	a^2	-	-

The empty entries here aren't physically relevant. For context, at $z \sim 10^3$ we expect that the dark matter perturbations are $\delta_m \sim 10^{-3}$, while primordial perturbations are of order 10^{-5}.

- In the ΛCDM model, the basic story is that fluctuations were sourced by inflation. Subhorizon dark matter perturbations grew slowly during radiation domination, and then after matter domination at $T_{\text{eq}} \approx 1\,\text{eV}$, grew quickly. The baryons and photons remain tightly coupled and hence behave like a relativistic fluid (i.e. the photons set the sound speed), and hence their perturbations begin to oscillate, rather than grow, once they enter the horizon; these are called baryon acoustic oscillations. **(sources differ on the description of this; does dark matter play a nontrivial role in these oscillations?)**

- At recombination at $T_{\text{rec}} \approx 0.3\,\text{eV}$, the photons and baryons decouple. The baryons are stuck in whatever phase of the baryon acoustic oscillation they were in, then fall into the potential well created by the dark matter; the two matter perturbations then grow together. Our formalism then breaks down at $z \sim 30$, where structure formation becomes significantly nonlinear.

- This story can be tested by measurements of the CMB temperature, which reflect the potential and hence the matter power spectrum at the time of recombination, and measurements of the current matter power spectrum. Note that dark matter has an important structure enhancing effect because $T_{\rm rec} < T_{\rm eq}$, which appears to be a coincidence.
- The fact that this story works places bounds on "hot" dark matter, i.e. dark matter that is relativistic at $T_{\rm eq}$. Such matter has a sizable pressure, but doesn't exert it on itself, so our formalism above does not apply – instead its constituents "free stream", wiping out density perturbations on scales smaller than ct, where t is the total time.
- As a simple check, we can compute the Jeans mass after photon-baryon decoupling, where $T \sim 1\,\text{eV}$. Since the Jeans length is parametrically c_s/H, we have

$$M_J \sim \rho \left(\frac{c_s}{H}\right)^3.$$

  Since this is shortly after matter-radiation equality, we can crudely estimate the energy density as $\rho \sim T^4$, and hence $H \sim \sqrt{\rho}/M_{\rm pl} \sim T^2/M_{\rm pl}$. The speed of sound is $c_s \sim \sqrt{T/m_p}$, so

$$M_J \sim M_{\rm pl} \frac{M_{\rm pl}^2}{m_p^{3/2} T^{1/2}} \sim 10^4 M_\odot$$

  which is indeed on the scale of the smallest dwarf galaxies.

Note. Treating perturbations larger than the Hubble scale properly requires general relativity, which we set up in the next section. These "superhorizon modes" become important when they enter the horizon, and also provide evidence for inflation. Without inflation, the (comoving) Hubble sphere only grows in the early universe, so there is no way to account for the origin of these superhorizon modes beyond simply specifying them as an initial condition. As we'll see, within inflation they can be formed during inflation in a calculable way, get stretched out beyond the horizon, then reenter it.

The main subtlety that appears in the relativistic case is that the apparent form of the results can depend on our choice of gauge. (This isn't an issue for subhorizon modes, because we can restrict to comoving coordinates at such scales; the point is that extending this beyond the horizon is ambiguous, because the coordinates start to "feel the curvature".) However, it turns out that the results we have derived above actually do apply on superhorizon scales, provided we work in comoving gauge. To avoid ambiguity, we will denote the comoving gauge perturbations by Δ when working relativistically, and perturbations in a general gauge by δ.

Finally, we connect these results to the observed power spectra.

- We can distill the evolution of a perturbation in terms of a transfer function,

$$\delta(\mathbf{k}, t_0) = T(k)\delta(\mathbf{k}, t_i)$$

  where t_i is usually taken to be just after the end of inflation.
- Using our previous results, we can infer the form of $T(k)$ for matter perturbations. The main difference is that subhorizon and superhorizon modes grow differently during radiation domination. Thus, $T(k)$ has nontrivial k-dependence only for modes that reenter the horizon during radiation domination.

- The modes with wavenumber $k_{\rm eq}$ reenter the horizon at matter-radiation equality,

$$k_{\rm eq} = (aH)_{\rm eq}.$$

For long wavelength modes with $k < k_{\rm eq}$, we have

$$T(k) = \left(\frac{a_{\rm eq}}{a_i}\right)^2 \frac{a_0}{a_{\rm eq}}$$

where we've ignored the slow logarithmic growth during radiation domination. But for $k > k_{\rm eq}$,

$$T(k) = \left(\frac{a_{\rm enter}}{a_{\rm eq}}\right)^2 \left(\frac{a_{\rm eq}}{a_i}\right)^2 \frac{a_0}{a_{\rm eq}} \propto a_{\rm enter}^2$$

where $a_{\rm enter}$ is the scale factor at horizon reentry,

$$k = (aH)_{\rm enter}.$$

During radiation domination, we have $a \propto t^{1/2}$ and $H \propto 1/t \propto 1/a^2$, so $k \propto 1/a_{\rm enter}$.

- Therefore, the transfer function has the form

$$T(k) \sim \begin{cases} 1 & k < k_{\rm eq} \\ k^{-2} & k > k_{\rm eq} \end{cases}.$$

One might also consider what happens for k so large that the mode never exits the horizon at all. However, perturbations in these modes are negligible, because the perturbations ultimately come from particle production effects associated with quantum field theory in curved spacetime, and on subhorizon scales the curvature has little effect.

- The spatial structure of the perturbations can be characterized in terms of the two-point correlation function,

$$\xi(|\mathbf{x}-\mathbf{y}|, t) = \langle \delta(\mathbf{x}, t)\delta(\mathbf{y}, t)\rangle$$

with a spatial average on the right-hand side. Note that isotropy tells us the correlation function depends only on $|\mathbf{x}-\mathbf{y}|$, while homogeneity tells us why it is a useful thing to compute at all.

- We take the same Fourier transform convention as the notes on Quantum Field Theory. In momentum space, we have

$$\langle \delta(\mathbf{k}, t)\delta(\mathbf{k}', t)\rangle = \not{\delta}(\mathbf{k}+\mathbf{k}')P(k, t), \quad P(k, t) = \int d\mathbf{r}\, e^{i\mathbf{k}\cdot\mathbf{r}}\xi(r, t)$$

where $P(k, t)$ is called the power spectrum. This is just the three-dimensional version of the Wiener–Khinchin theorem. More explicitly, expanding in spherical coordinates, we have

$$P(k, t) = \frac{4\pi}{k}\int_0^\infty dr\, r \sin(kr)\xi(r, t).$$

- The perturbations in our universe are predicted and observed to be approximately adiabatic and Gaussian, as we will see in detail in the following sections. Here, "adiabatic" means that the magnitudes of the perturbations of all species are simply related at the end of inflation, because

they all result from the fluctuations of the inflaton; this is why we haven't bothered putting subscripts on the fluctuations above. "Gaussian" means that the modes are independent, with

$$\mathbb{P}(\delta(\mathbf{k})) \propto \exp\left(-\frac{\delta(\mathbf{k})^2}{2P(k)}\right).$$

This means that the modes are, statistically, completely characterized by the two-point correlation function, or equivalently the power spectrum; thus measuring higher-point correlators is a test of Gaussianity.

- Currently, there are no observed deviations from adiabatic, Gaussian perturbations. This is in accordance with standard inflationary theory, where the non-Gaussianity (quantified by the three-point correlation function) is typically second order in the slow roll parameters, and hence quite small. If it were possible to measure this non-Gaussianity, we would gain information about the detailed dynamics of the inflaton.

- In many models, the initial power spectrum at the end of inflation has the form

$$P(k) \propto k^n$$

where n is called the spectral index. This is equivalent to $\xi(r) \propto 1/r^{n+3}$. Note that $n = 0$ corresponds to masses being sprinkled at random in space, while $n < -3$ makes the correlation function diverge at large separation, violating the cosmological principle.

- A particularly simple choice for the initial power spectrum is $n = 1$, which is known as the Harrison–Zel'dovich power spectrum. To see why this is a natural choice, first note that if we wanted to compute the total energy in perturbations, we would need to evaluate the integral

$$\int đ\mathbf{k}\, P(k,t) = \frac{4\pi}{(2\pi)^3}\int dk\, k^2 P(k,t) = \frac{1}{2\pi^2}\int d\log k\, k^3 P(k).$$

Thus, we are motivated to define the dimensionless power spectrum

$$\Delta(k) = \frac{k^3 P(k)}{2\pi^2}$$

which measures the energy per logarithmic interval in k.

- Next, it makes sense to think about perturbations in the gravitational potential Φ, because these are ultimately what lead to all other perturbations, in standard inflation. We can similarly define their power spectrum,

$$\langle \delta\Phi(\mathbf{k})\delta\Phi(\mathbf{k}')\rangle = đ(\mathbf{k}+\mathbf{k}')P_\Phi(k), \quad \Delta_\Phi = \frac{k^3 P_\Phi(k)}{2\pi^2}.$$

The Poisson equation tells us that $\nabla^2(\delta\Phi) \sim \delta$, which means

$$P_\Phi(k) \propto k^{-4}P(k), \quad P(k) \propto k\Delta_\Phi(k).$$

The Harrison–Zel'dovich spectrum thus corresponds to constant Δ_Φ, and is hence described as scale invariant. In reality, $n \approx 0.97$, as we'll see below.

- Putting this all together, the matter power spectrum today has the form

$$P(k,t_0) = T^2(k)P(k,t_i) \sim \begin{cases} k^n & k < k_{\text{eq}} \\ k^{n-4} & k > k_{\text{eq}} \end{cases}$$

  where $a_0 k_{\text{eq}} \sim 0.01\,\text{MPc}^{-1}$. The matter power spectrum can be measured from the CMB at small k, and from the distribution of matter in the universe today at large k. Baryon acoustic oscillations can be observed at k somewhat above k_{eq}.

- Measuring the matter power spectrum from matter itself is somewhat tricky. On cosmological scales, galaxies are effectively point masses, with infinite density; they are the products of nonlinear structure growth. To get a result that can possibly match with our linearized theory, we need to average the density over a cosmological scale. This is done by convolving it with a window function of radius R,

$$\delta(\mathbf{x};R) = \int d\mathbf{x}'\, W(\mathbf{x}-\mathbf{x}';R)\delta(\mathbf{x}'), \quad \delta(\mathbf{k};R) = \tilde{W}(\mathbf{k};R)\delta(\mathbf{k}).$$

  By symmetry, our $\tilde{W}(\mathbf{k};R)$ will only depend on kR, and we reflect this in the notation below.

- The spherical top hat is a sharp cutoff in real space,

$$W(\mathbf{x};R) = \left(\frac{4\pi}{3}R^3\right)^{-1}\theta(R-|\mathbf{x}|), \quad \tilde{W}(kR) = \frac{3}{(kR)^3}(\sin kR - kr\cos kR).$$

  Another option is a sharp cutoff in momentum space,

$$\tilde{W}(kR) = \theta(1/R-k), \quad W(\mathbf{x};R) = \frac{1}{2\pi^2 r^3}\left(\sin\frac{r}{R} - \frac{r}{R}\cos\frac{r}{R}\right).$$

  Yet another option is a Gaussian in both real and momentum space,

$$W(\mathbf{x};R) = \frac{1}{(2\pi)^{3/2}R^3}\exp\left(-\frac{r^2}{2R^2}\right), \quad \tilde{W}(kR) = \exp\left(-\frac{k^2R^2}{2}\right).$$

  In all cases, we have normalized to a unit integral over space, or equivalently $\tilde{W}(0) = 1$.

- Using a window function, we can define the mass within a sphere of radius R,

$$M(R) = \int d\mathbf{x}\, W(\mathbf{x};R)\rho(\mathbf{x}).$$

  The smoothed density contrast on scale R is then

$$\delta(\mathbf{x};R) = \frac{\delta M(\mathbf{x};R)}{\overline{M}(R)}$$

  where $\delta M = M - \overline{M}$, and the variance is $\sigma^2(R) = \langle\delta^2(\mathbf{x};R)\rangle$.

- Using the definition of the power spectrum, we have the relation

$$\sigma^2(R) = \frac{1}{2\pi^2}\int d\log k\, k^3\tilde{W}(kR)P(k).$$

The formal effect of the smoothing disappears as $R \to 0$, as expected. Conventionally, σ^2 is written as a function of $\overline{M}$, which is related to R on the average by

$$\overline{M} = \frac{4}{3}\pi R^2 \overline{\rho}\gamma, \quad \gamma = \begin{cases} 1 & \text{top hat} \\ 9\pi/2 & \text{sharp cutoff in } k \\ 3\sqrt{\pi/2} & \text{Gaussian} \end{cases}.$$

The standard choice of window function is the top hat with $R = 8h^{-1}\,\mathrm{Mpc}$. The resulting variance is called σ_8^2, where $\sigma_8 \approx 0.8$.

- Using the simple asymptotic form of the matter power spectrum above, we have

$$\sigma^2(M) \sim \begin{cases} 1/R^{3+n} \sim 1/M^{(n+3)/3} & k < k_{\mathrm{eq}} \\ 1/R^{n-1} \sim 1/M^{(n-1)/3} & k > k_{\mathrm{eq}} \end{cases}.$$

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

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

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. Simple models of inflation do 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, so

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

- 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

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

- 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

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

Here we have defined the non-adiabatic pressure perturbation

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

which vanishes for adiabatic fluctuations.

- Setting $\delta P_{\rm nad} = 0$, 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

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

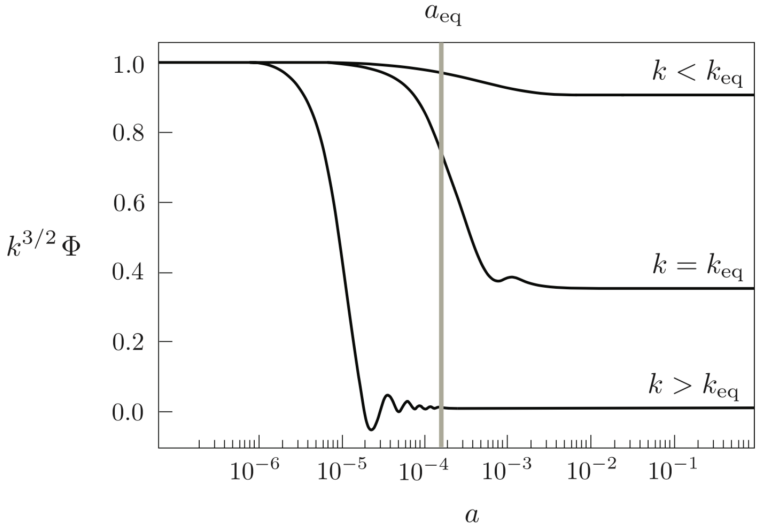
  Hence during the radiation era, subhorizon modes of Φ oscillate with frequency $k/\sqrt{3}$ and decay as $1/\tau^2 \propto 1/a^2$.

- 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 $k < k_{\rm eq}$, the mode is frozen completely, except for the 10% decrease during the transition to matter domination. For $k \sim k_{\rm eq}$, the mode is somewhat suppressed during the radiation era, while for $k \gg k_{\rm eq}$, 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.

- 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 $\delta_r = -4\Phi_{\rm MD}(k)$.

- 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 equation. 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.

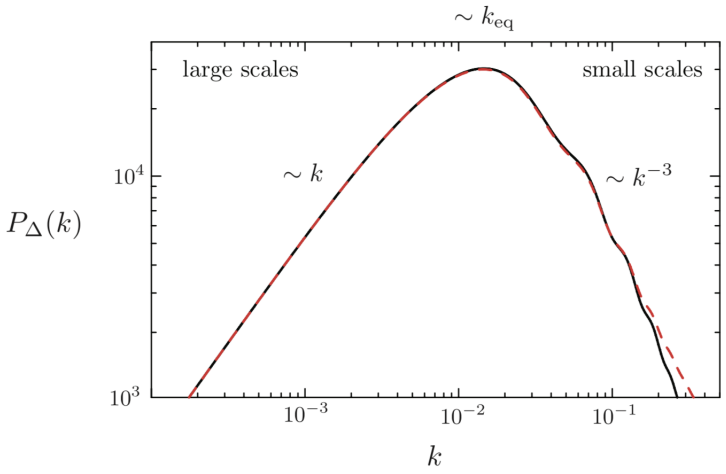
Note. The net effect of all the evolution above is that the primordial perturbations are "post-processed". We can express this in terms of a transfer function,

$$\Delta_{m,\mathbf{k}}(z) = T(k,z)\mathcal{R}_{\mathbf{k}}$$

for the matter fluctuations, where z is the redshift. The observed matter power spectrum is

$$P_\Delta(k,z) = |\Delta_{m,\mathbf{k}}(z)|^2 \sim \begin{cases} k & k \ll k_{\text{eq}}, \\ k^{-3} & k \gg k_{\text{eq}}. \end{cases}$$

Here the primordial perturbations satisfy $|\mathcal{R}_{\mathbf{k}}|^2 \propto 1/k^3$, and the asymptotic behavior follows from the exact same logic as in our Newtonian formalism. 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

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 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.
- The point here is that a quantum wavefunction differs from a stochastic mixture only because it can display interference effects – but once modes exit the horizon, such effects can't occur, so a stochastic treatment is acceptable. That is, the modes decohere upon horizon exit.
- 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

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

  The variance of curvature perturbations is therefore

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

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

Note. Before diving into the details, we give a heuristic overview of how inflationary perturbations of the inflaton affect observables today. A naive guess would be that changes in ϕ result in changes of the potential, so

$$\frac{\delta\rho}{\rho} \sim \frac{\delta V}{V} \sim \frac{V'}{V}\,\delta\phi.$$

This is incorrect; the real effect is that changes in ϕ result in changes of when inflation ends. Different regions will reheat to the same temperature once inflation ends locally, so regions where inflation

ends earlier will end up at a lower temperature because they have more time for the thermal bath to dilute. Parametrically, we have

$$\frac{\delta\rho}{\rho} \sim \frac{\delta a}{a} \sim \frac{\dot{a}}{a}\,\delta t \sim H\,\delta t \sim \frac{H}{\dot{\phi}}\,\delta\phi \sim \frac{H^2}{\dot{\phi}}.$$

This effect is thus stronger the slower ϕ rolls. Indeed, as we'll see below, the resulting power spectrum is proportional to $1/\epsilon$.

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

- We start from the inflaton action in conformal time,

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

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

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

  The normalization of f is chosen to account for Hubble friction, so that it will not have a friction term in the Lagrangian below; this will help us think about the quantization.

- 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

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

- 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

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

  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

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

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

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

- 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
$$\hat{f}_{\mathbf{k}}(\tau) = f_k(\tau)\hat{a}_{\mathbf{k}} + f_k^*(\tau)a_{\mathbf{k}}^\dagger$$
and the equations of motion imply that the modes $f_k(\tau)$ 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, during inflation, where the background is approximately de Sitter, there is a preferred vacuum state called the Bunch–Davies vacuum, corresponding to the ground state of the Hamiltonian.

- 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. 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

$$\langle|\hat{f}|^2\rangle = \int \frac{d\mathbf{k}}{(2\pi)^3}|f_k(\tau)|^2 = \int d\log k\,\frac{k^3}{2\pi^2}|f_k(\tau)|^2.$$

It is also useful to define the dimensionless power spectrum in the usual way,

$$\Delta_f^2(k,\tau) = \frac{k^3}{2\pi^2}|f_k(\tau)|^2.$$

- Scaling back to the original field, using the mode functions, and $\tau = -1/aH$, we have

$$\Delta_{\delta\phi}^2(k,\tau) = a^{-2}\Delta_f^2(k,\tau) = \left(\frac{H}{2\pi}\right)^2\left(1+\left(\frac{k}{aH}\right)^2\right)$$

which approaches the scale-invariant result $(H/2\pi)^2$, as each mode goes well outside the horizon.

- More precisely, the universe during inflation is not perfectly de Sitter, as H changes over time. In order to find the corrected result, we would in principle have to expand everything, from the very beginning, order by order in the slow-roll parameters. However, a decent approximation is to evaluate H for each mode k at the moment it leaves the horizon, giving

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

- In the approximations we have used, the distribution of each Fourier mode $\hat{f}_{\mathbf{k}}(\tau)$ is exactly Gaussian. To see this, note that

$$\langle\hat{f}_{\mathbf{k}}(\tau)^{2n}\rangle = \langle(f_k(\tau)\hat{a}_{\mathbf{k}} + f_k^*(\tau)\hat{a}_{\mathbf{k}}^\dagger)^{2n}\rangle = |f_k(\tau)|^{2n}\langle(\hat{a}_{\mathbf{k}} + \hat{a}_{\mathbf{k}}^\dagger)^{2n}\rangle$$

which shows that all nonvanishing moments scale by factors of $|f_k(\tau)|$. Since the modes are initially Gaussian distributed in the Bunch–Davies vacuum (because the ground state wavefunction of a quantum harmonic oscillator is), they remain so. Furthermore, modes of different $\mathbf{k}$ are uncorrelated. These two results lead to the observed "Gaussian" perturbations.

- Corrections to the Gaussian perturbations would have appeared if we had expanded to higher order in the perturbation f above, leading to nontrivial interactions. This is rather technical, as it also requires expanding in the metric perturbations.

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,

$$\Delta_{\mathcal{R}}^2(k) = \left(\frac{\mathcal{H}}{\overline{\phi}'}\right)^2\Delta_{\delta\phi}^2(k) = \frac{1}{2\epsilon}\frac{\Delta_{\delta\phi}^2}{M_{\rm pl}^2}$$

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

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

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

- 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_*,

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

  The amplitude A_s is of order 10^{-10}, since density perturbations are of order 10^{-5}.

- 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,

$$n_s - 1 \approx -3M_{\rm pl}^2\left(\frac{V'}{V}\right)^2 + 2M_{\rm pl}^2\frac{V''}{V} = -6\epsilon_V + 2\eta_V.$$

Next, we consider tensor perturbations to the metric.

- The tensor perturbations take the form

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

  This gives a second-order action variation of

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

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

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

  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,

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

  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.

- Tensor modes are a robust, highly model-independent prediction of inflation, and measuring their amplitude would give direct information about the inflationary scale H. (By contrast, the scalar modes are easier to measure, but only give indirect information, because they depend on ϵ.) For instance, we generically have the Lyth bound **(derive)**

$$\frac{\Delta\phi}{M_{\rm pl}} \sim \sqrt{\frac{r}{0.01}}$$

  where $\Delta\phi$ is the total field excursion during inflation.

- 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

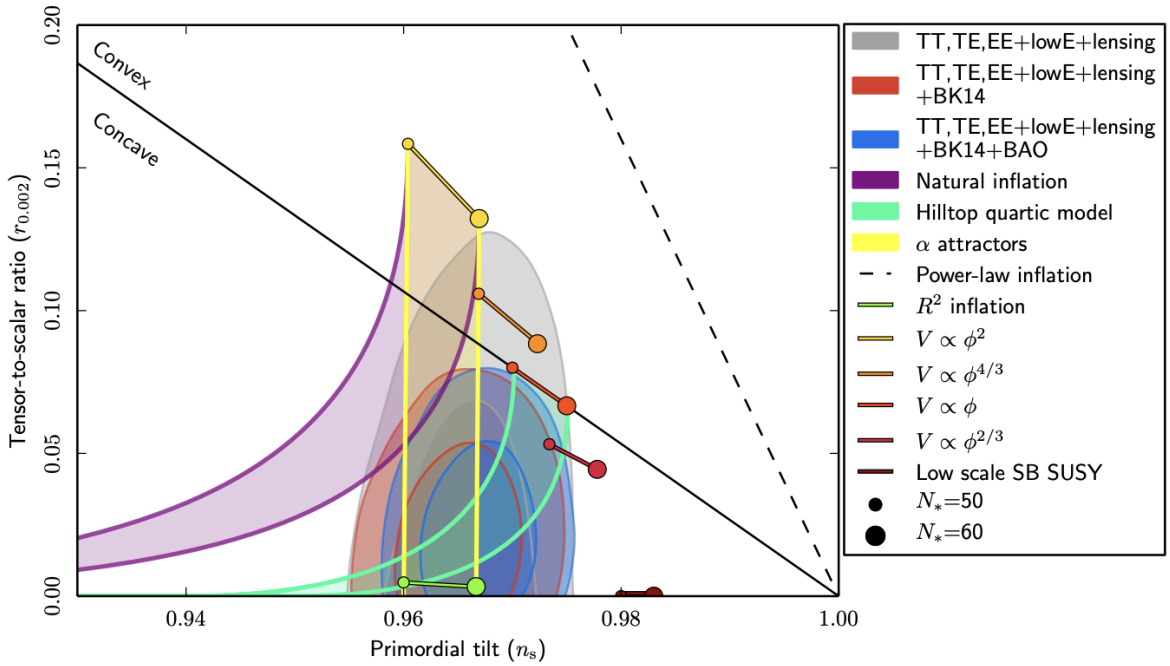
$$r = 16\epsilon, \quad n_t = -2\epsilon.$$

  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

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

  The results are shown graphically below.



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

- Tensor perturbations are detected in the CMB by the pattern of light polarization. **(understand in more detail)** 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
$$H^2 = \frac{V}{3M_{\mathrm{pl}}^2}, \quad 3H\dot{\phi} \approx -V'.$$
The slow roll parameters are
$$\epsilon = -\frac{\dot{H}}{H^2} = -\frac{d\log H}{dN} = \frac{\dot{\phi}^2/2}{M_{\mathrm{pl}}^2 H^2}, \quad \eta = \frac{d\log\epsilon}{dN} = \frac{\dot{\epsilon}}{H\epsilon}, \quad \delta = -\frac{\ddot{\phi}}{H\dot{\phi}}, \quad \eta = 2(\epsilon - \delta).$$
There are also parameters defined in terms of the potential,
$$\epsilon_v = \frac{M_{\mathrm{pl}}^2}{2}\left(\frac{V'}{V}\right)^2, \quad \eta_v = M_{\mathrm{pl}}^2 \frac{V''}{V}$$
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
$$a(t) = e^{Ht}, \quad \tau = -\frac{1}{Ha} = -\frac{1}{\mathcal{H}}.$$
Slow roll parameters are related to $\mathcal{H}$ and $\mathcal{H}'$ by
$$\epsilon = 1 - \frac{\mathcal{H}'}{\mathcal{H}^2}, \quad \eta = \frac{\epsilon'}{\mathcal{H}\epsilon}.$$

6 Dark Matter

6.1 History and Evidence

We now consider the evidence for dark matter (DM). For lack of space, the history below is drastically oversimplified; for more of the story, see *A History of Dark Matter*. We begin with one of the most straightforward pieces of evidence, galaxy rotation curves.

- In typical galaxies, we infer from the galactic luminosity distribution that the mass is strongly concentrated at the center. So 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.
- One should also account for interstellar gas in the calculation of $\rho(r)$, which extends beyond the stars. This can be measured from 21 cm line emission, but in any case is not nearly enough for galaxies.
- The curves can be fit by assuming a spherically symmetric DM halo. The halo extends far beyond the galactic disk, with significant density up to order 100 kpc from the center. 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 density profiles used for modeling, such as the Navarro-Frenk-White profile

  $$\rho(r) = \frac{\rho_0}{\tilde{r}(1+\tilde{r}^2)},$$

  the Einasto profile

  $$\rho(r) = \rho_0 \exp\left(-\frac{2}{\alpha}(\tilde{r}^\alpha - 1)\right),$$

  and the Hernquist profile. Note that a constant $v(r)$ at large radii corresponds to $\rho(r) \propto 1/r^2$.
- Such a wide DM distribution occurs naturally: objects can only gravitationally collapse if they can radiate away energy, while DM only weakly interacts. The collapsing visible matter then becomes a disc rather than a sphere because of the angular momentum barrier.
- We can also infer our local DM 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 often predicted to be "cuspy" (with a sharp increase at small radii), but observations indicate "cored" density profiles (flat at small radii). In addition, ΛCDM simulations seem to indicate more small scale structure than is actually observed. For example, there may be missing dwarf galaxies even though such galaxies should have inevitably formed; this is called the "too big to fail" problem. All of these issues are galactic in nature, and were recently reviewed here.

- These discrepancies might be explained by "bursty" outflows of baryonic matter, or by including DM self-interaction; such work is currently in progress. In particular, for GeV-scale DM, QCD-scale self-interaction cross sections are sufficient.

Note. MOND is an alternative hypothesis that explains galaxy rotation curves well. Here, 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 empirical 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 galaxy simulations with DM, but completely fails to match evidence from galaxy clusters and the CMB; the only MOND models that work here *also* use DM. Moreover, galaxy simulations are steadily improving with time, reducing the benefit that MOND provides. This explains the comparatively little interest in MOND: it only gives a minor benefit to the modeling of only one of many aspects of the universe. (However, there are theories of superfluid DM which reduce to MOND on galactic scales. Also, to be fair, one can get the right CMB power spectrum with MOND as long as one puts in a precisely chosen, bumpy initial power spectrum; in this case DM plus inflation has the advantage of working with generic initial conditions, i.e. being more predictive.)

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,

  $$\ddot{I} = 2T + V, \quad I = \frac{1}{2}\sum_i m_i \mathbf{r}_i \cdot \mathbf{r}_i$$

  where I is the virial, T is the total kinetic energy, and V is the total potential energy. In the steady state, $\ddot{I}$ averages to zero and we have

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

  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 multiply by 3 assuming spherical symmetry. We also assume $\rho(r)$ is proportional to the luminosity $L(r)$, allowing us to estimate the average potential energy. Combining these gives a total mass several times higher than the stellar mass inferred by a "census" estimate using luminosity.

- One also needs to know the distance to the cluster. Zwicky did this using the overall Doppler shift and Hubble's law. Though he used an incorrect value for the Hubble constant, overestimating the amount of DM, he did find the right qualitative conclusion.
- The common ground between cluster observations and rotation curves is that they use "tracers", visible collisionless objects that respond to the DM density. Stars are generally excellent tracers because they are very sparse, and hence collisionless.
- DM is also responsible for presence of the hot intracluster gas, which emits X-rays; if DM 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 indicate the same result.

Note. A proof of the virial theorem. We simply differentiate to find
$$\dot{I} = \sum_i \mathbf{r}_i \cdot \mathbf{p}_i, \quad \ddot{I} = \sum_i \mathbf{v}_i \cdot \mathbf{p}_i + \mathbf{r}_i \cdot \mathbf{F}_i.$$
The first term in $\ddot{I}$ is simply $2T$. To handle the second term, note that
$$\sum_i \mathbf{r}_i \cdot \mathbf{F}_i = -\sum_{i\neq j} \mathbf{r}_i \cdot \nabla_i V_{ji} = -\sum_{i<j} \mathbf{r}_i \cdot \nabla_i V_{ji} - \sum_{i>j} \mathbf{r}_i \cdot \nabla_i V_{ji}.$$
Swapping the index names on the second term, we get
$$-\sum_{i<j} \mathbf{r}_i \cdot \nabla_i V_{ji} + \mathbf{r}_j \cdot \nabla_j V_{ij} = -\sum_{i<j} (\mathbf{r}_i - \mathbf{r}_j) \cdot \nabla_i V_{ji}$$
where we used the fact that $V_{ij} = V_{ji}$ is a function of $\mathbf{r}_i - \mathbf{r}_j$. Now, for any power law potential $V_{ij} \propto (\mathbf{r}_i - \mathbf{r}_j)^n$, the term simplifies using the power rule to $-nV$, giving the result
$$\ddot{I} = 2T - nV.$$
But as long as we work in the CM frame, we expect $\dot{I}$ to be roughly constant over time, giving
$$\langle T \rangle = \frac{n}{2}\langle V \rangle.$$
Using the virial theorem, we can also estimate the speed of dark matter in the galactic halo,
$$v \sim \sqrt{\frac{GM_{\text{halo}}}{R_{\text{halo}}}} \sim \sqrt{G\,\frac{10^{12} M_\odot}{100\,\text{kpc}}} \sim 200\,\text{km/s}.$$
That is, we have $v/c \sim 10^{-3}$.

Note. For generic initial conditions, we estimate the virialization time to be on the order of the gravitational collapse timescale $t \sim 1/\sqrt{G\rho}$, which is on the order of the time it takes the galaxy to rotate; the Milky Way has rotated about 100 times since its formation. There is the possibility of a nonvirialized subcomponent to dark matter. For example, dark matter that has recently been pulled off a satellite galaxy is part of a "stream" with low velocity dispersion.

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

- Galaxies and galaxy clusters deflect light. This can be used to measure their masses, giving results in agreement with the observations above.
- Lensing observations can also be used to constrain the nature of dark matter. Baryonic DM could 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 DM density.
- 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 interstellar gas, as measured by X-ray spectroscopy, lagged behind most of the gravitational mass, indicating the presence of weakly-interacting DM. (The stars are effectively collisionless and kept going, mostly following the DM.) This was regarded by many as the "nail in the coffin" for MOND.
- However, note that there also are "anti-Bullet cluster" systems, such as the "train wreck cluster" Abell 520, where the DM is in the middle with the interstellar gas; it's not clear how these are to be interpreted, though the data are in some doubt. The Bullet cluster is also unusual, as it has exceptionally high velocity dispersion. Finally, bounds on DM self-interaction from the Bullet cluster are in tension with the self-interaction needed to explain galaxy anomalies.

Finally, cosmological observations give us crucial independent data.

- 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 DM.
- Stringent independent constraints on Ω_b come from the baryon acoustic oscillation peaks in the CMB, and from the results of BBN, which both yield the same number.

- The DM 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 yet 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. For instance, TeVeS does not produce stable stars.

6.2 Models of Dark Matter

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

- DM candidates can be divided into "cold" and "hot", depending on their typical velocities. Since the 1990s we have known that most, if not all DM must be cold, as hot DM could "free-stream" to smooth out small mass fluctuations. Concretely, with cold DM one can first form small structures that merge into larger ones; with hot DM one would first form large structures that fragment into smaller ones. Observations allow only 1% of the DM to be hot.
- One could also consider warm DM, which only suppresses small structures; one could test for this by looking for a cutoff in the matter power spectrum. Other constraints on warm DM come from Lyman α-forest absorption, which provides a map of nearby ($z \sim 5$) hydrogen gas.
- If it's assumed that DM is in thermal equilibrium with the SM thermal bath, then lighter DM particles would automatically be hot. However, there are many nonthermal DM production mechanisms, such as the misalignment mechanism for axions.
- The interactions of DM with the SM are so far consistent, astrophysically, with being purely gravitational. Any interactions present must be weak, as they would otherwise contradict, e.g. Bullet cluster observations, and so on. A rough limit for the scattering cross section is $\sigma/m \lesssim (100\,\mathrm{MeV})^{-3}$. However, adding weak interactions may increase the accuracy of numerical simulations of galaxies.
- Decay of DM must be slow. At the simplest level, it must decay slowly enough to still be present today; if the decay is to visible particles, one has stronger constraints from the non-observation of such particles, while if the decay is invisible, one also has stronger constraints due to bounds on how much the DM density can change over time. However, a small amount of decay or annihilation could be used to explain some standing anomalies involves excess detections of energetic SM particles. These should be taken with a grain of salt, because many such anomalies exist, and they frequently vanish.
- DM must have a very small charge, if any at all. The simplest reason is that we can't see it, but stronger constraints come from, e.g. the requirement that the DM be completely decoupled from the thermal bath at recombination. Depending on the mass range, the bound on charge is about $10^{-6}e$. Such a small charge could result from a normal-sized charge under a $U(1)_{A'}$ which weakly kinetically mixes with $U(1)_A$.
- In order to fit with CMB observations, DM must be present by $z \sim 1000$. This eliminates many candidates that form by collapse at late times. However, primordial black holes are a possibility, as they are formed right after inflation ends.

- One generic bound on fermionic DM comes from the exclusion principle. We have

$$\frac{n}{p^3} \sim \frac{\rho_{\mathrm{DM}}/m_{\mathrm{DM}}}{(m_{\mathrm{DM}}v)^3} \lesssim 2$$

  which yields $m_{\mathrm{DM}} \gtrsim 1\,\mathrm{keV}$. This is the Tremaine–Gunn bound. However, bosonic DM can be substantially lighter. On the other hand, composite DM can range up to planetary masses, above which there are lensing bounds, leaving a range of many orders of magnitude in mass.

We now briefly consider some dark matter candidates.

- One can put many candidates into two classes: light cold bosonic DM (below keV scale) and heavy WIMP-like DM (above MeV scale) produced by freeze-out. Some of these candidates are covered in more detail in the notes on the Standard Model.
- The lightest stable supersymmetric particle, typically a neutralino (but also possibly, e.g. a "mixed sneutrino" or a gravitino), 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 "natural" parameter space.
- Neutrinos could account for the DM density if their mass was several eV and they come into thermal equilibrium. This made neutrinos a leading DM candidate in the 1980s, when neutrino masses were being established. However, such neutrinos would be hot, and so could not make up the DM by themselves. While we don't know the masses of the three known neutrinos, they are probably much less than an eV. However, there could be sterile neutrinos at this mass.
- Sterile neutrinos that don't come into thermal equilibrium could account for the DM density if their mass is on the keV scale; they can be produced via neutrino oscillations in the Dodelson–Widrow or Shi–Fuller mechanisms. Resonant scattering off such sterile neutrinos has been used as an explanation for a deficit of 3.5 keV photons emitted from galaxy centers. But they are also strongly constrained, as their decay to normal neutrinos would produce X-rays that haven't been detected.
- Axions are typically extremely light, but can become cold DM because they have a nonthermal production mechanism, the misalignment mechanism. This produces axions in a BEC-like state, with very low velocity dispersion.
- One could also consider axions in a supersymmetric theory, in which case they are part of a chiral supermultiplet. There is another real scalar field, the saxion, and a fermion called the axino. If the axion is the phase of a complex scalar field, the saxion can be thought of as the magnitude. The saxion decays quickly, but the axino can be another WIMP-like DM candidate.
- We know that late-forming black holes can't be the DM, because this would be inconsistent with the baryon density measured earlier in the universe. However, "primordial" black holes, formed before BBN, could be the DM. They could be formed by collapse of large density fluctuations at the end of inflation, e.g. if there is a violent phase transition. There are many astrophysical and cosmological constraints, which rule out most but not all possible masses.

We now list some twists on the collisionless cold dark matter (CCDM) paradigm.

- Warm DM, as mentioned above, smooths out small scale structure.

- Self-interacting DM (SIDM) makes the later stages of structure formation significantly more complicated. Scattering can alleviate the cusp-core problem, and also strips the halos from small clumps of DM orbiting larger structures, reducing their number. It has also been invoked to explain the "train wreck cluster".
- In particle physics, SIDM can also stand for strongly interacting DM. The reason is that the cross sections needed to make self-interacting DM work are of order $\sigma/m \sim \Lambda_{\mathrm{QCD}}^{-3}$, which is typical for a particle interacting under the strong force.
- Self-annihilating DM, as mentioned above, can explain standing anomalies and helps alleviate the cusp-core problem, as does decaying DM.
- Fuzzy DM is extremely light bosonic DM with a de Broglie wavelength on the order of the galaxy size; this could alleviate the cusp-core problem. Using the uncertainty principle $\Delta x \Delta p \gtrsim 1$ with $\Delta p \sim m_{\mathrm{DM}} v$ and Δx the size of a dwarf galaxy halo gives $m_{\mathrm{DM}} \gtrsim 10^{-22}\,\mathrm{eV}$.
- Repulsive DM is light bosonic DM which forms a condensate with a repulsive self-interaction, also addressing the cusp-core problem. The idea of "superfluid DM" falls in this category.
- These options are distinguished by their detailed predictions. For example warm DM, fuzzy DM, and repulsive DM all pick out preferred length scales, while decay DM picks out a time scale, and self-interacting and self-annihilating DM pick out densities.

There are many DM production mechanisms beyond freeze-out, and we list some here.

- In the freeze-in scenario, we suppose that the DM is never in thermal equilibrium, but rather has zero abundance in the early universe; the abundance rises monotonically to the desired value throughout cosmological history. Such an initial condition can be achieved if the DM interaction with the SM is very weak, and it is not produced during reheating.
- The freeze-in scenario has different parametrics (the final abundance increases with interaction strength, as opposed to decreasing for freeze-out) and accommodates a wide range of masses. The relevant DM particles are called feebly interacting massive particles (FIMPs).
- Note that if the production rate Γ is set by a dimensionless coupling $g \ll 1$ rather than a suppression by a mass scale, then we have $\Gamma \sim g^2 T$. Since $H \sim T^2$, this means production gets more important later in cosmological evolution. This is nice because the final abundance isn't determined by the reheat temperature, which is free.
- One can modify the freeze-out scenario by considering more DM interactions. The relevant reactions are:
  - Annihilation: $\chi + \chi \leftrightarrow \mathrm{SM} + \mathrm{SM}$.
  - Elastic scattering: $\chi + \mathrm{SM} \leftrightarrow \chi + \mathrm{SM}$. (Note that this can decouple significantly later than annihilation, even though the amplitudes are related by crossing, because the rate is suppressed by only one power of n_χ. Here, "elastic" means that the colliding particles retain their identities after the collision.)
  - Self-annihilation or "cannibalization": $\chi + \chi \leftrightarrow n\chi$ with $n \geq 3$. (This typically also implies DM self-interactions $\chi + \chi \leftrightarrow \chi + \chi$, which are useful for astrophysical reasons.) For an isolated DM bath, this causes slow cooling, as rest energy is cannibalized to kinetic energy.

In the standard WIMP freeze-out scenario, self-annihilation isn't present. The abundance is fixed by when the annihilation process decouples. In several newer ideas, the annihilation process is weak and decouples early, while the other processes continue to be effective.

- In strongly interacting massive particle (SIMP) DM, the SM and DM baths are kept in thermal equilibrium by the elastic scattering process. Thus, as the self-annihilation process goes on, energy is removed from the dark sector by elastic scattering, until the self-annihilation process decouples. The final DM abundance is sensitive to the DM mass and the self-annihilation rate, with typical target masses of $m_{\rm DM} \sim 100\,{\rm MeV}$, motivating "light dark matter".
- In elastically decoupling relic (ELDER) DM, the elastic scattering process decouples before self-annihilation. Self-annihilation then keeps the DM bath at approximately a constant temperature for a long period, until it too decouples. It turns out this makes the final DM abundance sensitive to only the elastic scattering cross section, leading to different parametrics.
- Ultralight bosonic DM can be produced nonthermally through the misalignment mechanism, with its abundance thus fixed by initial conditions, or by inflationary fluctuations. Other nonthermal production mechanisms include the decay of heavy particles or topological defects.
- Asymmetric DM is the idea that the relic DM abundance arises from an asymmetry between $\overline{\rm DM}$ and DM. (For example, in "cogenesis", this asymmetry arises simultaneously with the baryon-antibaryon symmetry.) This is a tempting idea since the DM and baryon densities are equal up to an $O(1)$ factor. In the simplest version of the idea, $n_{\rm DM}$ and n_B could exactly cancel, with $m_{\rm DM} \approx 5m_p$.

Example. A bit more about SIMP DM.

- We focus on SIMP DM with the self-annihilation process having $n = 3$. The rates are

$$\Gamma_{\rm ann} \sim n_{\rm DM}\langle\sigma v\rangle_{\rm ann}, \quad \Gamma_{\rm kin} \sim n_{\rm SM}\langle\sigma v\rangle_{\rm kin}, \quad \Gamma_{3\to2} \sim n_{\rm DM}^2\langle\sigma_{3\to2}v^2\rangle.$$

  We can parametrize these rates in terms of couplings,

$$\langle\sigma v\rangle_{\rm kin} \sim \langle\sigma v\rangle_{\rm ann} \equiv \frac{\epsilon^2}{m_{\rm DM}^2}, \quad \langle\sigma v^2\rangle_{3\to2} \equiv \frac{\alpha_{\rm eff}^3}{m_{\rm DM}^5}$$

  where the $m_{\rm DM}$ dependence enters by dimensional analysis. We don't have to include powers of T because we have assumed the relevant rates are not dependent on the typical velocity v. This is just the usual assumption of s-wave annihilation for the two-body processes, and its analogue for the three-body annihilation.

- The story of SIMP DM holds as long as

$$\Gamma_{\rm kin} \gtrsim \Gamma_{3\to2} \gtrsim \Gamma_{\rm ann}$$

  until freeze-out. Assuming this is true, freeze-out occurs at

$$\Gamma_{3\to2} \sim H_F \sim \frac{T_F^2}{M_{\rm pl}}$$

  and by similar reasoning to the usual freeze-out calculation, means

$$m_{\rm DM} \sim \alpha_{\rm eff} T_{\rm eq}^{2/3} M_{\rm pl}^{1/3}$$

  where $T_{\rm eq} \sim 1\,{\rm eV}$ is the temperature of matter-radiation equality. Here we have used $T_F \sim m_{\rm DM}/20$ and dropped all numeric factors.

- Assuming the coupling is strong, $\alpha_{\text{eff}} \sim 1$, this points to masses

$$m_{\text{DM}} \sim 100\,\text{MeV}$$

as stated above.

- Next, we check the bounds above. After freeze-out, we have

$$\rho_{\text{DM}} = m_{\text{DM}} n_{\text{DM}} \sim T^4 \frac{T_{\text{eq}}}{T}$$

since the DM density and radiation density are on the same order at T_{eq}. Then at freeze-out,

$$n_{\text{SM}} \sim T_F^3, \quad n_{\text{DM}} \sim n_{\text{SM}} \frac{T_{\text{eq}}}{m_{\text{DM}}}$$

which gives

$$\Gamma_{\text{ann}} \sim \frac{\epsilon^2 T_F^3}{m_{\text{DM}}^2} \frac{T_{\text{eq}}}{m_{\text{DM}}}, \quad \Gamma_{\text{kin}} \sim \frac{\epsilon^2 T_F^3}{m_{\text{DM}}^2}, \quad \Gamma_{3\to 2} \sim \frac{\alpha_{\text{eff}}^3 T_F^6 T_{\text{eq}}^2}{m_{\text{DM}}^7}.$$

- By using $T_F \sim m_{\text{DM}}/20$ and plugging in our expression for m_{DM} above, our bounds become

$$\alpha_{\text{eff}}^{1/2} \left(\frac{T_{\text{eq}}}{M_{\text{pl}}}\right)^{1/3} \lesssim \epsilon \lesssim \alpha_{\text{eff}} \left(\frac{T_{\text{eq}}}{M_{\text{pl}}}\right)^{1/6}.$$

For $\alpha_{\text{eff}} \sim 1$ this gives a possible range of three orders of magnitude for ϵ.

We now say a bit more about the relatively new paradigm of dark/hidden sectors.

- In the standard WIMP paradigm, we think of the DM as part of a sector that interacts strongly with the SM, such as the superpartners in SUSY. By contrast, a "dark sector" is a group of related particles, one of which is the DM, which interacts only weakly with the SM.
- In principle, a dark sector doesn't have to interact with the SM at all, except gravitationally. This is the worst-case scenario; on the other hand, there is reason for optimism, because concrete production mechanisms generally require nongravitational interactions.
- For concreteness, let's suppose that pairs of DM particles couple to a mediator, which in turn couples to pairs of SM particles. In this scenario we can have freeze-out production, where the mediator takes the place of the weak boson for WIMPs.
- The relic abundance scales as $m_{\text{DM}}^2/\epsilon^2 g_D^2$, where ϵ is the mediator's coupling to the SM, and g_D is its coupling to DM. Taking g_D to be $O(1)$ and adjusting ϵ, we can accommodate a range of DM masses, from the keV to TeV scale. In SUSY, which in this context is just a concrete way to construct the hidden sector, this was called the "WIMPless miracle". In principle g_D can be adjusted too, but lowering it just increases ϵ, so taking a strong dark sector coupling is a conservative assumption.
- For $m_{\text{DM}} \lesssim$ MeV, DM annihilates after neutrinos decouple, which would be detected through N_{eff}. Thus, the window for "light" DM is MeV to GeV.

- Light DM is difficult to search for in direct detection experiments, because such particles deposit little energy in a collision. And since the couplings are low, it is difficult to see the production of such DM in the collider "energy frontier". Instead, for colliders one needs to go to the "intensity frontier", e.g. in beam dump experiments.

- Stepping back, we can ask which interactions between the hidden sector and the SM are "generic". One answer is that it corresponds to the lowest-dimension allowed operators:

  - For a vector, there is the dark photon coupling $\epsilon B_{\mu\nu} F'^{\mu\nu}$ where F' is the gauge field strength of a dark photon, and B is the $U(1)_Y$ gauge field strength. More general possibilities are also possible with more model building; for example, there are several motivations to gauge $U(1)_{B-L}$: it is non-anomalous in the SM provided sterile neutrinos are introduced, and it can be used to construct R-parity in SUSY models. However, the standard dark photon is distinguished by its minimality.
  - There is nothing similar for a pseudovector, because the SM has no conserved axial currents; coupling to a nonconserved current would yield issues with unitarity. Of course, a dark pseudovector can be accommodated with further model building, which is used for the X17 boson to explain the Beryllium anomaly.
  - For a pseudoscalar, there are again no options in the SM. However, if we allow dimension 5 operators, we have the axion couplings $(a/f_a)F_{\mu\nu}\tilde{F}^{\mu\nu}$, $(a/f_a)G_{\mu\nu}\tilde{G}^{\mu\nu}$, and $(\partial_\mu a/f_a)\overline{\psi}\gamma^\mu\gamma^5\psi$. Since the axion is so well-motivated, this is usually included in the standard list.
  - For scalars, we have the dark Higgs couplings, $(\mu S + \lambda S^2)H^\dagger H$, where the first term isn't allowed for complex scalars. This leads to couplings to SM fermions proportional to their mass. If we allow dimension 5 operators, we can multiply the scalar with any marginal SM operator; in this context, the scalar is often called a "dilaton" in analogy to string theory, since in both cases its value shifts fundamental constants.
  - For fermions we have $y_N LHN$, where N is called a sterile neutrino, or heavy neutral lepton.

  The dark photon, axion, dark Higgs, and sterile neutrino are the "portals". For concreteness, we'll focus on dark photons.

- The dark photon coupling kinetically mixes the SM photon and the dark photon. Upon diagonalizing the kinetic terms with a field redefinition, we can show that the physical degrees of freedom are the familiar photon, along with a massive vector which couples strongly to DM particles, and to SM particles proportional to $\epsilon e Q$.

- The way we would search for such a dark sector depends on the mass of the dark photon.

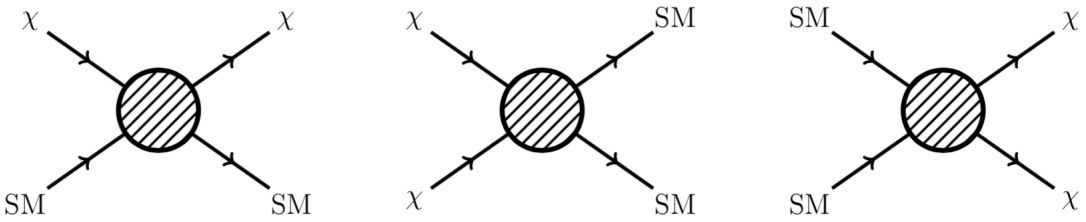
  - If it's in the same "light" mass range as the DM particle, then we could look for its production in a collider experiment. In this case the DM particle isn't so important, except that it sets a target for the dark photon coupling.
  - If it's very light, then we can integrate it out. The result is that the DM particles effectively have very small "millicharges". The dark photon itself can also be searched for directly, e.g. through fifth force searches.
  - In this case, the dark photon itself could be the DM, produced by a nonthermal production mechanism like the axion. This "ultralight" DM could be searched for through precision experiments, such as DM Radio.

- To get a specific target for ϵ, note that if there is a heavy particle that couples to both the photon and dark photon, then it can induce this coupling at one-loop. Integrating it out gives $\epsilon \sim 1/16\pi^2 \sim 10^{-3}$, which gives a mass right in the middle of the light DM range.
- A dark photon kinetically mixing with $U(1)_{B-L}$ would be searched for in much the same ways as an ordinary dark photon, but the details of the couplings would be different, e.g it would couple equally strongly to neutrons and protons. We could also break family symmetry, e.g. by gauging $U(1)_{L_\mu - L_\tau}$, which would produce no coupling to electrons. On the other hand, these exotic dark photons still pick up a mixing with the SM photon at one loop, albeit a smaller one.
- A similar story holds for a "dark Higgs". Since it would couple to SM fermions proportionally to their mass, it is best probed by the decays of heavy mesons.
- In the standard hidden sector freeze-out scenario, we assume the DM is heavier than the mediator. Forbidden DM is a variant on freeze-out where the DM can only annihilate into *heavier* hidden sector particles. In this case, we have $\langle \sigma v \rangle \sim e^{-\Delta m/T}/m_{\rm DM}^2$, which can accommodate light dark matter if Δm is chosen correctly.
- Another twist is if the DM is heavier than the mediator, but pairs of DM particles couple to *pairs* of mediator particles. Then DM can annihilate to the mediator, which later decays to SM particles. In this case, the relic density scales as $m_{\rm DM}^2/g_D^4$, where g_D is the mediator-DM coupling. This is called "secluded" annihilation because the extra step means there are much weaker constraints on $m_{\rm DM}$ and g.
- Things could of course be much more complicated. DM could be strongly interacting in the dark sector, forming composite objects, such as dark atoms, dark molecules, dark hadrons, or dark nuclei. For example, the dark sector could be a confining pure Yang–Mills theory, in which case the DM is in the form of glueballs.

6.3 Direct Dark Matter Detection

Next, we briefly consider dark matter detection.

- The three main paradigms for DM 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 DM particles. Such experiments are placed deep underground to shield cosmic rays. They were first explored in the mid-1980s by Goodman and Witten. Many experiments have been done since then, placing stringent bounds on WIMP-nucleon cross sections.

- 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 DM 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 DM couping is strongly spin-dependent.

- Astrophysicists typically use indirect detection methods, searching for the produces of annihilating DM. Such signatures include an excess of antimatter, mono-energetic gamma rays, and high-energy neutrinos. The sensitivities of these experiments tend to be complementary to direct direction experiments, e.g. they are more sensitive to higher masses.

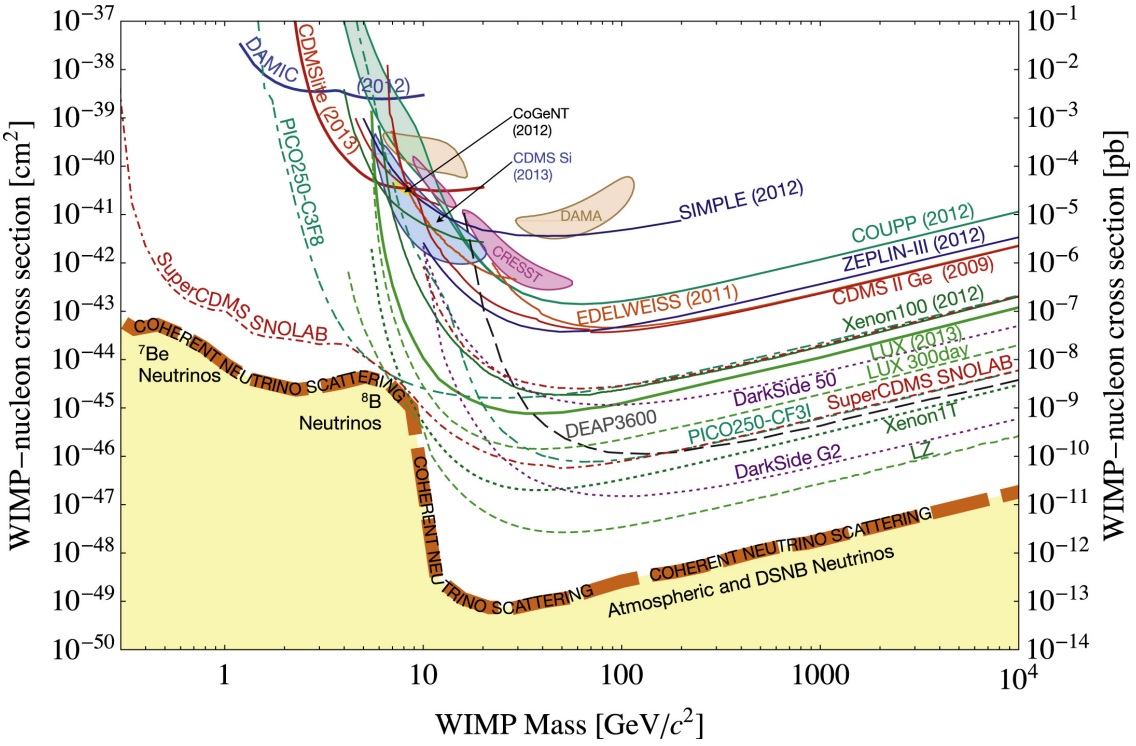
- TeV scale DM could also be produced at the LHC, with the classic signal being missing energy or momentum, such as in the production of a "mono-X" event, where X is a photon, jet, W, Z, h, t or so on. Here, one can take a "top-down" approach, computing results from a full theory such as the MSSM. Alternatively, one could use a simplified model, which gives more generic constraints but which may miss peculiarities that can occur in more complex models. This is done starting from a full model by just removing irrelevant particles, or setting couplings equal or to zero.

- For practicality, LHC results are often phrased in terms of these simplified models, which include the DM particle and the most important mediator. This leaves a reasonable number of parameters, primarily the DM mass, the mediator mass, and, e.g. the mediator's coupling to quarks, which fix the rates for processes such as mono-jets, mediator pair production, or qq scattering via a mediator.

- Various combinations are then possible for the spin and parity of the DM and mediator. For example, a typical figure caption is "axial mediator, Dirac DM, $g_q = 0.25$, $g_\chi = 1.0$" with the masses of the mediator and DM on the axes.

Next, we consider direct detection in more detail.

- Standard WIMP direct detection experiments consist of looking for recoils of WIMPs off nuclei. An exclusion plot is shown below.



The basic intuition is that at higher masses, the sensitivity to the cross section falls as $1/m$ simply because there are fewer DM particles. At lower masses, the sensitivity falls off because the energy deposited by each DM particle is smaller, so not all impacts can be seen; the curve here depends on the details of the high-velocity tail.

- Since the WIMPs would rotate along with the galaxy, there is both a daily modulation of the event rate, due to the spin of the Earth, and a yearly modulation, due to the orbit of the Earth. This is an important way to confirm a WIMP signal, as there are many possible backgrounds, such as radioactivity and cosmic rays. One can also directly measure the WIMP "wind" if the detector is directionally sensitive.

- Suppose a WIMP with mass m_χ and speed v hits a stationary nucleus with mass M. If the nucleus recoils at an angle θ from the original velocity of the WIMP, then some basic classical mechanics gives a recoil energy of

$$E = \frac{2\mu^2 v^2 \cos^2\theta}{M}, \quad \mu = \frac{M m_\chi}{M + m_\chi}.$$

The maximum recoil energy has the limits

$$E_{\max} \sim M v^2 \, \min(1, m_\chi^2/M^2)$$

For a standard weak-scale WIMP recoiling off a heavy nucleus, the two arguments above are of the same order. Taking $v \sim 10^{-3}$, we have $E_{\max} \sim 10 - 100\,\text{keV}$.

- Depending on the search, the impact is detected by the resulting ionization, light emitted, calorimetry (i.e. energy deposited into phonons/heat), or a combination of these options. These have varying advantages; for instance, calorimetry reliably gets all the deposited energy, but it is by far the slowest.

- Calculating the differential event rate dR/dE uses input from several fields of physics.
  - First, compute the amplitude for WIMPs to scatter off quarks or gluons. This is highly model-dependent.
  - Next, use results from nuclear physics to compute the scattering amplitude for protons and neutrons. Usually, we take these to be the same for simplicity; the final experimental bound will get reported in terms of that quantity.
  - Infer the amplitude for scattering off the heavy nuclei by summing over the nucleons. At finite momentum transfer, this requires using nuclear form factors, such as the standard simple Helm form factor.
  - Compute the scattering rate using the WIMP number density and velocity distribution, e.g. using the standard halo model
    $$f(v) = \frac{4}{\sqrt{\pi}} \frac{v^2}{v_0^3} e^{-v^2/v_0^2}.$$
    This is not exactly true, especially in the high velocity tail; for example, one crude correction would be to set it to zero for velocities greater than the escape velocity of the Milky Way. There is a great variety of halo models, which mostly differ in the tail.

  One can substantially change the results by upsetting some of the assumptions made above; this can be used to make the DAMA results consistent with other exclusions.

- An important distinction is between spin-independent and spin-dependent scattering. For spin-independent scattering, in the limit of low momentum transfers, the amplitude contributions of the nucleons add coherently, enhancing the cross section by A^2 where A is the atomic mass number. This works for a decent range of momentum transfers because of the low WIMP velocity. (However, for higher momentum transfers, we must account for a form factor suppression.) The standard WIMP exclusion plots assume spin-independent scattering.

- On the other hand, for spin-dependent scattering, the contributions of nucleons with opposite spins cancel. Since heavy nuclei generally have spin much less than A, this leads to a huge suppression. (DAMA's NaI target is much more sensitive to spin-dependent couplings than other targets, which is why its positive result is not flatly impossible.)

- In general, it's important to have good resolution at low energies, because dR/dE is always monotonically decreasing. It is especially important for experiments detecting light DM, which require light targets and very low energy thresholds.

Example. Suppose the WIMP χ is a Majorana fermion, e.g. the lightest neutralino in a SUSY model. There are two possible leading couplings to quarks: the spin-independent scalar coupling
$$\mathcal{L} \supset \overline{\chi}\chi\overline{q}q$$
which could arise from Higgs exchange, and the spin-dependent axial-vector coupling
$$\mathcal{L} \supset \overline{\chi}\gamma^\mu\gamma_5\chi\overline{q}\gamma_\mu\gamma_5 q$$
which could arise from Z exchange. If χ is a Dirac fermion, we could also have the spin-independent vector coupling
$$\mathcal{L} \supset \overline{\chi}\gamma_\mu\chi\overline{q}\gamma^\mu q.$$
One can extend this discussion to spin 0 or spin 1 WIMPs. Standard WIMP exclusion plots assume a spin-independent scalar coupling.

6.4 Indirect Dark Matter Detection

Next, we consider indirect detection in more detail.

- Generically, two-body annihilation dominates since the density is low, and we assume it produces some SM particles which cascade decay to long-lived known particles. However, three-body or higher annihilation could be important if two-body annihilation is forbidden for some reason, or if we have "strongly-interacting massive particles" (SIMPs).
- Two-body annihilation scales quadratically with the DM density. Assuming a typical WIMP,

$$\langle \sigma v \rangle \sim \frac{1}{(100\,\mathrm{TeV})^2}$$

  from the WIMP miracle calculation, and in the simplest situations we can use this to infer the present-day annihilation rate. Of course, there are other options. The cross section could have strong velocity dependence due to resonances, or effects like the Sommerfeld enhancement, or the WIMP could have "coannihilated" with a particle which is no longer abundant.
- Another possibility is a one-body decay, which scales only linearly with the density. However, this is trickier because we don't have a benchmark for how fast this decay should be.
- One could get rough estimates using an EFT approach with GUTs. For example, if a weak-scale WIMP decays by a dimension 5/6/7 operator from the GUT scale, we would have

$$\tau \sim \begin{cases} m_{\mathrm{GUT}}^2/m_{\mathrm{DM}}^3 \sim 0.1\,\mathrm{s} & \text{dimension 5} \\ m_{\mathrm{GUT}}^4/m_{\mathrm{DM}}^5 \sim 10^{25}\,\mathrm{s} & \text{dimension 6} \\ m_{\mathrm{GUT}}^6/m_{\mathrm{DM}}^7 \sim 10^{51}\,\mathrm{s} & \text{dimension 7} \end{cases}$$

  which are definitely ruled out, possibly observable, and completely unobservable.
- There are a few qualitatively distinct options for the decay/annihilation products. If colored particles or taus are produced, they hadronize and produce many pions, which in turn decay to produce a continuum spectrum of gamma rays. If electrons or muons are produced, we end up with charged leptons and neutrinos, with only a few photons. Photons can be produced directly, giving a sharp spectral peak, but the branching ratio is expected to be small. Neutrinos can also be produced, but they're very hard to detect.
- Charged particles get strongly deflected by galactic magnetic fields, so experiments looking for such particles are effectively direction blind. Photons and neutrinos aren't deflected, but the detection rates are generically expected to be lower.
- Historically, many indirect detection experiments have reported anomalies which could be explained by DM, but then have been explained by astrophysical effects. This is especially difficult for charged particles, since they could be produced by essentially any nearby new astrophysical phenomenon, such as an unseen population of pulsars. For the uncharged particles, these sources can be localized and ruled out.

Now we focus on detection of uncharged particles.

- Consider DM annihilation in a volume element dV at a distance R from the observer. If the spectrum produced per annihilation is dN/dE, then the spectrum produced per time is

$$\frac{1}{2}\frac{\rho^2}{m_{\text{DM}}^2}\langle\sigma v\rangle\frac{dN}{dE}\,dV$$

  where ρ is the DM density and the $1/2$ is a symmetry factor. (However, if DM comes in particles and antiparticles symmetrically, then we lose the symmetry factor, but we convert ρ^2 to $(\rho/2)^2$, so we get an overall factor of $1/2$. These results also change for "asymmetric" DM.)

- The spectrum per unit time incident on a detector per unit area dA is found by dividing by $4\pi R^2$ and integrating over dV. This quantity splits neatly into a "particle physics" part and "astrophysics" part,

$$\frac{dN_{\text{obs}}}{dE\,dt\,dA} = \frac{1}{8\pi m_{\text{DM}}^2}\langle\sigma v\rangle\frac{dN}{dE}\int\rho^2\,dR\,d\Omega$$

  where the integral is called the J-factor. For decaying DM, a similar argument gives

$$\frac{dN_{\text{obs}}}{dE\,dt\,dA} = \frac{1}{4\pi m_{\text{DM}}}\frac{1}{\tau}\frac{dN}{dE}\int\rho\,dR\,d\Omega$$

  Note that conventions differ on where to put the factors of $1/8\pi$ or $1/4\pi$.

- Here we have assumed that the cross-section does not depend too strongly on the velocity; otherwise we have to put the $\langle\sigma v\rangle$ under the integral. Also, if we look at distant targets, one should also account for redshift; the line of sight integral dR should be recast as a dz integral.

- Computations of the J-factor depend on the halo model used, especially whether it is cuspy or cored. This makes a significant difference for annihilation because scales quadratically with the density, and hence dominates in the galactic center region. One can estimate local DM densities using oscillations of stars above and below the galactic plane, but this is difficult at the galactic center because one must subtract off a large baryonic matter background.

- For concreteness, plugging in typical WIMP numbers with a cuspy NFW profile gives

$$\frac{dN_{\text{obs}}}{dE\,dt} \sim 10^{-9}\,\text{cm}^{-2}\text{s}^{-1}$$

  for the galactic center. We need about a year with a $100\,\text{cm}^2$ detector to see one photon.

- There are several places that one could look, with various advantages and disadvantages.
  - Dwarf galaxies are nearby and have low background, and thus often give the strongest constraints on annihilation.
  - The galactic center has high signal, giving it a high potential for discovery. However, it has high background, and the J-factor is uncertain.
  - The galactic halo is large and nearby, but has complex backgrounds.
  - Other galaxies and galaxy clusters carry a huge amount of DM and also hold redshift information, but are further away. Since they contain a huge total amount of DM, they often give the strongest constraints on decay. However, the J-factor is uncertainty since it depends on DM substructure in the cluster.

  - Seemingly empty space could produce a signal, depending on the DM substructure, but it's hard to know where to look.
  - Extragalactic background radiation also carries substantial redshift information and averages over a huge region.

- For gamma rays, there are a variety of experiments providing exclusion bounds.
  - From 100 MeV to 1 TeV: Fermi and DAMPE. These are in space, with a relatively small area but wide field of view.
  - From 100 GeV to 10 TeV: HESS, VERITAS, MAGIC. These are ground-based telescopes, with a narrow field of view.
  - From 1 TeV to 100 TeV: HAWC, and later CTA and LHAASO. These ground-based apparatuses have a wide field of view, since they search for the Cherenkov radiation produced when the gamma rays pass through a tank of water.

- For neutrinos, there's also a variety of experiments.
  - SuperK probes neutrinos from a few MeV to 1 TeV.
  - ANTARES reaches 100 GeV to 100 TeV.
  - IceCube reaches 100 GeV to 10^9 GeV.

- The strongest constraints on GeV-scale DM that annihilates to photon-rich channels come from Fermi looking at gamma rays from dwarf galaxies. The analysis is done by fitting "templates" (i.e. Poisson distributions due to background and candidate signal) using maximum likelihood. A combined analysis rules out thermal relic WIMP DM for masses up to ~ 100 GeV, subject to standard assumptions, though the precise bound depends on the annihilation channel.

- Constraints on decaying DM can can probe DM masses up to 10^{10} GeV, or even Hawking radiation from primordial black holes. We can also cover masses down to the keV range using X-ray telescopes such as Chandra, XMM-Newton, INTEGRAL, HXMT, Astrosat, NuSTAR, and Suzaku, or the proposed XRISM, XPoSat, STROBE-X, Athena, and Lynx.

- According to Fermi data, there is a highly significant excess of GeV-scale photons from the galactic center, first noticed by Goodenough and Hooper in 2009. Particle explanations of this excess would have masses in the range $7 - 10$ GeV and are generally called "hooperons", but the excess could also be explained by a large population of pulsars. Emission from pulsars would be more "clumpy", but whether the data reflects this is a matter of current controversy which depends on fine statistical details.

- There is also an outstanding 4σ excess of 3.5 keV photons from galaxy clusters, found by XMM-Newton in 2014, which has been confirmed by other experiments. The simplest DM explanation is a decaying 7 keV sterile neutrino, but that hypothesis is in tension with other constraints. In addition, there are many possible non-DM sources in this energy range.

Next, we consider detection of charged particles.

- Charged cosmic rays diffuse through the galactic magnetic field, losing energy as they propagate and escaping at the boundaries; both of these effects soften the spectrum. Protons are usually

in the "diffusion-dominated" regime, while electrons and positrons are in the "loss-dominated" regime, sampling only a local regime of about 1 kpc. In practice, one uses specialized programs such as GALPROP or DRAGON, or pregenerated results such as PPPC4DMID.

- It is generally better to look for antimatter because DM decay should produce it equally as often, but the background should be much lower; examples include antiprotons, antideuterons, and positrons. Antimatter can be distinguished from matter by how it curves in a magnetic field. If we're using matter, we can still distinguish signal from background, but we need to use spectral features. In general, the backgrounds for charged particles are much harder to model than for photons, so while the bounds tend to be stronger, systematic uncertainties are high.
- The leading experiment for cosmic rays of GeV energies and up is AMS-02, on the International Space Station, and for the purposes of DM detection its sensitivity to antiprotons and positrons is most important.
- Some surprisingly strong limits for sub-GeV cosmic rays come from the Voyager spectrometer. These cosmic rays are usually deflected by the solar wind, but Voyager can see them since it is beyond the heliopause. This provides the best limits on 10 MeV to 1 GeV DM annihilating or decaying to electrons and positrons.
- Even just considering AMS-02, there are many outstanding anomalies.
  - There is a very large, long-standing excess of TeV-scale positrons, which was also seen by AMS-02's predecessor Pamela. But it's so large that it's hard to explain with DM, so it's likely due to a nearby population of pulsars.
  - There is an excess of antiprotons around 100 GeV at 4.5σ, but the statistical significance is disputed because of possible correlations between energy bins.
  - There are tentatively a few events with antihelium, which is supposed to be extremely rare from astrophysical sources. However, it's also tough to get an antihelium signal from DM.

We can also consider "very indirect" DM detection.

- Annihilating or decaying DM would have steadily injected energy into the universe throughout its history. This can modify the results of BBN, and also change the ionization and temperature history, which affects the CMB and 21 cm radiation. Such limits are clean because they don't depend on galactic astrophysics.
- Extra ionization during the cosmic dark ages ($z \sim 10 - 1000$) acts as a "screen" for CMB photons. To estimate the resulting bound, note that during radiation domination and after freeze-out, the fraction of DM that annihilates per Hubble time is

$$n\langle\sigma v\rangle H \propto T$$

  and this quantity is order 1 at freeze-out itself.
- Thus, at recombination the fraction is small, $T_{\text{rec}}/T_f \sim 10^{-9}$. On the other hand, the mass-energy of one DM particle is enough to ionize $(1\,\text{GeV})/(13.6\,\text{eV}) \sim 10^8$ hydrogen atoms, assuming all the energy goes into ionization, so at matter-radiation equality a tiny fraction of the DM particles annihilating gives a large effect.

- Planck is sensitive to changes in the ionization of order $\sim 10^{-4}$, so ionization provides a strong bound, especially for lower DM masses. Planck rules out the thermal relic benchmark for masses below 10 GeV, though the precise bound depends on the annihilation channel. Ionization also places constraints on decaying DM, which are especially strong from MeV to GeV.

- Annihilation can also heat the universe, and energy injection during recombination distorts the blackbody spectrum of the CMB. However, at matter-radiation equality the change in temperature would be of order $T_{\rm rec}/T_f \sim 10^{-9}$, which is too small for Planck to probe. More promisingly, the energy can also go into heating the baryons in the form of hydrogen gas, which are vastly outnumbered by the photons.

- We can measure the gas temperature over time by measuring the redshifted 21 cm line, with more emission corresponding to a higher temperature. Without a DM effect, we expect the gas to be colder than the CMB after recombination, since nonrelativistic matter cools down more quickly, then heat back up at the cosmic dark ages; DM would appear as a higher temperature at earlier times.

- This line is currently probed by EDGES, HERA, LOFAR, MWA, PAPER, SARAS, SCI-HI, and will be probed further by DARE, LEDA, PRIZM, and SKA. EDGES has seen a deep absorption trough at $z \sim 17$, which is extremely surprising, but the measurement is very difficult and has not yet been confirmed by other experiments.

- The Lyman-alpha forest is a set of emission lines due to the Lyman-alpha transition in neutral hydrogen. It can thus be used to probe the amount of neutral hydrogen. While the CMB is better for probing extra ionization near the beginning of the cosmic dark ages, the Lyman-alpha forest is better for probing extra ionization near the end, before the usual reionization starts. It is not as useful for earlier times because when most hydrogen is neutral, such photons are very efficiently absorbed.

7 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.
- The temperature variations can be used to infer the matter distribution on the surface of last scattering, and hence the matter power spectrum at that time. However, this is somewhat subtle. Matter and radiation perturbations are related, since we have adiabatic perturbations,

  $$\delta_r = \frac{4}{3}\delta_m, \quad \delta_r = \frac{\delta\rho_r}{\rho_r} = 4\frac{\delta T}{T}$$

  which tells us that

  $$\frac{\delta T}{T} = \frac{1}{3}\delta_m.$$
- However, we must also account for the fact that the photons redshift when they climb out of the gravitational potential. Naively, we would have $\delta T/T = \delta\Phi/c^2$, where Φ is the potential, and we restored factors of c. However, part of this is cancelled by the simultaneous expansion of the universe, giving

  $$\frac{\delta T}{T} = \frac{\delta\Phi}{3c^2}.$$

  This is called the Sachs–Wolfe effect.
- Setting $c = 1$ again, $\delta\Phi$ and δ_m are related by the Poisson equation,

  $$\delta\Phi(\mathbf{k}) = -\frac{4\pi G}{k^2}\bar{\rho}a^2\delta_m(\mathbf{k})$$

  which means the Sachs–Wolfe effect dominates for modes with $k \lesssim aH$ at last scattering, i.e. superhorizon modes.
- Isotropy tells us that it is useful to expand $\delta T/T$ in spherical harmonics,

  $$\frac{\delta T(\hat{\mathbf{n}})}{T} = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} a_{lm}Y_{lm}(\theta,\phi), \quad Y_{lm}(\theta,\phi) = N_{lm}e^{im\phi}P_l^m(\cos\theta)$$

  where the N_{lm} make the spherical harmonics normalized on the sphere. Intuitively, the spherical harmonics have $O(1)$ values on the sphere, but vary on angular scales $\theta \sim \pi/l$. The Sachs–Wolfe effect dominates for $l \lesssim 50$.

- If we have Gaussian perturbations as stated above, the coefficients a_{lm} will be uncorrelated,

$$\langle a_{lm} a^*_{l'm'} \rangle = C_l \delta_{ll'} \delta_{mm'}$$

  where the multipole moments C_l have no m-dependence by isotropy.

- To relate this to the two-point correlation function of the temperature, note that for two directions with $\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}' = \cos\theta$,

$$\frac{\langle \delta T(\hat{\mathbf{n}}) \delta T(\hat{\mathbf{n}}') \rangle}{T^2} = \sum_{lml'm'} \langle a_{lm} a_{l'm'} \rangle Y_{lm}(0,0) Y_{l'm'}(\theta, 0) = \sum_l \frac{2l+1}{4\pi} C_l P_l(\cos\theta)$$

  where we used $P_l^m(1) = \delta_{m0}$ and $N_{l0} = \sqrt{(2l+1)/4\pi}$.

- For some intuition, the functions $P_l(\cos\theta)$ are about 1 for $\theta \lesssim 1/l$, and oscillate rapidly with lower magnitude for $\theta \gtrsim 1/l$. Thus, the temperature correlation function above is parametrically $l_0^2 C_{l_0}$ where $l_0 \sim 1/\theta$. Alternatively, the variance of the temperature goes roughly as $\int l^2 C_l \, d\log l$.

We now relate this to the observed CMB data.

- As motivated above, we plot the combination $l(l+1)C_l$. Note that the error bars are larger at small l. This "cosmic variance" effect is because there are fewer of the coefficients a_{lm} available to average over.

- For small l, the Sachs–Wolfe effect dominates. By decomposing $\delta\Phi(\mathbf{k})$ into spherical harmonics, one can show that

$$C_l = \frac{16\pi T^2}{9} \int dk \, k^2 P_\Phi(k) j_l^2(kr)$$

  where $j_l(kr)$ is a spherical Bessel function. For the Harrison–Zel'dovich spectrum, this implies that $C_l \sim 1/l(l+1)$, which is indeed observed in the left part of the plot above, where the curve is flat. The slight tilt of the curve can be used to establish $n_s \approx 0.97$, in agreement with direct measurements of the matter power spectrum.

- At higher l, we see a pattern of peaks and troughs, associated with baryon acoustic oscillations. The first peak at $l \approx 200$ represents an acoustic wave that had time to undergo a single compression before decoupling. Its l value gives us information about the curvature of the universe, and hence tells us that the universe is nearly flat.

- Historically, this was known by the 1990s, and at the time, it was regarded as a puzzle because the amounts of matter and radiation did not add up to the critical density. Thus, CMB observations lent early evidence to dark energy, which was regarded as established by supernova redshift measurements in 1998, though the CMB itself doesn't directly say anything about it.

- The next two peaks represent oscillations that decoupled at the first rarefaction, and the second compression, respectively. (Oscillations that decoupled in between a rarefaction and a compression, i.e. when at uniform density, correspond to the troughs.) These give information about the amount of baryonic and dark matter in the universe. More dark matter lowers all of the peaks, while more baryonic matter enhances the odd-numbered peaks **(why?)** .

- For high $l \gtrsim 10^3$, the power spectrum is suppressed. This is because photon decoupling takes a finite time, and during this process, photon diffusion wipes out small-scale fluctuations.

- Additional information can be extracted from the polarization pattern of the CMB, which is decomposed into E-modes and B-modes, as mentioned earlier. The E-modes arise automatically from Thomson scattering in an inhomogeneous plasma, and are in accordance with the standard theory of inflationary perturbations. The B-modes are expected to be much weaker, and haven't been observed yet, but would give information about the scale of inflation.

Note. More about the discovery of dark energy. Type Ia supernova begin as white dwarfs, which slowly siphon matter from a companion star. When the mass passes the Chandrasekhar limit, a supernova occurs. Since the initial masses of all type Ia supernova are similar, they can be used as standard candles. Strictly speaking, they don't all have the same luminosity, but much of the variance can be accounted for by examining how the luminosity changes over time, in a so-called light-curve shape analysis. Supernova are detected by looking for momentary brightenings of distant galaxies. Their Doppler shift can be used to compute a redshift, while their luminosity can be used to compute a distance. In a uniformly expanding universe, these two are proportional.

Each data point is quite noisy. However, in the 1990s, supernova surveys investigated about 100 supernova at redshifts up to about $z = 1.0$, and observed a deviation from linearity at 5σ, indicating accelerating expansion due to our current period of dark energy dominance. (Note that these observations take place substantially later in the universe than anything else covered in these notes!) Dark energy can be distinguished from a systematic effect, such as absorption of distant light by dust (i.e. a "tired light" hypothesis), because ΛCDM predicts that for $z \gtrsim 1.0$ we have matter domination, during which the expansion *decelerates*. However, measurements at such huge distances are extremely difficult, and ongoing. Another line of evidence for dark energy is that the accelerated expansion stunts the growth of galaxies late in the universe, which is indeed observed.

These results remain rather controversial to many. To theorists, dark energy appears difficult to account for in ordinary quantum field theory without tuning, leading to the cosmological constant problem, and it appears difficult to embed into string theory, according to the dS Swampland conjecture. Philosophers are convinced that dark energy is objectionable because of a great variety of vague, metaphysical reasons. Some cosmologists have objected to the original dark energy discovery papers, on the grounds that their analyses overestimated their statistical significance. Finally, late-time measurements of the Hubble constant (including supernova Ia measurements, but also measurements of quasars) indicate a higher value than that extrapolated from CMB measurements, leading to the "Hubble tension".

Note. None of the discussion above gives direct information about the topology of the universe as a whole. For example, it's possible that the universe is toroidal (i.e. "wraps around") on scales far larger than our Hubble patch. This is scientifically uninteresting, because it produces no observable consequences while making the math more complicated; hence the default assumption is a trivial topology. However, if the topology is on a smaller scale, predictions can be made. In 2003, it was conjectured that the small C_l at $l \lesssim 3$ could be explained by a finite, "Poincare dodecahedral" space, whose volume would be too small for such perturbations to "fit in". This leads to the prediction of "matched circles" in the temperature variations of the CMB, where we see the same part of the surface of last scattering in two different directions; however, a search in 2004 failed to find this.