The Physics of Inflation

A Course for Graduate Students in Particle Physics and Cosmology

by

Daniel Baumann

Contents

Preface

Unless we accept fine-tuning of initial conditions, the standard Big Bang cosmology suffers from the so-called horizon problem. This refers to the fact that the cosmic microwave background (CMB) at the time of decoupling naively consisted of about 10^4 causally disconnected patches. Yet, we observe an almost perfectly uniform CMB temperature field across super-horizon scales at recombination. A key goal of modern cosmology is to explain this large-scale uniformity without resorting to fine-tuning.

Cosmological inflation, an early period of accelerated expansion, solves the horizon problem dynamically and allows our universe to arise from generic initial conditions. At the same time, quantum fluctuations during inflation produce small inhomogeneities. The primordial seeds that grew into galaxies, clusters of galaxies and the temperature anisotropies of the CMB were thus planted during the first moments of the universe's existence. By measuring the anisotropies in the microwave background and the large-scale distribution of galaxies in the sky, we can infer the spectrum of the primordial perturbations laid down during inflation, and thus probe the underlying physics of this era. Over the next decade, the inflationary era – perhaps 10^{-30} seconds after the Big Bang – will thus join nucleosynthesis (3 minutes) and recombination (380,000 years) as observational windows into the primordial universe. However, while the workings of recombination and nucleosynthesis depend on well-tested laws of atomic and nuclear physics respectively, the 'physics of inflation' remains speculative. The Standard Model of particle physics almost certainly does not contain the right type of fields and interactions to source an inflationary epoch. To describe inflation we therefore have to leave to comfort of the Standard Model and explore 'new physics' possibly far above the TeV scale. Some of the boldest and most profound ideas in particle physics come into play at these scales.

Overview

The aim of this course is two-fold: In **Part I**, we give a first-principles introduction to inflation. We show how inflation classically solves the horizon problem, while quantum mechanically providing a mechanism to generate the primordial seeds for the large-scale structure of the universe. However, despite this phenomenological success of inflation, the microphysical cause for the inflationary expansion remains mysterious. In **Part II**, we discuss our current understanding of the physics of inflation. We will use this as a welcome excuse to learn some fascinating physics such as effective field theory, supersymmetry and aspects of string theory.

As an orientation for the reader, we give a brief road map of the course:

Part I: The Quantum Origin of Large-Scale Structure

The structure of **Part I** of the course is summarized in fig. 1. Let me make a few comments on the individual chapters:

At the heart of the horizon problem lies the fact that in the conventional Big Bang evolution the comoving Hubble radius $(aH)^{-1}$ is an increasing function of time. Inflation solves the horizon problem by reversing this behavior, making $(aH)^{-1}$ temporarily a decreasing function of time (see fig. 1). Fluctuations that naively seem out of causal contact at recombination hence become causally connected before inflation. In **Chapter 1**, we describe this solution to the horizon problem and the classical dynamics of inflation which underlies it.

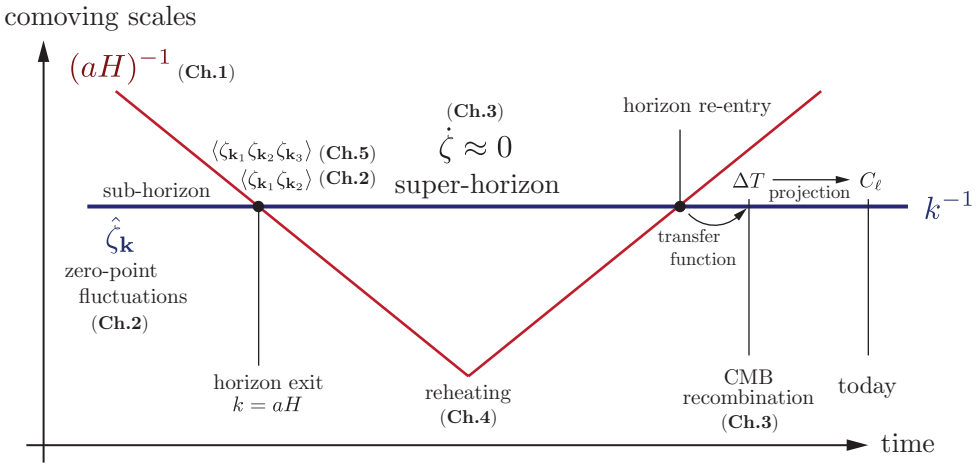


Figure 1: *A road map for Part I of the lectures.*

The shrinking comoving Hubble radius is also the key feature of inflation that allows quantum zero-point fluctuations on sub-horizon scales to become classical fluctuations on super-horizon scales (see fig. 1). Remarkably, the quantum-mechanical treatment of inflation leads to primordial fluctuations that are in striking agreement with the observed CMB anisotropies. In **Chapter 2**, we will learn about the intricacies of quantum field theory in curved spacetimes, such as the inflationary quasi-de Sitter spaces. This will culminate in a derivation of the primordial density fluctuations predicted by inflation. We present the calculation in full detail and try to avoid 'cheating' and approximations. We also explain why inflation predicts a stochastic background of gravitational waves.

To make contact with observations, the primordial inflationary fluctuations need to be evolved to late times, when they seed the CMB and the formation of large-scale structures (LSS). In **Chapter 3**, we provide this important link between the correlation functions computed at horizon crossing during inflation and late-time observables. We review Weinberg's proof that the curvature perturbation ζ freezes on superhorizon scales during single-field inflation. This allows us to relate our computation of fluctuations at horizon exit (high energies) to horizon re-entry (low energies), while remaining ignorant about the details of the physics of the time in between. The subsequent subhorizon evolution of the fluctuations is well-understood and computable in perturbation theory.

Although, reheating has no (or little) effect on the CMB anisotropies, it is an important component of any complete theory of inflation. In **Chapter 4**, we digress to review our current understanding of the reheating era. As we will see, reheating can be understood as a wonderful

application of quantum field theory in a time-dependent background and involves rich physics such as Bose condensation, parametric resonance and Schrödinger scattering.

One of the most active areas of current research in theoretical cosmology is the possibility of *non-Gaussianity* of the primoridial fluctuations. In **Chapter 5**, we give a detailed discussion of the computation of higher-order correlation functions during inflation. We discuss both quantum-mechnically sources of non-Gaussianity (computed in the *in-in* formalism) and classical sources of non-Gaussianity (computed in the δN formalism).

Part II: The Physics of Inflation

Part II of the course moves to a critical discussion of the mystery of the physical origin of the inflationary era:

One of the most powerful organizing principles of theoretical physics is the concept of *effective field theories* (EFT). In the absence of a UV-complete theory of high-energy physics, EFTs give us a way to discuss low-energy physics in a model-independent way, while including possible high-energy corrections systematically. In **Chapter 6**, we give an extensive discussion of the basic principles of EFT. Although we will later use inflation as our main example, much of the techniques that we will develop in this chapter have a much wider range of applicability. The slogan "if you can't understand it in effective field theory, then you haven't understood it" applies almost universally in theoretical physics. We end our treatment of EFT with a discussion of technical naturalness. We illustrate the concept of naturalness with the hierarchy problem of the Standard Model and the analogous eta problem of inflationary cosmology.

In **Chapter 7**, we give a brief discussion of the challenges of inflation as an effective theory. We describe the eta problem, large-field inflation and single-field non-Gaussianity, highlighting the extraordinary Planck-scale sensitivity of inflation. We discuss the role of global symmetries in considerations of naturalness in inflation.

In **Chapter 8**, we discuss supersymmetry (SUSY) as an attractive solution to the radiative instability of scalar fields. If the symmetry is unbroken, it controls dangerous radiative effects by enforcing exact cancellations between boson and fermion loops. Even if the supersymmetry is broken at low energies – as it has to be if it is to describe the real world – the appearance of supersymmetry at high energies can still help to regulate loop effects. We explain that in inflation, SUSY is often required to achieve technical naturalness of the quantum-corrected inflaton action. We illustrate these considerations with a detailed case study of supersymmetric hybrid inflation.

We end the course, in **Chapter 9**, we a brief survey of the state-of-the-'art' of inflation in string theory. We explain that the Planck-scale sensitivity of inflation provides both the primary motivation and the central theoretical challenge for realizing inflation in string theory. We illustrate these issues through two case studies: warped D-brane inflation and axion monodromy inflation.

Digressions, exercises and computational details are separated from the main text by horizontal lines. These parts can be omitted without loss of continuity, but they often illuminate the underlying physics.

Appendix C collects a number of instructive exercises.

Recommended Books and Resources

- **Books**:

  - Mukhanov, *Physical Foundations of Cosmology*.
  - Dodelson, *Modern Cosmology*.
  - Weinberg, *Cosmology*.
  - Mukhanov and Winitzki, *Introduction to Quantum Effect in Gravity*.
  - Terning, *Modern Supersymmetry*.
  - Peskin and Schroeder, *Introduction to Quantum Field Theory*.

- **Videos**:

  - Many of the lectures at the recent PiTP schools are excellent:
  http://video.ias.edu/pitp-2010 (e.g. Seiberg's lectures on SUSY) and
  http://video.ias.edu/pitp-2011 (e.g. Zaldarriaga's lectures on the CMB).
  - TASI 2009 videos can be found here:
  http://physicslearning2.colorado.edu/tasi/tasi_2009/tasi_2009.htm

- **Reviews**:

  - My TASI lectures (arXiv:0907.5424) have been superseded by the present notes.
  - Skiba, *TASI Lectures on Effective Field Theory*, (arXiv:1006.2142).
  - Luty, *TASI Lectures on SUSY Breaking* (arXiv:hep-th/0509029).
  - Chen, *Primordial Non-Gaussianity from Inflation Models* (arXiv:1002.1416).

- **Papers**:

  Many of the original papers on inflation are well worth reading:

  - Guth's famous paper (http://www.slac.stanford.edu/cgi-wrap/getdoc/slac-pub-2576.pdf) has a very clear description of the Big Bang puzzles.
  - Maldacena's paper (arXiv:astro-ph/0210603) contains the first rigorous computation of the three-point function for slow-roll inflation.
  - Kofman, Linde, and Starobinsky (arXiv:hep-ph/9704452) worked out everything we now know about (p)reheating.

Acknowledgements

I am most grateful to my friends and collaborators for explaining many of the concepts described in these notes to me, especially Liam McAllister, Daniel Green, and Matias Zaldarriaga.

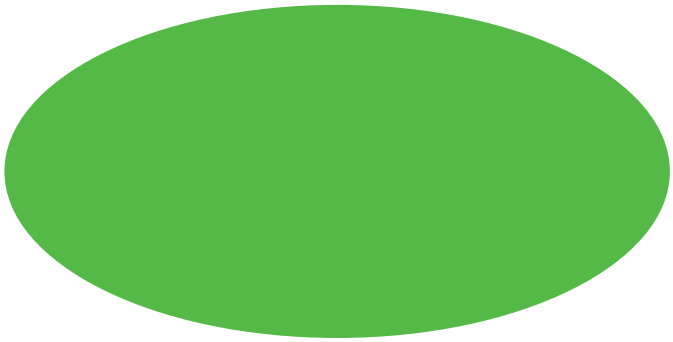
Part I

The Quantum Origin of Large-Scale Structure

1 Classical Dynamics of Inflation

1.1 Introduction

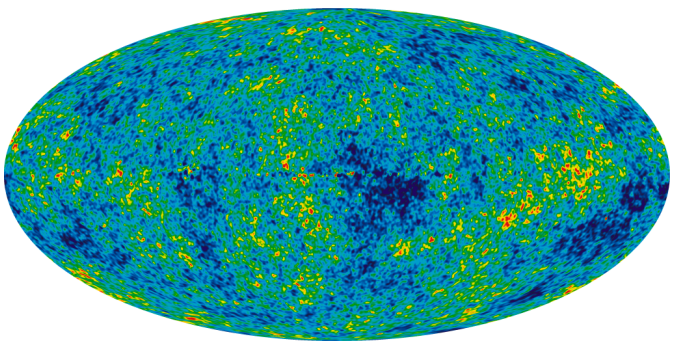
Running the expansion of the universe back in time, the uniformity of the CMB becomes a mystery. It is a famous fact that in the conventional Big Bang cosmology the CMB at the time of decoupling consisted of about 10^4 causally independent patches. Two points on the sky with an angular separation exceeding 2 degrees, should never have been in causal contact, yet they are observed to have the same temperature to extremely high precision:



This puzzle is the *horizon problem*.

As we will see, the horizon problem in the form stated above assumes that *no new physics* becomes relevant for the dynamics of the universe at early times (and extremely high energies). In this chapter, I will explain how a specific form of new physics may lead to a negative pressure component and quasi-exponential expansion. This period of *inflation* produces the apparently acausal correlations in the CMB and hence solves the horizon problem.

Remarkably, inflation also explains why the CMB has small inhomogeneities:



Quantum mechanical zero-point fluctuations during inflation are promoted to cosmic significance

as they are stretched outside of the horizon. When the perturbations re-enter the horizon at later times, they seed the fluctuations in the CMB. Through explicit calculation one finds that the primordial fluctuations from inflation are just of the right type—Gaussian, scale-invariant and adiabatic—to explain the observed spectrum of CMB fluctuations. This remarkable story will be told in the next chapter. This chapter will be setting the stage by explaining how the classical dynamics during inflation solves the horizon problem. An effort was made to keep this description as concise as possible. More details may be found in my previous lectures on the topic[1], as well as in the standard textbooks.[2]

1.2 The Horizon Problem

We start with a lightning review of FRW cosmology and the horizon problem[3] of the standard Big Bang scenario.

1.2.1 FRW Spacetimes

Modern cosmology begins with two observational facts: i) the universe is *expanding* and ii) on scales larger than 300 million light years the matter distribution is *homogeneous* and *isotropic*. In fact, as we go back in time the approximation of homogeneity and isotropy is expected to become increasingly accurate. The average spacetime is then described by the Friedmann-Robertson-Walker (FRW) metric[4]

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\Omega^2\right] \ , \tag{1.2.1}$$

where $k = 0$, $k = +1$ and $k = -1$ for flat, positively curved and negatively curved spacelike 3-hypersurfaces, respectively. For ease of notation we will restrict most of our discussion to the case $k = 0$.[5] In that case, the Friedmann equations for the evolution of the scale factor $a(t)$ are

$$H^2 = \frac{1}{3M_{\rm pl}^2}\,\rho \qquad \text{and} \qquad \dot{H} + H^2 = -\frac{1}{6M_{\rm pl}^2}(\rho + 3p) \ , \tag{1.2.2}$$

where $H \equiv \partial_t \ln a$ is the Hubble parameter and ρ and p are the density and pressure of background stress-tensor (here assumed to be a perfect fluid). To study the propagation of light (and hence the causal structure of the FRW universe) it is convenient to define *conformal time* τ via the relation

$$\mathrm{d}\tau = \frac{\mathrm{d}t}{a(t)} \ . \tag{1.2.3}$$

The FRW metric then factorizes into a static Minkowski metric $\eta_{\mu\nu}$ multiplied by a time-dependent conformal factor $a(\tau)$,

$$\mathrm{d}s^2 = a^2(\tau)\left[-\mathrm{d}\tau^2 + \mathrm{d}r^2 + r^2\mathrm{d}\Omega^2\right] \ \equiv \ a^2(\tau)\,\eta_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu \ . \tag{1.2.4}$$

[1] D. Baumann, *TASI Lectures on Inflation* (arXiv:0907.5424).

[2] Mukhanov, *Physical Principles of Cosmology*; Dodelson, *Modern Cosmology*; Weinberg, *Cosmology*.

[3] Inflation is sometimes motivated by listing a host of other problems, such as the flatness problem, the monopole problem, the entropy problem, etc. However, in my opinion, these are all just close cousins of the horizon problem. In other words, any solution to the horizon problem is likely to solve these secondary problems as well.

[4] Throughout these notes I will set the speed of light equal to unity, $c \equiv 1$.

[5] A flat universe is in fact favored by present observations (see fig. 1.5). Furthermore, as we will explain, it is a fundamental prediction of 60 e-folds of inflationary expansion.

1.2.2 Causal Structure

The radial propagation of light is characterized by the following two-dimensional line element

$$\mathrm{d}s^2 = a^2(\tau)\left[-\mathrm{d}\tau^2 + \mathrm{d}r^2\right] \ . \tag{1.2.5}$$

Just like in Minkowski space, the null geodesics of photons ($\mathrm{d}s^2 \equiv 0$) are straight lines at $\pm 45^\circ$ angles in the τ-r plane

$$r(\tau) = \pm\tau + const. \tag{1.2.6}$$

The maximal distance a photon (and hence any particle) can travel between an initial time t_i and later time $t > t_i$ is

$$\Delta r = \Delta\tau \equiv \tau - \tau_i = \int_{t_i}^{t} \frac{\mathrm{d}t'}{a(t')} \ , \tag{1.2.7}$$

i.e. the maximal distance travelled is equal to the amount of conformal time elapsed during the interval $\Delta t = t - t_i$. The initial time is often taken to be the 'origin of the universe', $t_i \equiv 0$, defined formally by the initial singularity[6] $a_i \equiv a(t_i = 0) \equiv 0$. We then get

$$\Delta r_{\max}(t) = \int_0^t \frac{\mathrm{d}t'}{a(t')} = \tau(t) - \tau(0) \ . \tag{1.2.8}$$

We call this the *comoving (particle) horizon*.

The integral defining conformal time may be re-written in the following illuminating way

$$\tau \equiv \int \frac{\mathrm{d}t}{a(t)} = \int (aH)^{-1}\, \mathrm{d}\ln a \ . \tag{1.2.9}$$

This shows that the elapsed conformal time depends on the evolution of the *comoving Hubble radius* $(aH)^{-1}$. For example, for a universe dominated by a fluid with equation of state $w \equiv p/\rho$, we find that this evolves as

$$(aH)^{-1} \propto a^{\frac{1}{2}(1+3w)} \ . \tag{1.2.10}$$

Note the dependence of the exponent on the combination $(1 + 3w)$. All familiar matter sources satisfy the strong energy condition (SEC), $1+3w > 0$, so it was reasonable for post-Hubble physicists to assume that the comoving Hubble radius increases as the universe expands. Performing the integral in (1.2.9) gives

$$\tau \propto \tfrac{2}{(1+3w)}\, a^{\frac{1}{2}(1+3w)} \ , \tag{1.2.11}$$

up to an irrelevant integration constant. For conventional matter sources the initial singularity is therefore at $\tau_i = 0$,[7]

$$\tau_i \propto a_i^{\frac{1}{2}(1+3w)} = 0 \ , \qquad \text{for} \quad w > -\tfrac{1}{3} \ , \tag{1.2.12}$$

and the comoving horizon (1.2.8) is finite,

$$\Delta r_{\max}(t) \propto a(t)^{\frac{1}{2}(1+3w)} \ , \qquad \text{for} \quad w > -\tfrac{1}{3} \ . \tag{1.2.13}$$

[6] Of course, the concept of a classical spacetime (and hence the FRW metric) has broken down by that time. We will get back to that below.

[7] Of course, the actual value of τ_i is a matter of definition. The invariant statement is that for conventional matter sources the integral in (1.2.8) is dominated by the upper limit and receives vanishing contributions from early times.

1.2.3 Shock in the CMB

A moment's thought will convince the reader that the finiteness of the conformal time elapsed between $t_i = 0$ and the time of CMB decoupling $t_{\rm rec}$ implies a serious problem: most spots in the CMB have non-overlapping past light-cones and hence never were in causal contact (see fig. 1.1). Why aren't there order-one fluctuations in the CMB temperature?

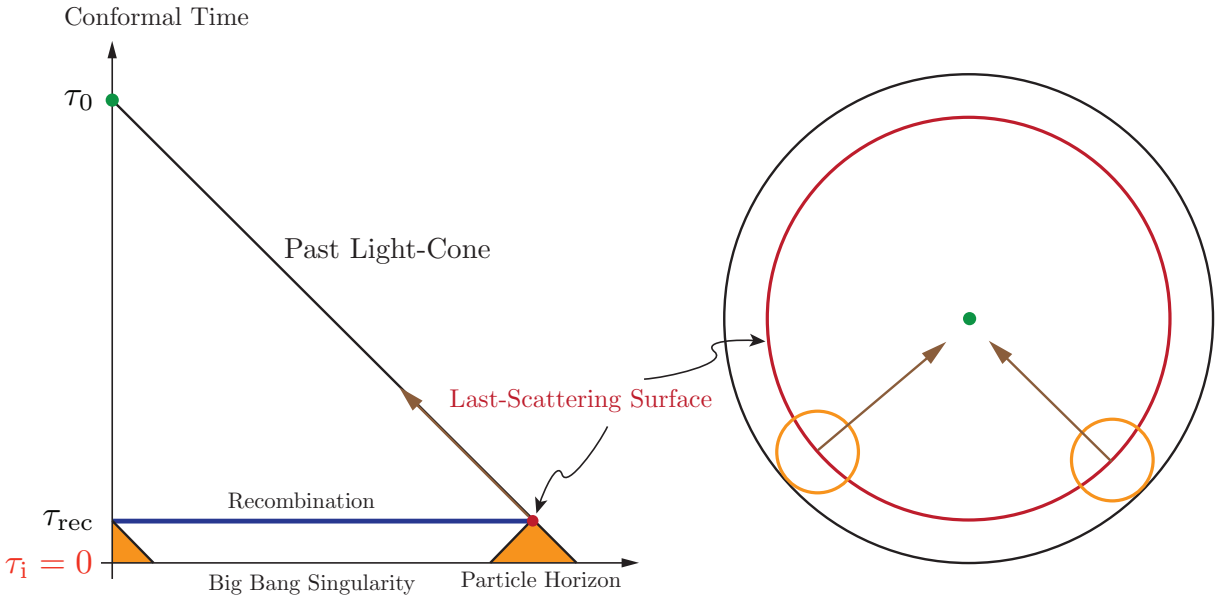


Figure 1.1: *Conformal diagram for the standard FRW cosmology.*

CMB correlations. Let us compute the angle subtended by the comoving horizon at recombination. This is defined as the ratio of the comoving particle horizon at recombination and the comoving angular diameter distance from us (an observer at redshift $z = 0$) to recombination ($z \simeq 1090$) (cf. fig. 1.1)

$$\theta_{\rm hor} = \frac{d_{\rm hor}}{d_A} \,. \tag{1.2.14}$$

A fundamental quantity is the comoving distance between redshifts z_1 and z_2

$$\tau_2 - \tau_1 = \int_{z_1}^{z_2} \frac{{\rm d}z}{H(z)} \equiv \mathcal{I}(z_1, z_2) \,. \tag{1.2.15}$$

The comoving particle horizon at recombination is

$$d_{\rm hor} = \tau_{\rm rec} - \tau_i \approx \mathcal{I}(z_{\rm rec}, \infty) \,. \tag{1.2.16}$$

In a flat universe, the comoving angular diameter distance from us to recombination is

$$d_A = \tau_0 - \tau_{\rm rec} = \mathcal{I}(0, z_{\rm rec}) \,. \tag{1.2.17}$$

The angular scale of the horizon at recombination therefore is

$$\theta_{\rm hor} \equiv \frac{d_{\rm hor}}{d_A} = \frac{\mathcal{I}(z_{\rm rec}, \infty)}{\mathcal{I}(0, z_{\rm rec})} \,. \tag{1.2.18}$$

Using

$$H(z) = H_0 \sqrt{\Omega_m (1+z)^3 + \Omega_\gamma (1+z)^4 + \Omega_\Lambda} \,, \tag{1.2.19}$$

where $\Omega_m = 0.3$, $\Omega_\Lambda = 1 - \Omega_m$, $\Omega_\gamma = \Omega_m/(1 + z_{\rm eq})$ and $z_{\rm eq} = 3400$, we can numerically evaluate the integrals $\mathcal{I}(0, z_{\rm rec})$ and $\mathcal{I}(z_{\rm rec}, \infty)$, to find

$$\theta_{\rm hor} = 1.16^\circ \,. \tag{1.2.20}$$

Causal theories should have vanishing correlation functions for

$$\theta > \theta_{\rm c} \equiv 2\theta_{\rm hor} = 2.3^\circ \,. \tag{1.2.21}$$

Inflation explains why we observe correlations in the CMB even for $\theta \gtrsim 2^\circ$.

1.2.4 Quantum Gravity Hocus-Pocus?

Let me digress briefly to make an important qualifier: when we inferred that the total conformal time between the singularity and decoupling is finite and small, we included times in the integral in (1.2.9) that were arbitrarily close to the initial singularity:

$$\Delta\tau = \underbrace{\int_0^{\delta t} \frac{{\rm d}t'}{a(t')}}_{\rm QG?} + \underbrace{\int_{\delta t}^{t} \frac{{\rm d}t'}{a(t')}}_{\rm inflation?} \quad . \tag{1.2.22}$$

However, in the first integral in (1.2.22) we have *no* reason to trust the classical geometry (1.2.1). By stating the horizon problem as we did, we were hence implicitly assuming that the breakdown of General Relativity in the regime close to the singularity does *not* lead to large contributions to the conformal time: $\delta\tau \ll \Delta\tau$. This assumption may be incorrect and there may, in fact, be no horizon problem in a complete theory of quantum gravity.[8] In the absence of an alternative solution to the horizon problem this is a completely reasonable attitude to take. However, I will now show that inflation provides as very simple and computable solution to the horizon problem. Effectively, this is achieved by modifying the scale factor evolution in the second integral in (1.2.22), i.e. in the *classical* regime. I then leave it to the reader to decide if this solution or a version of 'quantum gravity hocus-pocus' is preferable.

1.3 The Shrinking Hubble Sphere

Our presentation of the horizon problem has highlighted the fundamental role played by the growing Hubble sphere of the standard Big Bang cosmology. A simple solution to the horizon problem therefore suggests itself: conjecture a phase of *decreasing Hubble radius* in the early history of the universe,

$$\frac{d}{dt}(aH)^{-1} \;<\; 0 \,. \tag{1.3.23}$$

If this lasts long enough, the horizon problem may be avoided.

1.3.1 Solution of the Horizon Problem

As noted earlier, a decreasing Hubble radius requires a violation of the SEC, $1 + 3w < 0$, cf. (1.2.10). We then notice that the Big Bang singularity is now pushed to *negative conformal time*,[9]

$$\tau_i \propto \tfrac{2}{(1+3w)}\, a_i^{\frac{1}{2}(1+3w)} = -\infty \,, \qquad \text{for} \quad w < -\tfrac{1}{3} \,. \tag{1.3.24}$$

[8] I thank Erik Verlinde for an interesting debate on this important issue.

[9] The integral in (1.2.8) is now dominated by the lower limit.

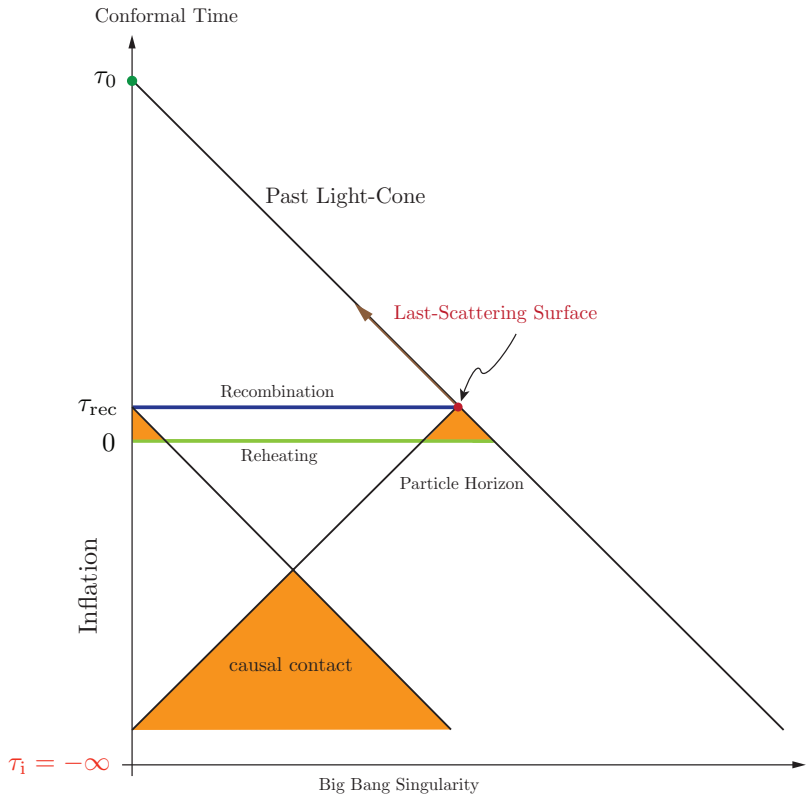


Figure 1.2: *Conformal diagram for inflationary cosmology.*

This implies that there was much more conformal time between the singularity and decoupling than we had thought! Fig. 1.2 shows the new conformal diagram. The past light cones of widely separated points in the CMB now had time to intersect before the time $\tau = 0$. In inflationary cosmology, $\tau = 0$ isn't the initial singularity, but instead becomes the time of reheating. There is time both before and after $\tau = 0$.

A decreasing comoving horizon means that large scales entering the present universe were inside the horizon before inflation (see fig. 1.3). Causal physics before inflation therefore had time to establish spatial homogeneity. With a period of inflation, the uniformity of the CMB is not a mystery anymore.

1.3.2 Solution of the Flatness Problem*

In foonote 3, I advertised that any solution to the horizon problem also solves the other Big Bang puzzles. Let me therefore demonstrate that a shrinking Hubble sphere indeed solves the flatness problem.

Consider adding spatial curvature to the Friedmann equation (1.2.2),

$$H^2 = \frac{\rho}{3M_{\rm pl}^2} - \frac{k}{a^2} \ . \tag{1.3.25}$$

Dividing both sides by the Hubble parameter, we can write this as

$$1 - \Omega(a) = \frac{-k}{(aH)^2} \ , \tag{1.3.26}$$

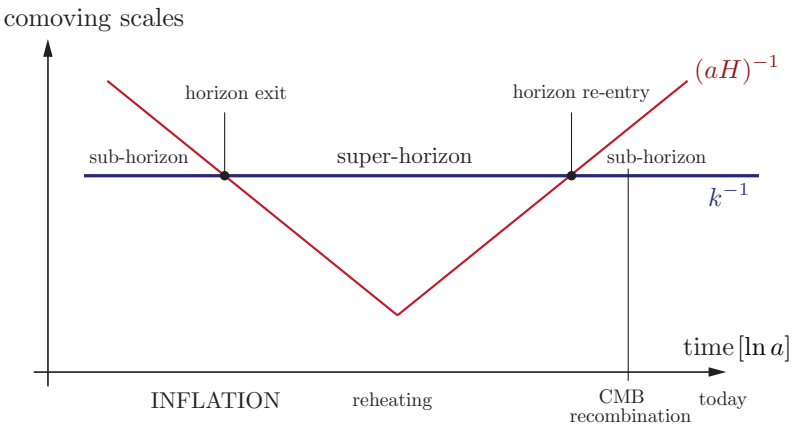


Figure 1.3: *Solution of the horizon problem. Scales of cosmological interest were larger than the Hubble radius until $a \sim 10^{-5}$ (where today is at $a(t_0) \equiv 1$). However, at very early times, before inflation operated, all scales of interest were smaller than the Hubble radius and therefore susceptible to microphysical processing. Similarly, at very late times, the scales of cosmological interest are back within the Hubble radius.*

where

$$\Omega(a) \equiv \frac{\rho(a)}{\rho_{\rm crit}(a)} \,, \qquad \rho_{\rm crit} \equiv 3M_{\rm pl}^2 H^2 \,. \tag{1.3.27}$$

The deviation of the normalized density parameter Ω from unity is a measure of the curvature of the universe. From observations we know that today $|1-\Omega(a_0)| \lesssim 0.01$. However, if there was some amount of spatial curvature in the early universe, we have to worry that it will grow with time. Conversely, in order to explain the flatness of the universe today, we have to explain a much more extreme flatness at early times, e.g. $|1-\Omega(a_{\rm GUT})| \lesssim 10^{-55}$. From eq. (1.3.26) we see that the time evolution of the curvature parameter $|1-\Omega(a)|$ again relates to the time evolution of the comoving Hubble radius $(aH)^{-1}$. Whenever $(aH)^{-1}$ is an increasing function of time, curvature grows. In contrast, during inflation, when $(aH)^{-1}$ decreases, the universe is driven towards flatness. This solves the flatness problem. The solution $\Omega = 1$ is an attractor during inflation.

Exercise. Show that

$$\frac{d\Omega}{d\ln a} = (1+3w)\Omega(\Omega-1)\,. \tag{1.3.28}$$

This makes it apparent that $\Omega = 1$ is an unstable fixed point if the strong energy condition is satisfied, but becomes an attractor during inflation.

1.3.3 Conditions for Inflation

Decreasing comoving horizon. I like the shrinking Hubble sphere as the fundamental definition of inflation since it most directly relates to the horizon problem and is key for the inflationary mechanism of generating fluctuations.

However, before we move to a description of the physics that can lead to a shrinking Hubble sphere, we show that this definition of inflation is equivalent to other popular ways of describing

inflation:

$$\frac{d}{dt}(aH)^{-1} < 0 \quad \Rightarrow \quad \varepsilon \equiv -\frac{\dot{H}}{H^2} < 1 \quad \Leftrightarrow \quad \frac{d^2a}{dt^2} > 0 \quad \Leftrightarrow \quad \rho + 3p < 0 \ .$$

Accelerated expansion. From the relation

$$\frac{d}{dt}(aH)^{-1} = \frac{d}{dt}(\dot{a})^{-1} = -\frac{\ddot{a}}{(\dot{a})^2} \ , \tag{1.3.29}$$

we see that a shrinking comoving Hubble radius implies accelerated expansion

$$\frac{d^2a}{dt^2} > 0 \ . \tag{1.3.30}$$

This explains why inflation is often defined as a period of accelerated expansion.

Slowly-varying Hubble parameter. Alternatively, we may write

$$\frac{d}{dt}(aH)^{-1} = -\frac{\dot{a}H + a\dot{H}}{(aH)^2} = -\frac{1}{a}(1-\varepsilon) \ , \qquad \text{where} \quad \varepsilon \equiv -\frac{\dot{H}}{H^2} > 0 \ . \tag{1.3.31}$$

The shrinking Hubble sphere therefore also corresponds to

$$\varepsilon = -\frac{\dot{H}}{H^2} = -\frac{d\ln H}{dN} < 1 \ . \tag{1.3.32}$$

Here, we have defined $\mathrm{d}N \equiv \mathrm{d}\ln a = H\mathrm{d}t$, which measures the number of e-folds N of inflationary expansion. Eq. (1.4.36) implies that the fractional change of the Hubble parameter per e-fold is small. Moreover, to solve the cosmological problems we want inflation to last for a sufficiently long time (usually at least $N \sim 40$ to 60 e-folds). To achieve this requires ε to remain small for a sufficiently large number of Hubble times. This condition is measured by a second parameter

$$\eta \equiv \frac{\dot{\varepsilon}}{H\varepsilon} = \frac{d\ln\varepsilon}{dN} \ . \tag{1.3.33}$$

For $|\eta| < 1$ the fractional change of ε per Hubble time is small and inflation persists.

Negative pressure. What forms of stress-energy source accelerated expansion? Assuming a perfect fluid with pressure p and density ρ, we consult the Friedmann equations in (1.2.2),

$$\dot{H} + H^2 = -\frac{1}{6M_{\rm pl}^2}(\rho + 3p) = -\frac{H^2}{2}\left(1 + \frac{3p}{\rho}\right) \ , \tag{1.3.34}$$

to find that

$$\varepsilon = -\frac{\dot{H}}{H^2} = \frac{3}{2}\left(1 + \frac{p}{\rho}\right) < 1 \quad \Leftrightarrow \quad w \equiv \frac{p}{\rho} < -\tfrac{1}{3} \ , \tag{1.3.35}$$

i.e. inflation requires negative pressure or a violation of the strong energy condition. How this can arise in a physical theory will be explained in the next section. We will see that there is nothing sacred about the strong energy condition and that it can easily be violated.

1.4 The Physics of Inflation

We have shown that a given FRW background with time-dependent Hubble parameter $H(t)$ corresponds to cosmic acceleration if and only if

$$\varepsilon \equiv -\frac{\dot{H}}{H^2} < 1 \ . \tag{1.4.36}$$

For this condition to be sustained for a sufficiently long time, requires

$$|\eta| \equiv \frac{|\dot{\varepsilon}|}{H\varepsilon} \ll 1 \ , \tag{1.4.37}$$

i.e. the fractional change of ε per Hubble time is small. In this section, we discuss what microscopic physics can lead to these conditions.

1.4.1 False Vacuum Inflation

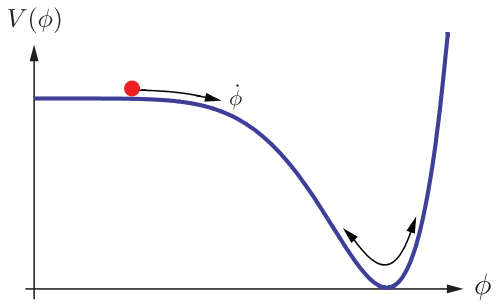
The first version of inflation considered a universe dominated by the constant energy density of a metastable false vacuum. This leads to an exponentially expanding de Sitter space with $H = const.$, and hence $\varepsilon = \eta = 0$. However, classically, false vacuum inflation never ends. Quantum-mechanically, tunnelling from the false vacuum to the true vacuum ends inflation locally, but the post-inflationary universe looks nothing like our universe. The universe is either empty or much too inhomogeneous. This is the graceful exit problem of old inflation. Any successful inflationary mechanism has to include a way of ending inflation and successfully reheating the universe. We will have to work a bit harder.

1.4.2 Slow-Roll Inflation

Consider a scalar field ϕ, the *inflaton*, minimally coupled to Einstein gravity[10]

$$S = \int \mathrm{d}^4x\sqrt{-g}\left[\frac{M_{\rm pl}^2}{2}\mathcal{R} - \frac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - V(\phi)\right] \ , \tag{1.4.38}$$

where $\mathcal{R}$ is the four-dimensional Ricci scalar derived from the metric $g_{\mu\nu}$ and $V(\phi)$ is so far an arbitrary function:



[10] In principle, we could imagine a non-minimal coupling between the inflaton and the graviton, however, in practice, non-minimally coupled theories can be transformed to minimally coupled form by a field redefinition. Similarly, we could entertain the possibility that the Einstein-Hilbert part of the action is modified at high energies. However, the simplest examples for this UV-modification of gravity, so-called $f(\mathcal{R})$ theories, can again be transformed into a minimally coupled scalar field with potential $V(\phi)$.

The time evolution of the homogeneous mode of the inflaton $\phi(t)$ is governed by the Klein-Gordon equation

$$\ddot{\phi} + 3H\dot{\phi} = -V' \ , \tag{1.4.39}$$

where the size of the Hubble friction is determined by the Friedmann equation

$$H^2 = \frac{1}{3M_{\rm pl}^2}\left[\frac{1}{2}\dot{\phi}^2 + V\right] \ . \tag{1.4.40}$$

From (1.4.39) and (1.4.40) we derive the continuity equation

$$\dot{H} = -\frac{1}{2}\frac{\dot{\phi}^2}{M_{\rm pl}^2} \ . \tag{1.4.41}$$

Substituting this into the definition of ε, we find

$$\varepsilon = \frac{\frac{1}{2}\dot{\phi}^2}{M_{\rm pl}^2 H^2} \ . \tag{1.4.42}$$

Inflation therefore occurs if the potential energy, V, dominates over the kinetic energy, $\frac{1}{2}\dot{\phi}^2$. For this condition to persist the acceleration of the scalar field has to be small. To assess this, it is useful to define the dimensionless acceleration per Hubble time

$$\delta \equiv -\frac{\ddot{\phi}}{H\dot{\phi}} \ . \tag{1.4.43}$$

Taking the time-derivative of (1.4.42) we find

$$\eta = 2(\varepsilon - \delta) \ . \tag{1.4.44}$$

Hence, if $\{\varepsilon, |\delta|\} \ll 1$ then both H and ε have small fractional changes per e-fold: $\{\varepsilon, |\eta|\} \ll 1$.

So far, no approximations have been made. We simply noted that in a regime where $\{\varepsilon, |\delta|\} \ll 1$, inflation persists. We now use these conditions to simplify the equations of motion. This is called the *slow-roll approximation*. The condition $\varepsilon = \frac{1}{2}\frac{\dot{\phi}^2}{M_{\rm pl}^2 H^2} \ll 1$ implies $\frac{1}{2}\dot{\phi}^2 \ll V$ and hence leads to the following simplification of the Friedmann equation

$$H^2 \approx \frac{V}{3M_{\rm pl}^2} \ . \tag{1.4.45}$$

The condition $|\delta| = \frac{|\ddot{\phi}|}{H|\dot{\phi}|} \ll 1$ simplifies the Klein-Gordon equation to

$$3H\dot{\phi} \approx -V' \ . \tag{1.4.46}$$

Substituting (1.4.45) and (1.4.46) into (1.4.42) gives

$$\varepsilon = -\frac{\dot{H}}{H^2} = \frac{\frac{1}{2}\dot{\phi}^2}{M_{\rm pl}^2 H^2} \approx \frac{M_{\rm pl}^2}{2}\left(\frac{V'}{V}\right)^2 \equiv \epsilon_{\rm v} \ . \tag{1.4.47}$$

Furthermore, taking the time-derivative of (1.4.46),

$$3\dot{H}\dot{\phi} + 3H\ddot{\phi} = -V''\dot{\phi} \ , \tag{1.4.48}$$

leads to

$$\delta + \varepsilon = -\frac{\ddot{\phi}}{H\dot{\phi}} - \frac{\dot{H}}{H^2} \approx M_{\rm pl}^2 \frac{V''}{V} \equiv \eta_{\rm v} \ . \tag{1.4.49}$$

Hence, a convenient way to assess a potential $V(\phi)$ is to compute the *potential slow-roll parameters*[11] $\epsilon_{\rm v}$ and $\eta_{\rm v}$. When these are small, *slow-roll inflation* occurs:

$$\epsilon_{\rm v} \equiv \frac{M_{\rm pl}^2}{2}\left(\frac{V'}{V}\right)^2 \ll 1 \qquad \text{and} \qquad |\eta_{\rm v}| \equiv M_{\rm pl}^2 \frac{|V''|}{V} \ll 1 \ . \tag{1.4.50}$$

The amount of inflation is measured by the number of e-folds of accelerated expansion

$$N \equiv \int_{a_i}^{a_f} {\rm d}\ln a = \int_{t_i}^{t_f} H(t)\,{\rm d}t \ , \tag{1.4.51}$$

where t_i and t_f are defined as the times when $\varepsilon(t_i) = \varepsilon(t_f) \equiv 1$. In the slow-roll regime we can use

$$H{\rm d}t = \frac{H}{\dot{\phi}}\,{\rm d}\phi \approx -\frac{3H}{V'}\cdot H{\rm d}\phi \approx \frac{1}{\sqrt{2\epsilon_{\rm v}}}\frac{{\rm d}\phi}{M_{\rm pl}} \tag{1.4.52}$$

to write (1.4.51) as an integral in the field space of the inflaton

$$N = \int_{\phi_i}^{\phi_f} \frac{1}{\sqrt{2\epsilon_{\rm v}}}\frac{{\rm d}\phi}{M_{\rm pl}} \ , \tag{1.4.53}$$

where ϕ_i and ϕ_f are defined as the boundaries of the interval where $\epsilon_{\rm v} < 1$. The largest scales observed in the CMB are produced some 40 to 60 e-folds before the end of inflation

$$N_{\rm cmb} = \int_{\phi_{\rm cmb}}^{\phi_f} \frac{1}{\sqrt{2\epsilon_{\rm v}}}\frac{{\rm d}\phi}{M_{\rm pl}} \approx 40 - 60 \ . \tag{1.4.54}$$

A successful solution to the horizon problem requires at least $N_{\rm cmb}$ e-folds of inflation.

Case study: $m^2\phi^2$ inflation. As an example, let us give the slow-roll analysis of arguably the simplest model of inflation: single field inflation driven by a mass term

$$V(\phi) = \frac{1}{2}m^2\phi^2 \, . \tag{1.4.55}$$

The slow-roll parameters are

$$\epsilon_{\rm v}(\phi) = \eta_{\rm v}(\phi) = 2\left(\frac{M_{\rm pl}}{\phi}\right)^2 \, . \tag{1.4.56}$$

To satisfy the slow-roll conditions $\epsilon_{\rm v}, |\eta_{\rm v}| < 1$, we therefore need to consider super-Planckian values for the inflaton

$$\phi > \sqrt{2}M_{\rm pl} \equiv \phi_f \, . \tag{1.4.57}$$

The relation between the inflaton field value and the number of e-folds before the end of inflation is

$$N(\phi) = \frac{\phi^2}{4M_{\rm pl}^2} - \frac{1}{2} \, . \tag{1.4.58}$$

Fluctuations observed in the CMB are created at

$$\phi_{\rm cmb} = 2\sqrt{N_{\rm cmb}}\, M_{\rm pl} \sim 15 M_{\rm pl} \, . \tag{1.4.59}$$

[11] In contrast, the parameters ε and η are often called the *Hubble slow-roll parameters*. During slow-roll the parameters are related as follows: $\epsilon_{\rm v} \approx \varepsilon$ and $\eta_{\rm v} \approx 2\epsilon - \frac{1}{2}\eta$.

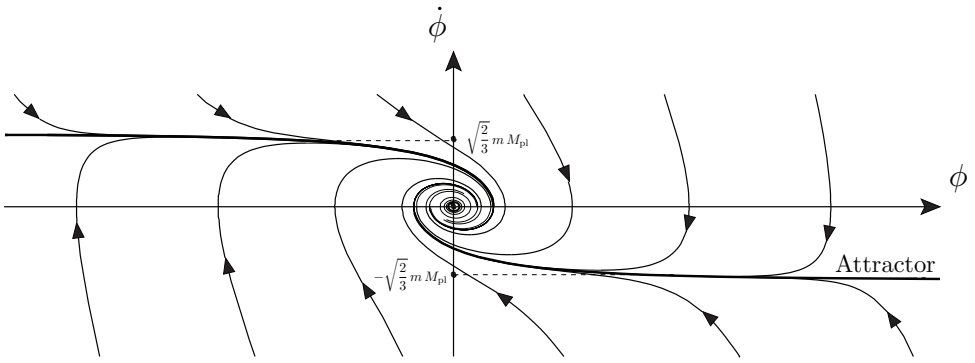


Figure 1.4: *Phase space diagram of $m^2\phi^2$ inflation.*

Finally, let us comment that slow-roll inflation for the $m^2\phi^2$ potential is an *attractor* solution. To see this you should study the phase space diagram using[12]

$$\frac{d\dot\phi}{d\phi} = -\frac{\sqrt{\frac{3}{2}}\frac{1}{M_{\rm pl}^2}(\dot\phi^2 + m^2\phi^2)^{1/2}\dot\phi + m^2\phi}{\dot\phi} \ . \tag{1.4.60}$$

The result is portrayed in fig. 1.4.[13]

1.4.3 Hybrid Inflation

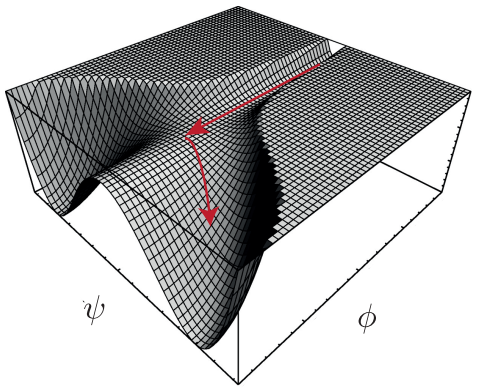
In single-field slow-roll inflation, a single field governs both the inflationary dynamics ($\varepsilon \ll 1$) and the exit from inflation ($\varepsilon \to 1$).[14] In contrast, in hybrid inflation the shape of the inflaton potential is decoupled from the exit from inflation. This is achieved by coupling the inflaton field ϕ to a second field ψ. Consider, for example, the following Lagrangian

$$\mathcal{L} = -\frac{1}{2}(\partial_\mu\phi)^2 - \frac{1}{2}(\partial_\mu\psi)^2 - V(\phi) - \frac{1}{4\lambda}(M^2 - \lambda\psi^2)^2 - \frac{g^2}{2}\phi^2\psi^2 \ , \tag{1.4.61}$$

where $V(\phi) \ll \frac{M^4}{4\lambda}$, so that the dominant contribution to the inflationary energy density is coming from the false vacuum energy of the symmetry breaking potential $V(\psi) \equiv \frac{1}{4\lambda}(M^2 - \lambda\psi^2)^2$. The coupling between ϕ and ψ induces an effective mass of the second field that depends on the value of the inflaton

$$m_\psi^2(\phi) = -M^2 + g^2\phi^2 \ . \tag{1.4.62}$$

For large values of the inflaton field $\phi > \phi_c \equiv M/g$ the field ψ is stabilized at its only minimum at $\psi = 0$. During that phase, ψ is very massive and can be integrated out, so that the theory reduces to that of single-field slow-roll inflation. However, for $\phi < \phi_c$, the waterfall field ψ becomes tachyonic and ends inflation.



[12] To arrive at eq. (4.3.64) we substituted $\ddot\phi = \dot\phi\frac{d\dot\phi}{d\phi}$ into the Klein-Gordon equation.

[13] Figure reproduced from V. Mukhanov, *Physical Foundations of Cosmology*.

[14] This then often requires that the field moves over a super-Planckian field range during the 60 e-folds of inflation.

1.4.4 K-Inflation

Slow-roll inflation assumes that the kinetic term is canonical, i.e. $\mathcal{L}_{\rm s.r.} = X - V(\phi)$, where $X \equiv -\frac{1}{2}(\partial_\mu\phi)^2$. As we have seen, this puts strong constraints on the shape of the potential $V(\phi)$ via the potential slow-roll conditions, $\{\epsilon_{\rm v}, |\eta_{\rm v}|\} \ll 1$. However, these conditions for inflation are not absolute, but assume the slow-roll approximations. In contrast, the Hubble slow-roll conditions, $\{\varepsilon, |\eta|\} \ll 1$ don't make any approximations, and allow for a larger spectrum of inflationary backgrounds. In particular, the constraints on the inflationary potential can potentially be relaxed if higher-derivative corrections to the kinetic term were dynamically relevant during inflation, i.e. $|\dot{H}| \ll H^2$ not because the theory of potential-dominated, but because it allows non-trivial dynamics.

A useful way to describe these effects is by the following action,

$$S = \int \mathrm{d}^4x\sqrt{-g}\left[\frac{M_{\rm pl}^2}{2}\mathcal{R} + P(X,\phi)\right] , \tag{1.4.63}$$

where

$$P(X,\phi) = \sum_n c_n(\phi)\frac{X^n}{\Lambda^{4n-4}} . \tag{1.4.64}$$

For $X \ll \Lambda^4$, the dynamics reduces to that of slow-roll inflation, so we are now interested in the limit $X \sim \Lambda^4$. Naively, this looks like playing with fire, since X/Λ^4 controls the derivative expansion in (1.4.64). In particular, in the limit $X \to \Lambda^4$ we have to worry about the appearance of unstable 'ghost' states and the stability under radiative corrections. Specifically, in the absence of symmetries, there is no way to protect the coefficients c_n in (1.4.64) from quantum corrections. The predictions derived from (1.4.64) then can't be trusted. However, sometimes the theory is equipped with a symmetry that forbids large renormalizations of these coefficients. This is the case, for instance in Dirac-Born-Infeld (DBI) inflation, where a higher-dimensional boost symmetry protects the special form of the Lagrangian

$$P(X,\phi) = -\Lambda^4(\phi)\sqrt{1 + \frac{X}{\Lambda^4(\phi)}} - V(\phi) . \tag{1.4.65}$$

In this case, the boost symmetry forces quantum corrections to involve the two-derivative combination $\nabla\nabla\phi$. We stress that only when they come with protective symmetries are $P(X)$–theories really interesting and predictive theories. The stress-energy tensor arising from (1.4.63) has pressure P and energy density

$$\rho = 2XP_{,X} - P , \tag{1.4.66}$$

where $P_{,X}$ denotes a derivative with respect to X. The inflationary parameter (1.4.36) becomes

$$\varepsilon = -\frac{\dot{H}}{H^2} = \frac{3XP_{,X}}{2XP_{,X} - P} . \tag{1.4.67}$$

The condition for inflation is still $\varepsilon \ll 1$, which now is a condition on the functional form of $P(X)$. However, it should be remembered that unless a protective symmetry is identified, there is no guarantee that the $P(X)$–theory is radiatively stable.

The fluctuations in $P(X)$–theories have a number of interesting features. First, in the limit $X \sim \Lambda^4$, they propagate with a non-trival speed of sound

$$c_s^2 = \frac{dP}{d\rho} = \frac{P_{,X}}{P_{,X} + 2XP_{,XX}} \ll 1 . \tag{1.4.68}$$

The limit of small sound speed implies enhanced interactions in the cubic and quartic Lagrangian. This leads to large amount of non-Gaussianity. In Chapter 5, we will explain how those non-Gaussianities are calculated.

1.5 Outlook

In this chapter, we discussed the classical dynamics of inflation ($\hbar = 0$) and explained how it provides a simple solution to the horizon problem. Inflation therefore explains the large-scale homogeneity, isotropy and flatness of the universe. In the next chapter, I will present the quantum limit of inflation ($\hbar \neq 0$) and show that it provides a beautiful mechanism to explain the observed CMB fluctuations. The evolution of the Hubble sphere in fig. 1.3 will play a fundamental role in this story. It allows quantum zero-point fluctuations of the inflaton field to lead to primordial density fluctuations of precisely the right type to account for the observed CMB anisotropies.

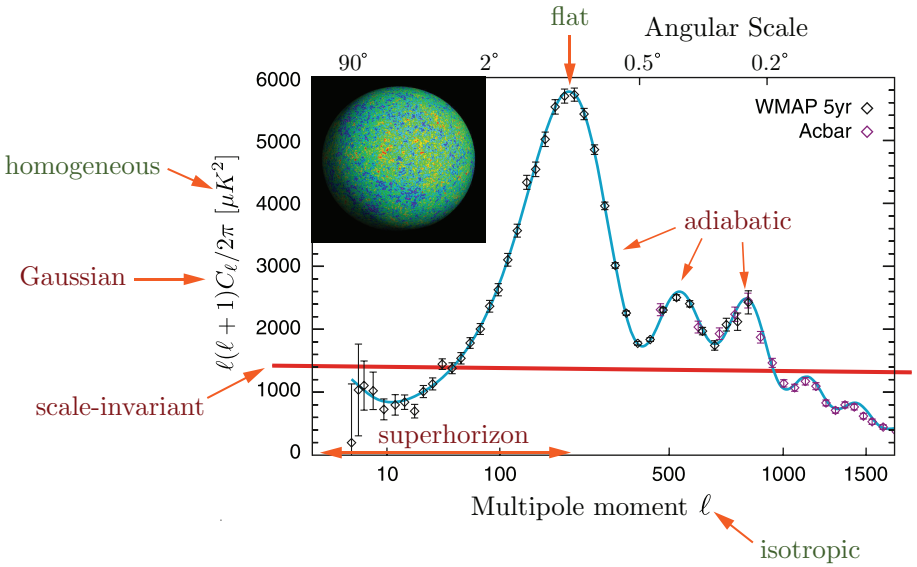


Figure 1.5: *The observational evidence for inflation.*

A compact representation of CMB data is in terms of the angular power spectrum (see fig. 1.5)

$$C_\ell \equiv \frac{1}{2\ell+1}\sum_m |a_{\ell m}|^2 \ , \qquad \text{where} \quad \frac{\Delta T(\theta,\phi)}{\bar{T}} = \sum_{\ell,m} a_{\ell m} Y_{\ell m}(\theta,\phi) \ . \tag{1.5.69}$$

All the predictions of inflation are directly (or indirectly) encoded in the CMB power spectrum: On large scales the universe is

1a) *homogenous*: the temperature fluctuations are small: $\Delta T \lesssim 10\mu K$,

1b) *isotropic*: little information is lost by the sum over $a_{\ell m}$'s in (1.5.69),

1c) *flat*: the first peak of the power spectrum is at $\ell \sim 200$.

Its small-scale fluctuations are

2a) *superhorizon*: the power doesn't vanish for $\theta > 2°$,

2b) *scale-invariant*: the primordial power is nearly independent of scale,

2c) *Gaussian*: little information is lost by reducing the data to the power spectrum,

2d) *adiabatic*: the presence of the acoustic peaks constrains isocurvature fluctuations.

2 Quantum Fluctuations during Inflation

2.1 Motivation

In this chapter and the next, we discuss the primordial origin of the temperature variations in the CMB. The main goal will be to show how quantum fluctuations in quasi-de Sitter space produce a spectrum of fluctuations that accurately matches the observations.

The reason why inflation inevitably produces fluctuations is simple: as we have seen in the previous chapter, the inflaton evolution $\phi(t)$ governs the energy density of the early universe $\rho(t)$ and hence controls the end of inflation. Essentially, ϕ plays the role of a local clock reading off the amount of inflationary expansion remaining. Because microscopic clocks are quantum-mechanical objects with necessarily some variance (by the uncertainty principle), the inflaton will have spatially varying fluctuations $\delta\phi(t, \mathbf{x}) \equiv \phi(t, \mathbf{x}) - \bar{\phi}(t)$. These fluctuations imply that different regions of space inflate by different amounts. In other words, there will be local differences in the time when inflation end $\delta t(\mathbf{x})$. Moreover, these differences in the local expansion histories lead to differences in the local densities after inflation. In quantum theory, local fluctuations in $\delta\rho(t, \mathbf{x})$ and hence ultimately in the CMB temperature $\Delta T(\mathbf{x})$ are therefore unavoidable. The main purpose of this chapter is to compute this effect. It is worth remarking that the theory wasn't engineered to produce the CMB fluctuations, but their origin is instead a natural consequence of treating inflation quantum mechanically.

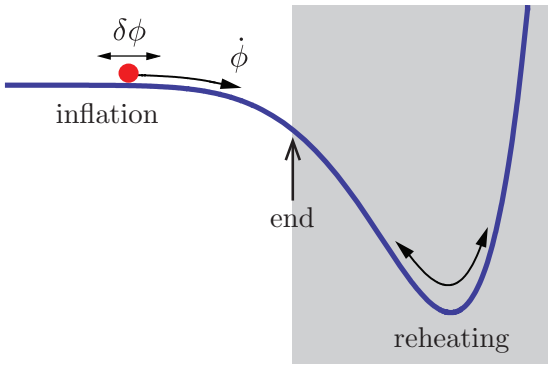


Figure 2.1: *Quantum fluctuations $\delta\phi(t, \mathbf{x})$ around the classical background evolution $\bar{\phi}(t)$. Regions acquiring a negative frozen fluctuations $\delta\phi$ remain potential-dominated longer than regions with positive $\delta\phi$. Different parts of the universe therefore undergo slightly different evolutions. After inflation, this induces relative density fluctuations $\delta\rho(r, \mathbf{x})$.*

2.2 Classical Perturbations

For concreteness, we will consider single-field slow-roll models of inflation

$$S = \int \mathrm{d}^4x\sqrt{-g}\left[\frac{1}{2}\mathcal{R} - \frac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - V(\phi)\right] \ , \tag{2.2.1}$$

where $M_{\rm pl} \equiv 1$. We will study both scalar and tensor fluctuations. For the scalar modes we have to be careful to identify the true physical degrees of freedom. A priori, we have 5 scalar modes: 4 metric perturbations – $\delta g_{00}, \delta g_{ii}, \delta g_{0i} \sim \partial_i B$ and $\delta g_{ij} \sim \partial_i\partial_j H$ – and 1 scalar field perturbation $\delta\phi$. Gauge invariances associated with the invariance of (2.2.1) under scalar coordinate transformations – $t \to t + \epsilon_0$ and $x_i \to x_i + \partial_i\epsilon$ – remove two modes. The Einstein constraint equations remove two more modes, so that we are left with 1 physical scalar mode. Deriving the quadratic action for this mode is the aim of this section.

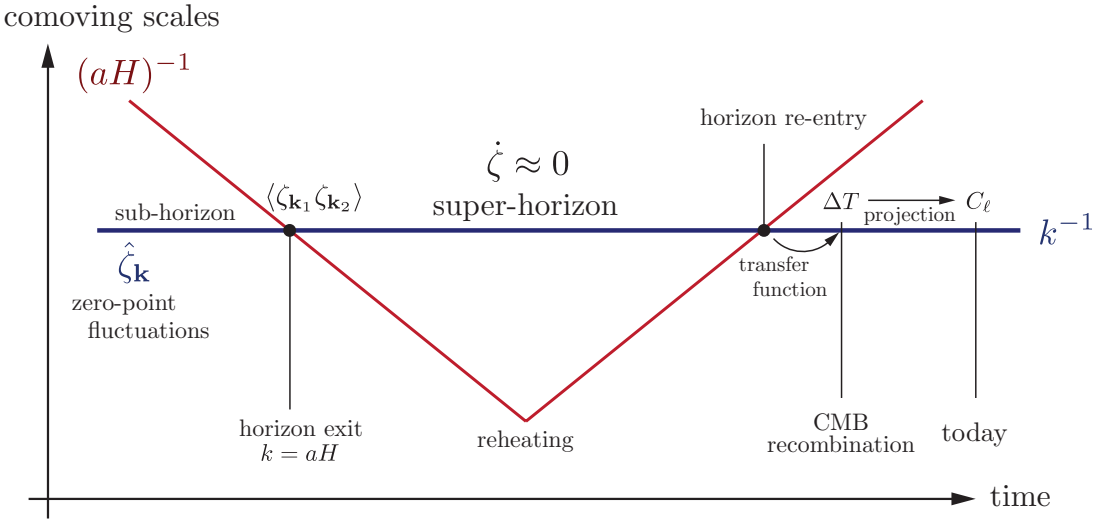


Figure 2.2: *Curvature perturbations during and after inflation: The comoving horizon $(aH)^{-1}$ shrinks during inflation and grows in the subsequent FRW evolution. This implies that comoving scales k^{-1} exit the horizon at early times and re-enter the horizon at late times. While the curvature perturbations ζ are outside of the horizon they don't evolve, so our computation for the correlation function $\langle\zeta_{\mathbf{k}_1}\zeta_{\mathbf{k}_2}\rangle$ at horizon exit during the early de Sitter phase can be related directly to CMB observables at late times.*

2.2.1 Comoving Gauge

We will work in a fixed gauge throughout. For a number of reason it will be convenient to work in *comoving gauge*, defined by the vanishing of the momentum density, $\delta T_{0i} \equiv 0$. For slow-roll inflation, this becomes[1]

$$\delta\phi \ = \ 0 \ . \tag{2.2.2}$$

In this gauge, perturbations are characterized purely by fluctuations in the metric,

$$\delta g_{ij} \ = \ a^2(1-2\zeta)\delta_{ij} + a^2 h_{ij} \ . \tag{2.2.3}$$

[1]Below we use the Einstein equations to replace the additional (non-dynamical) metric perturbations δg_{00} and δg_{0i} in terms of ζ. This results in an action purely for ζ which is why, for now, we can afford to be a bit implicit about the remaining metric perturbations.

Here, h_{ij} is a transverse ($\nabla_i h^{ij} = 0$), traceless ($h^i_i = 0$) tensor and ζ is a scalar. One can show that the comoving spatial slices $\phi = const.$ have three-curvature $R_{(3)} = \frac{4}{a^2}\nabla^2\zeta$. Hence, ζ is referred to as the (comoving) *curvature perturbation.*[2]

The perturbation ζ has the crucial property that (for adiabatic matter fluctuations) it is time-independent on superhorizon scales (see fig. 2.2):

$$\lim_{k\ll aH} \dot{\zeta}_{\mathbf{k}} = 0 \ . \tag{2.2.4}$$

We will prove this important fact in the next chapter. The constancy of ζ on superhorizon scales allows us to relate CMB observations directly to the inflationary dynamics (at the time when a given fluctuation crosses the horizon) while allowing us to be completely ignorant about the high-energy physics during the intervening history of the universe.

2.2.2 Constraint Equations

Solving the Einstein equations for the non-dynamical metric perturbations δg_{00} and δg_{0i} in terms of ζ is a bit tedious. Readers who don't want to be distracted by these technical details are advise to skip to the next subsection and proceed to the result for the quadratic action for the perturbation ζ. Otherwise, the details can be found in the following digression:

Free field action for ζ. The constraint equations are solved most conveniently in the ADM formalism, where the metric fluctuations become non-dynamical Lagrange multipliers. In ADM, spacetime is sliced into three-dimensional hypersurfaces

$$\mathrm{d}s^2 = -N^2\mathrm{d}t^2 + g_{ij}(\mathrm{d}x^i + N^i\mathrm{d}t)(\mathrm{d}x^j + N^j\mathrm{d}t) \, . \tag{2.2.5}$$

Here, g_{ij} is the three-dimensional metric on slices of constant t. The lapse function $N(\mathbf{x})$ and the shift function $N_i(\mathbf{x})$ appear as non-dynamical Lagrange multipliers in the action, i.e. their equations of motion are purely algebraic. For our purposes this is the main advantage of the ADM formalism. The action (2.2.1) becomes

$$\begin{aligned} S &= \frac{1}{2}\int \mathrm{d}^4x\sqrt{-g}\Big[NR_{(3)} - 2NV + N^{-1}(E_{ij}E^{ij} - E^2) \\ &\qquad + N^{-1}(\dot{\phi} - N^i\partial_i\phi)^2 - Ng^{ij}\partial_i\phi\partial_j\phi - 2V\Big] \, , \end{aligned} \tag{2.2.6}$$

where E_{ij} is the extrinsic curvature of the three-dimensional spatial slices

$$E_{ij} \equiv \frac{1}{2}(\dot{g}_{ij} - \nabla_i N_j - \nabla_j N_i) \ , \qquad E = E^i_i \ , \tag{2.2.7}$$

and the perturbed three-metric is

$$g_{ij} = a^2(1-2\zeta)\delta_{ij} \ . \tag{2.2.8}$$

The ADM action (2.2.6) implies the following constraint equations for the Lagrange multipliers N and N^i

$$\nabla_i[N^{-1}(E^i_j - \delta^i_j E)] = 0 \, , \tag{2.2.9}$$

$$R_{(3)} - 2V - N^{-2}(E_{ij}E^{ij} - E^2) - N^{-2}\dot{\phi}^2 = 0 \, . \tag{2.2.10}$$

To solve the constraints, we split the shift vector N_i into irrotational (scalar) and incompressible (vector) parts

$$N_i \equiv \psi_{,i} + \tilde{N}_i \, , \qquad \text{where} \qquad \partial_i\tilde{N}_i = 0 \, , \tag{2.2.11}$$

[2] Sometimes the notation $\mathcal{R}$ is used for the comoving curvature perturbation, to distinguish it from the curvature perturbation on uniform density hypersurfaces, which sometimes is also denoted by ζ. We will continue to use ζ for the comoving curvature perturbation.

and define the lapse perturbation as

$$N \equiv 1 + \alpha \,. \tag{2.2.12}$$

The quantities α, ψ and $\tilde{N}_i$ then admit expansions in powers of ζ,

$$\begin{aligned} \alpha &= \alpha_1 + \alpha_2 + \dots \,, \\ \psi &= \psi_1 + \psi_2 + \dots \,, \\ \tilde{N}_i &= \tilde{N}_i^{(1)} + \tilde{N}_i^{(2)} + \dots \,, \end{aligned} \tag{2.2.13}$$

where, e.g. $\alpha_n = \mathcal{O}(\zeta^n)$. The constraint equations may then be solved order-by-order: To get the quadratic action, we only need to solve the constraint equations to first order. Eq. (2.2.10) then implies

$$\alpha_1 = \frac{\dot{\zeta}}{H} \,, \qquad \partial^2 \tilde{N}_i^{(1)} = 0 \,, \tag{2.2.14}$$

where $\tilde{N}_i^{(1)} \equiv 0$ with an appropriate choice of boundary conditions. Furthermore, at first order eq. (2.2.9) gives

$$\psi_1 = -\frac{\zeta}{H} + \frac{a^2}{H} \epsilon_{\rm v} \, \partial^{-2} \dot{\zeta} \,, \tag{2.2.15}$$

where ∂^{-2} is defined via $\partial^{-2}(\partial^2 f) = f$.

Substituting the first-order solutions for N and N_i back into the action, one finds the following second-order action

$$S_2 = \frac{1}{2} \int \mathrm{d}^4x \, a^3 \frac{\dot{\phi}^2}{H^2} \left[\dot{\zeta}^2 - a^{-2} (\partial_i \zeta)^2 \right] \,. \tag{2.2.16}$$

To arrive at this result requires integration by parts and use of the background equations of motion.

2.2.3 Quadratic Action

Substituting $\delta g_{00}(\zeta)$ and $\delta g_{0i}(\zeta)$ into (2.2.1) and expanding in powers of ζ, we find

$$S = \frac{1}{2} \int \mathrm{d}t \, \mathrm{d}^3\mathbf{x} \; a^3 \, \frac{\dot{\phi}^2}{H^2} \left[\dot{\zeta}^2 - \frac{1}{a^2} (\partial_i \zeta)^2 \right] + \cdots \tag{2.2.17}$$

The ellipses in (2.2.17) refer to terms that are higher order in ζ. Being interested only in the quadratic action of ζ we will now drop these terms. We will come back to these terms when we discuss higher-order correlations and non-Gaussianity in Chapter 5. We define the canonically-normalized Mukhanov variable

$$v \equiv z\zeta \,, \tag{2.2.18}$$

where

$$z^2 \equiv a^2 \frac{\dot{\phi}^2}{H^2} = 2a^2 \varepsilon \,. \tag{2.2.19}$$

Switching to conformal time, we get

$$S = \frac{1}{2} \int \mathrm{d}\tau \, \mathrm{d}^3\mathbf{x} \left[(v')^2 - (\partial_i v)^2 + \frac{z''}{z} v^2 \right] \,. \tag{2.2.20}$$

We recognize this as the action of an *harmonic oscillator with a time-dependent mass*

$$m_{\rm eff}^2(\tau) \equiv -\frac{z''}{z} = -\frac{H}{a\dot{\phi}} \frac{\partial^2}{\partial \tau^2} \frac{a\dot{\phi}}{H} \,. \tag{2.2.21}$$

The time-dependence of the effective mass accounts for the interaction of the scalar field ζ with the gravitational background. Notice that both $a(t)$ and $\phi(t)$ contribute to $m_{\rm eff}(\tau)$.

2.2.4 Mukhanov-Sasaki Equation

Varying the action S, we arrive at the *Mukhanov-Sasaki equation*[3]

$$v_{\mathbf{k}}'' + \underbrace{\left(k^2 - \frac{z''}{z}\right)}_{\equiv \omega_k^2(\tau)} v_{\mathbf{k}} = 0 \ , \tag{2.2.22}$$

where we defined the Fourier modes,

$$v_{\mathbf{k}}(\tau) \equiv \int \mathrm{d}^3\mathbf{x}\, e^{-i\mathbf{k}\cdot\mathbf{x}}\, v(\tau, \mathbf{x}) \ . \tag{2.2.23}$$

In de Sitter space, $a = -(H\tau)^{-1}$, the effective frequency reduces to

$$\omega_k^2(\tau) = k^2 - \frac{2}{\tau^2} \qquad \text{(de Sitter)}. \tag{2.2.24}$$

To guide our intuition we consider special limits of (2.2.22): For modes with wavelengths much smaller than the horizon, $k^2 \gg |z''/z|$, we get

$$v_{\mathbf{k}}'' + k^2 v_{\mathbf{k}} = 0 \qquad \text{(subhorizon)}. \tag{2.2.25}$$

This leads to oscillating solutions: $v_{\mathbf{k}} \propto e^{\pm ik\tau}$. For modes with wavelengths much larger than the horizon, $k^2 \ll |z''/z|$, we find instead

$$\frac{v_{\mathbf{k}}''}{v_{\mathbf{k}}} = \frac{z''}{z} \approx \frac{2}{\tau^2} \qquad \text{(superhorizon)}. \tag{2.2.26}$$

This has the growing solution[4] $v_{\mathbf{k}} \propto z \propto \tau^{-1}$ (and the decaying solution $v_{\mathbf{k}} \propto \tau^2$). This implies that ζ indeed freezes on superhorizon scales: $\zeta_{\mathbf{k}} = z^{-1} v_{\mathbf{k}} \propto const.$

2.2.5 Mode Expansion

Since the frequency $\omega_k(\tau)$ in (2.2.22) depends only on $k \equiv |\mathbf{k}|$, the most general solution of (2.2.22) can be written as[5]

$$v_{\mathbf{k}} \equiv a_{\mathbf{k}}^- v_k(\tau) + a_{-\mathbf{k}}^+ v_k^*(\tau) \ . \tag{2.2.27}$$

Here, $v_k(\tau)$ and its complex conjugate $v_k^*(\tau)$ are two linearly independent solutions of (2.2.22). As indicated by dropping the vector notation $\mathbf{k}$ on the subscript, the mode functions, $v_k(\tau)$ and $v_k^*(\tau)$, are the same for all Fourier modes with $k \equiv |\mathbf{k}|$. The Wronskian of the mode functions is

$$W[v_k, v_k^*] \equiv v_k' v_k^* - v_k v_k^{*\prime} = 2i\, \mathrm{Im}(v_k' v_k^*) \ . \tag{2.2.28}$$

From the equation of motion (2.2.22) it follows that $W[v_k, v_k^*]$ is time-independent. Furthermore, by rescaling the mode functions as $v_k \to \lambda v_k$ (giving $W[v_k, v_k^*] \to |\lambda|^2 W[v_k, v_k^*]$) we can always normalize v_k such that

$$W[v_k, v_k^*] = v_k' v_k^* - v_k v_k^{*\prime} \equiv -i \ . \tag{2.2.29}$$

[3]The Mukhanov-Sasaki equation is hard to solve in full generality since the function $z(\tau)$ depends on the background dynamics. For a given inflationary background, $\phi(\tau)$ and $a(\tau)$, one may of course solve eq. (2.2.22) numerically. However, to gain a more intuitive understanding of the solutions, we will discuss approximate analytical solutions in the pure de Sitter limit, as well as in the slow-roll expansion of quasi-de Sitter space.

[4]Recall that τ runs from large negative values to zero during inflation.

[5]The $-\mathbf{k}$ on $a_{-\mathbf{k}}^+$ was chosen for later convenience.

The reason for this particular choice of normalization will become clear momentarily.

The two time-independent integration constants $a^{\pm}_{\mathbf{k}}$ in (2.2.27) are

$$a^{-}_{\mathbf{k}} = \frac{v^{*\prime}_k v_{\mathbf{k}} - v^*_k v'_{\mathbf{k}}}{v^{*\prime}_k v_k - v^*_k v'_k} = \frac{W[v^*_k, v_{\mathbf{k}}]}{W[v^*_k, v_k]} \qquad \text{and} \qquad a^{+}_{\mathbf{k}} = (a^{-}_{\mathbf{k}})^* \ , \tag{2.2.30}$$

where the relation between $a^{+}_{\mathbf{k}}$ and $a^{-}_{\mathbf{k}}$ follows from the reality of v. Note that the constants $a^{\pm}_{\mathbf{k}}$ may depend on the direction of the wave vector $\mathbf{k}$.

Finally, Fourier transforming (2.2.27) gives

$$v(\tau, \mathbf{x}) = \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^{3/2}} \left[a^{-}_{\mathbf{k}} v_k(\tau) + a^{+}_{-\mathbf{k}} v^*_k(\tau)\right] e^{i\mathbf{k}\cdot\mathbf{x}} \tag{2.2.31}$$

$$= \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^{3/2}} \left[a^{-}_{\mathbf{k}} v_k(\tau) e^{i\mathbf{k}\cdot\mathbf{x}} + a^{+}_{\mathbf{k}} v^*_k(\tau) e^{-i\mathbf{k}\cdot\mathbf{x}}\right] \ , \tag{2.2.32}$$

where the second line is manifestly real, since $a^{+}_{\mathbf{k}} = (a^{-}_{\mathbf{k}})^*$.

2.3 Quantum Origin of Cosmological Perturbations

Our task now is to quantize the field v. This is not much more complicated than quantizing the simple harmonic oscillator in quantum mechanics, except for a small subtlety in the vacuum choice arising from the time-dependence of the oscillator frequencies $\omega_k(\tau)$.[6]

2.3.1 Canonical Quantization

The canonical quantization procedure proceeds in the standard way: the field v and its canonically conjugate momentum $\pi \equiv v'$ are promoted to quantum operators $\hat{v}$ and $\hat{\pi}$, which satisfy the standard equal-time commutation relations[7]

$$[\hat{v}(\tau, \mathbf{x}), \hat{\pi}(\tau, \mathbf{y})] = i\delta(\mathbf{x} - \mathbf{y}) \ , \tag{2.3.33}$$

and

$$[\hat{v}(\tau, \mathbf{x}), \hat{v}(\tau, \mathbf{y})] = [\hat{\pi}(\tau, \mathbf{x}), \hat{\pi}(\tau, \mathbf{y})] = 0 \ . \tag{2.3.34}$$

It follows from (2.2.22) that the commutation relation (2.3.33) holds at all times if it holds at any one time. The Hamiltonian is

$$\hat{H}(\tau) = \frac{1}{2} \int \mathrm{d}^3\mathbf{x} \left[\hat{\pi}^2 + (\partial_i \hat{v})^2 + m^2_{\mathrm{eff}}(\tau) \hat{v}^2\right] \ . \tag{2.3.35}$$

The constants of integration $a^{\pm}_{\mathbf{k}}$ in the mode expansion of v become operators $\hat{a}^{\pm}_{\mathbf{k}}$, so that the field operator $\hat{v}$ is expanded as

$$\hat{v}(\tau, \mathbf{x}) = \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^{3/2}} \left[\hat{a}^{-}_{\mathbf{k}} v_k(\tau) e^{i\mathbf{k}\cdot\mathbf{x}} + \hat{a}^{+}_{\mathbf{k}} v^*_k(\tau) e^{-i\mathbf{k}\cdot\mathbf{x}}\right] \ . \tag{2.3.36}$$

Substituting (2.3.36) into (2.3.33) and (2.3.34) implies

$$[\hat{a}^{-}_{\mathbf{k}}, \hat{a}^{+}_{\mathbf{k}'}] = \delta(\mathbf{k} - \mathbf{k}') \qquad \text{and} \qquad [\hat{a}^{-}_{\mathbf{k}}, \hat{a}^{-}_{\mathbf{k}'}] = [\hat{a}^{+}_{\mathbf{k}}, \hat{a}^{+}_{\mathbf{k}'}] = 0 \ . \tag{2.3.37}$$

[6] For a nice treatment of quantum field theory in curved backgrounds I strongly recommend: V. Mukhanov and S. Winitzki, *Introduction to Quantum Effects in Gravity*.

[7] Here, we defined $\hbar \equiv 1$.

We realize that our normalization for the mode functions (2.2.29) was wisely chosen to make (2.3.37) simple. The operators $\hat{a}^+_{\mathbf{k}}$ and $\hat{a}^-_{\mathbf{k}}$ may then be interpreted as creation and annihilation operators, respectively. As usual, quantum states in the Hilbert space are constructed by defining the vacuum state $|0\rangle$ via

$$\hat{a}^-_{\mathbf{k}}|0\rangle = 0 \ , \tag{2.3.38}$$

and by producing excited states by repeated application of creation operators

$$|m_{\mathbf{k}_1}, n_{\mathbf{k}_2}, \cdots\rangle = \frac{1}{\sqrt{m!n!\cdots}}\left[(a^+_{\mathbf{k}_1})^m (a^+_{\mathbf{k}_2})^n \cdots\right]|0\rangle \ . \tag{2.3.39}$$

2.3.2 Non-Uniqueness of the Vacuum

An unambiguous physical interpretation of the states in (2.3.38) and (2.3.39) arises only after the mode functions $v_k(\tau)$ are selected.[8] However, the normalization (2.2.29) is not sufficient to completely fix the solutions $\chi_k(\tau)$ to the second-order ODE (2.2.22). An unambiguous definition of the vacuum still requires additional physical input.

To illustrate this ambiguity explicitly, consider the following functions

$$u_k(\tau) = \alpha_k v_k(\tau) + \beta_k v_k^*(\tau) \ , \tag{2.3.40}$$

where α_k and β_k are complex constants. The functions $u_k(\tau)$ of course also satisfy the equation of motion (2.2.22). Moreover, they satisfy the normalization (2.2.29), i.e. $W[u_k, u_k^*] = -i$, if the coefficients α_k and β_k obey

$$|\alpha_k|^2 - |\beta_k|^2 = 1 \ . \tag{2.3.41}$$

At this point there is therefore nothing that permits us to favor $v_k(\tau)$ over $u_k(\tau)$ in our choice of mode functions. In terms of $u_k(\tau)$ the expansion of $\hat{v}$ takes the form

$$\hat{v}(\tau, \mathbf{x}) = \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^{3/2}}\left[\hat{b}^-_{\mathbf{k}} u_k(\tau) e^{i\mathbf{k}\cdot\mathbf{x}} + \hat{b}^+_{\mathbf{k}} u_k^*(\tau) e^{-i\mathbf{k}\cdot\mathbf{x}}\right] \ , \tag{2.3.42}$$

where $\hat{b}^\pm_{\mathbf{k}}$ are alternative creation and annihilation operators satisfying (2.3.37). Comparing (2.3.42) to (2.3.36) leads to the *Bogolyubov transformation* between $\hat{b}^\pm_{\mathbf{k}}$ operators and $\hat{a}^\pm_{\mathbf{k}}$ operators:

$$\hat{a}^-_{\mathbf{k}} = \alpha_k^* \hat{b}^-_{\mathbf{k}} + \beta_k \hat{b}^+_{-\mathbf{k}} \qquad \text{and} \qquad \hat{a}^+_{\mathbf{k}} = \alpha_k \hat{b}^+_{\mathbf{k}} + \beta_k^* \hat{b}^-_{-\mathbf{k}} \ . \tag{2.3.43}$$

Both sets of operators can be used to construct a basis of states in the Hilbert space:

$$\hat{a}^-_{\mathbf{k}}|0\rangle_a = 0 \qquad \hat{b}^-_{\mathbf{k}}|0\rangle_b = 0 \ , \tag{2.3.44}$$

and

$$|m_{\mathbf{k}_1}, n_{\mathbf{k}_2}, \cdots\rangle_a = \frac{1}{\sqrt{m!n!\cdots}}\left[(a^+_{\mathbf{k}_1})^m (a^+_{\mathbf{k}_2})^n \cdots\right]|0\rangle_a \ , \tag{2.3.45}$$

$$|m_{\mathbf{k}_1}, n_{\mathbf{k}_2}, \cdots\rangle_b = \frac{1}{\sqrt{m!n!\cdots}}\left[(b^+_{\mathbf{k}_1})^m (b^+_{\mathbf{k}_2})^n \cdots\right]|0\rangle_b \ . \tag{2.3.46}$$

[8] Changing $v_k(\tau)$ while keeping $\hat{v}$ fixed, changes $\hat{a}^\pm_{\mathbf{k}}$ [cf. (2.2.30)] and hence changes the vacuum $|0\rangle$ and the excited states $|m, n, \cdots\rangle$.

It should be clear that the b-states are in general *different* form the a-states. In particular, the b-vacuum contains a-particles:

$$
\begin{aligned}
{}_b\langle 0|\hat{N}^{(a)}_{\mathbf{k}}|0\rangle_b &= {}_b\langle 0|\hat{a}^+_{\mathbf{k}}\hat{a}^-_{\mathbf{k}}|0\rangle_b && (2.3.47)\\
&= {}_b\langle 0|(\alpha_k\hat{b}^+_{\mathbf{k}} + \beta^*_k\hat{b}^-_{-\mathbf{k}})(\alpha^*_k\hat{b}^-_{\mathbf{k}} + \beta_k\hat{b}^+_{-\mathbf{k}})|0\rangle_b && (2.3.48)\\
&= |\beta_k|^2\delta(0)\ . && (2.3.49)
\end{aligned}
$$

The divergent factor $\delta(0)$ arises because we are considering an infinite spatial volume, but the mean density of a-particles in the b-vacuum is finite (and typically not zero):

$$
n \equiv \int \mathrm{d}^3\mathbf{k}\ n_{\mathbf{k}} = \int \mathrm{d}^3\mathbf{k}\ |\beta_k|^2\ . \tag{2.3.50}
$$

2.3.3 Choice of the Physical Vacuum

Clearly, we are still missing some essential physical input to define the unique vacuum state.

Vacuum in Minkowski Space

How do we usually do this? In a *time-independent* spacetime a preferable set of mode functions and thus an unambiguous physical vacuum can be defined by requiring that the expectation value of the Hamiltonian in the vacuum state is minimized. To illustrate this let us consider the Mukhanov-Sasaki equation in Minkowski space (i.e. the $a \equiv 0$ limit of (2.2.22)):

$$
v_k'' + k^2 v_k = 0\ . \tag{2.3.51}
$$

We aim to find the mode functions v_k that minimize the expectation value of the Hamiltonian in the vacuum. We will therefore compute ${}_v\langle 0|\hat{H}|0\rangle_v$ for an arbitrary mode function v and then find the preferred function v that minimize the result. In terms of our mode expansion, the Hamiltonian (2.3.35) becomes

$$
\hat{H} = \frac{1}{2}\int \mathrm{d}^3\mathbf{k}\Big[\hat{a}^-_{\mathbf{k}}\hat{a}^-_{-\mathbf{k}}F^*_k + \hat{a}^+_{\mathbf{k}}\hat{a}^+_{-\mathbf{k}}F_k + \big(2\hat{a}^+_{\mathbf{k}}\hat{a}^-_{\mathbf{k}} + \delta(0)\big)\,E_k\Big]\ , \tag{2.3.52}
$$

where

$$
\begin{aligned}
E_k &\equiv |v_k'|^2 + k^2|v_k|^2\ , && (2.3.53)\\
F_k &\equiv v_k'^2 + k^2 v_k^2\ . && (2.3.54)
\end{aligned}
$$

Since $\hat{a}^-_{\mathbf{k}}|0\rangle_v = 0$, we have

$$
{}_v\langle 0|\hat{H}|0\rangle_v = \frac{\delta(0)}{4}\int \mathrm{d}^3\mathbf{k}\,E_k\ . \tag{2.3.55}
$$

Dividing out the uninteresting divergence, $\delta(0)$, we infer that the energy density in the vacuum state is

$$
\varepsilon = \frac{1}{4}\int \mathrm{d}^3\mathbf{k}\,E_k\ . \tag{2.3.56}
$$

It is clear that this is minimized if each k–mode E_k is minimized separately. We therefore need to determine the v_k and v_k' that minimize the expression

$$
E_k = |v_k'|^2 + k^2|v_k|^2\ . \tag{2.3.57}
$$

We mustn't forget that the mode functions χ_k satisfy the normalization (2.2.29),

$$v_k' v_k^* - v_k v_k^{*\prime} = -i \ . \tag{2.3.58}$$

Using the parameterization $v_k = r_k e^{i\alpha_k}$, for real r_k and α_k, (2.3.58) becomes

$$r_k^2 \alpha_k' = -\frac{1}{2} \ , \tag{2.3.59}$$

and (2.3.57) gives

$$\begin{aligned} E_k &= r_k'^2 + r_k^2 \alpha_k'^2 + k^2 r_k^2 & (2.3.60) \\ &= r_k'^2 + \frac{1}{4r_k^2} + k^2 r_k^2 \ . & (2.3.61) \end{aligned}$$

It is easily seen that (2.3.61) is minimized if $r_k' = 0$ and $r_k = \frac{1}{\sqrt{2k}}$. Integrating (2.3.59) gives $\alpha_k = -k\tau$ (up to an irrelevant constant that doesn't affect any observables; e.g. this constant phase factor drops out in the computation of the power spectrum) and hence

$$v_k(\tau) = \frac{1}{\sqrt{2k}} e^{-ik\tau} \ . \tag{2.3.62}$$

This defines the preferred mode functions for fluctuations in Minkowski space. For these mode functions we find $E_k = k \equiv \omega_k$ and $F_k = 0$, so the Hamiltonian is

$$\hat{H} = \int \mathrm{d}^3\mathbf{k} \ \omega_k \left[\hat{a}_{\mathbf{k}}^+ \hat{a}_{\mathbf{k}}^- + \frac{1}{2}\delta(0) \right] \ . \tag{2.3.63}$$

Hence, the Hamiltonian is diagonal in the eigenbasis of the occupation number operator $\hat{N}_{\mathbf{k}} \equiv \hat{a}_{\mathbf{k}}^+ \hat{a}_{\mathbf{k}}^-$.

Vacuum in Time-Dependent Spacetimes

The vacuum prescription which we just applied to Minkowski space does *not* generalize straightforwardly to *time-dependent* spacetimes. In this case the mode equation (2.2.22) involves time-dependent frequencies $\omega_k(\tau)$ and the 'minimum-energy vacuum' depends on the time τ_0 at which it is defined. Repeating the above argument, one can nevertheless determine the vacuum which *instantaneously* minimizes the expectation value of the Hamiltonian at some time τ_0. One finds that the initial conditions

$$v_k(\tau_0) = \frac{1}{\sqrt{2\omega_k(\tau_0)}} e^{-i\omega_k(\tau_0)\tau_0} \ , \qquad v_k'(\tau_0) = -i\omega_k(\tau_0)\chi_k(\tau_0) \tag{2.3.64}$$

select the preferred mode functions which determine the vacuum $|0\rangle_{\tau_0}$. However, since $\omega_k(\tau)$ changes with time, the mode functions satisfying (2.3.64) at $\tau = \tau_0$ will typically be different from the mode functions that satisfy the same conditions at a different time $\tau_1 \neq \tau_0$. This implies that $|0\rangle_{\tau_1} \neq |0\rangle_{\tau_0}$ and the state $|0\rangle_{\tau_0}$ is not the lowest-energy state at a later time τ_1.

Bunch-Davies Vacuum

How do we resolve this ambiguity for the inflationary quasi-de Sitter spacetime?

From fig. 2.2 we note that a sufficiently early times (large negative conformal time τ) all modes of cosmological interest were deep inside the horizon:

$$\frac{k}{aH} \sim |k\tau| \gg 1 \qquad \text{(subhorizon)} \ . \tag{2.3.65}$$

This means that in the remote past all observable modes had time-independent frequencies; e.g. in perfect de Sitter space:

$$\omega_k^2 = k^2 - \frac{2}{\tau^2} \to k^2 \ . \tag{2.3.66}$$

The corresponding modes are therefore not affected by gravity and behave just like in Minkowski space:

$$v_k'' + k^2 v_k = 0 \ . \tag{2.3.67}$$

The two independent solutions of (2.3.67) are $v_k \propto e^{\pm ik\tau}$. As we have seen above only the positive frequency mode $v_k \propto e^{-ik\tau}$ is the 'minimal excitation state', cf. eq. (2.3.62).

Given that all modes have time-independent frequencies at sufficiently early times, we can now avoid the ambiguity in defining the initial conditions for the mode functions that afflicts the treatment in more general time-dependent spacetimes. In practice, this means solving the Mukhanov-Sasaki equation with the (Minkowski) initial condition

$$\lim_{\tau\to-\infty} v_k(\tau) = \frac{1}{\sqrt{2k}} e^{-ik\tau} \ . \tag{2.3.68}$$

This defines a preferable set of mode functions and a unique physical vacuum, the *Bunch-Davies vacuum.*

2.3.4 Zero-Point Fluctuations in De Sitter

We are now ready to apply the formalism to de Sitter space.

De Sitter Mode Functions

Recall that the Mukhanov-Sasaki equation in de Sitter is

$$v_k'' + \left(k^2 - \frac{2}{\tau^2}\right) v_k = 0 \ . \tag{2.3.69}$$

This has the following exact solution

$$v_k(\tau) = \alpha \frac{e^{-ik\tau}}{\sqrt{2k}} \left(1 - \frac{i}{k\tau}\right) + \beta \frac{e^{ik\tau}}{\sqrt{2k}} \left(1 + \frac{i}{k\tau}\right) \ . \tag{2.3.70}$$

The initial condition (2.3.68) fixes $\beta = 0$, $\alpha = 1$, and, hence, the unique mode function is

$$v_k(\tau) = \frac{e^{-ik\tau}}{\sqrt{2k}} \left(1 - \frac{i}{k\tau}\right) \ . \tag{2.3.71}$$

This determines the future evolution of the mode including its superhorizon dynamics:

$$\lim_{k\tau\to 0} v_k(\tau) = \frac{1}{i\sqrt{2}} \cdot \frac{1}{k^{3/2}\,\tau} \ . \tag{2.3.72}$$

Zero-Point Fluctuations

Knowledge of the mode functions for canonically-normalized fields in de Sitter space allows us to compute the effect of quantum zero-point fluctuations:

$$\begin{aligned}
\langle \hat{v}_{\mathbf{k}} \hat{v}_{\mathbf{k}'} \rangle &= \langle 0 | \hat{v}_{\mathbf{k}} \hat{v}_{\mathbf{k}'} | 0 \rangle \\
&= \langle 0 | \left(a^-_{\mathbf{k}} v_k + a^+_{-\mathbf{k}} v^*_k \right) \left(a^-_{\mathbf{k}'} v_{k'} + a^+_{-\mathbf{k}'} v^*_{k'} \right) | 0 \rangle \\
&= v_k v^*_{k'} \langle 0 | a^-_{\mathbf{k}} a^+_{-\mathbf{k}'} | 0 \rangle \\
&= v_k v^*_{k'} \langle 0 | \left[a^-_{\mathbf{k}} \, , \, a^+_{-\mathbf{k}'} \right] | 0 \rangle \\
&= |v_k|^2 \, \delta(\mathbf{k} + \mathbf{k}') \\
&\equiv P_v(k) \, \delta(\mathbf{k} + \mathbf{k}') \, .
\end{aligned}$$

On superhorizon scales this approaches [cf. eq. (2.3.72)]

$$P_v = \frac{1}{2k^3} \frac{1}{\tau^2} = \frac{1}{2k^3} (aH)^2 \, . \tag{2.3.73}$$

All power spectra for fields in de Sitter space are simple rescalings of this power spectrum for the canonically-normalized field. For example, the power spectrum of curvature perturbations is

$$P_\zeta = \frac{1}{z^2} P_v \, . \tag{2.3.74}$$

2.4 Curvature Perturbations from Inflation

Strictly speaking, the curvature fluctuations $\zeta = z^{-1} v$ are ill-defined in perfect de Sitter since $z^2 = 2a^2 \varepsilon$ vanishes in that limit. This is just a reflection of the fact that for perfect de Sitter inflation never ends, so ζ is meaningless. In reality, we know that inflation has to end and that the spacetime during inflation has to deviate from the de Sitter idealization. This deviation is described by the small but finite slow-roll parameter ε.

2.4.1 Results for Quasi-De Sitter

In quasi-de Sitter space, the curvature perturbation ζ is well-defined, and its power spectrum follows directly from eq. (2.3.73),

$$P_\zeta = \frac{1}{z^2} P_v = \frac{1}{4k^3} \frac{H^2}{\varepsilon} = \frac{1}{2k^3} \frac{H^4}{\dot{\phi}^2} \, . \tag{2.4.75}$$

Since ζ freezes at horizon crossing, we may evaluate the r.h.s. at $k = aH$. The power spectrum then becomes a function purely of k:

$$P_\zeta(k) = \left. \frac{1}{4k^3} \frac{H^2}{\varepsilon} \right|_{k=aH} \, , \tag{2.4.76}$$

or in dimensionless form

$$\Delta^2_s(k) \equiv \frac{k^3}{2\pi^2} P_\zeta(k) = \left. \frac{1}{8\pi^2} \frac{H^2}{\varepsilon} \right|_{k=aH} \, . \tag{2.4.77}$$

Since H and possibly ε are now functions of time, the power spectrum will deviate slightly from the scale-invariant form $\Delta^2_s \sim k^0$. The common way to quantify the deviation from scale-invariance is via the scalar spectral index n_s:

$$n_s - 1 \equiv \frac{d \ln \Delta^2_s}{d \ln k} \, . \tag{2.4.78}$$

We split the r.h.s. into two factors

$$\frac{d\ln\Delta_{\rm s}^2}{d\ln k} = \frac{d\ln\Delta_{\rm s}^2}{dN} \times \frac{dN}{d\ln k} \ . \tag{2.4.79}$$

The derivative with respect to e-folds is

$$\frac{d\ln\Delta_{\rm s}^2}{dN} = 2\,\frac{d\ln H}{dN} - \frac{d\ln\varepsilon}{dN} \ . \tag{2.4.80}$$

The first term is just -2ε and the second term is $-\eta$ (see Chapter 1). The second factor in eq. (2.4.79) is evaluated by recalling the horizon crossing condition $k = aH$, or

$$\ln k = N + \ln H \ . \tag{2.4.81}$$

Hence,

$$\frac{dN}{d\ln k} = \left[\frac{d\ln k}{dN}\right]^{-1} = \left[1 + \frac{d\ln H}{dN}\right]^{-1} \approx 1 + \varepsilon \ . \tag{2.4.82}$$

To first order in the Hubble slow-roll parameters we therefore find

$$n_{\rm s} - 1 = -2\varepsilon - \eta \, . \tag{2.4.83}$$

The parameter n_s is an interesting probe of the inflationary dynamics. It measures deviations of the perfect de Sitter limit: H, $\dot{H}$, and $\ddot{H}$.

2.4.2 Systematic Slow-Roll Expansion*

The same results may be derived as a systematic expansion in slow-roll parameters:

$$\varepsilon \equiv -\frac{\dot{H}}{H^2} \ , \qquad \eta \equiv \frac{\dot{\varepsilon}}{H\varepsilon} \ , \qquad \kappa \equiv \frac{\dot{\eta}}{H\eta} \ . \tag{2.4.84}$$

This will involve a slow-roll expansion of the Mukhanov-Sasaki equation (2.2.22):

$$v_k'' + \left(k^2 - \frac{z''}{z}\right) v_k = 0 \ . \tag{2.4.85}$$

Given $z^2 = 2a^2\varepsilon$ we find

$$\frac{z'}{z} \ = \ (aH)\left[1 + \frac{1}{2}\eta\right] \qquad \text{(exact)} \ , \tag{2.4.86}$$

$$\frac{z''}{z} \ = \ (aH)^2\left[2 - \varepsilon + \frac{3}{2}\eta - \frac{1}{2}\varepsilon\eta + \frac{1}{4}\eta^2 + \eta\kappa\right] \qquad \text{(exact)} \ . \tag{2.4.87}$$

Despite the appearance of the slow-roll parameters, both expressions above are exact. From the definition of ε we furthermore get

$$\frac{d}{d\tau}\left(\frac{1}{aH}\right) = \varepsilon - 1 \qquad \text{(exact)} \ . \tag{2.4.88}$$

Expanding the expressions to first order in the slow-roll parameters, $\{\varepsilon, |\eta|, |\kappa|\} \ll 1$, gives

$$aH = -\frac{1}{\tau}(1 + \varepsilon) \qquad \text{(first order in SR)} \ , \tag{2.4.89}$$

and

$$\frac{z''}{z} = \frac{1}{\tau^2}\left[2 + 3\left(\varepsilon + \frac{1}{2}\eta\right)\right] \equiv \frac{\nu^2 - \frac{1}{4}}{\tau^2} \qquad \text{(first order in SR)} \ , \tag{2.4.90}$$

where

$$\nu \equiv \frac{3}{2} + \varepsilon + \frac{1}{2}\eta \ . \tag{2.4.91}$$

For constant ν, the Mukhanov-Sasaki equation,

$$v_k'' + \left(k^2 - \frac{\nu^2 - \frac{1}{4}}{\tau^2}\right) v_k = 0 \ , \tag{2.4.92}$$

has an exact solution in terms of Hankel functions of the first and second kind:

$$v_k(\tau) = \sqrt{-\tau}\left[\alpha H_\nu^{(1)}(-k\tau) + \beta H_\nu^{(2)}(-k\tau)\right] \ . \tag{2.4.93}$$

To impose the Bunch-Davies boundary condition at early times, we consider the limit

$$\lim_{k\tau\to-\infty} v_k(\tau) = \sqrt{\frac{2}{\pi}}\left[\alpha\frac{1}{\sqrt{k}}e^{-ik\tau} + \beta\frac{1}{\sqrt{k}}e^{ik\tau}\right] \ , \tag{2.4.94}$$

where we used

$$\lim_{k\tau\to-\infty} H_\nu^{(1,2)}(-k\tau) = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{-k\tau}}e^{\pm ik\tau}e^{\pm i\frac{\pi}{2}(\nu+\frac{1}{2})} \ , \tag{2.4.95}$$

and dropped the unimportant phase factors $e^{\pm i\frac{\pi}{2}(\nu+\frac{1}{2})}$. Comparing (2.4.94) to (2.3.68) we find

$$\beta = 0 \qquad \text{and} \qquad \alpha = \sqrt{\frac{\pi}{2}} \ . \tag{2.4.96}$$

Hence, the Bunch-Davies mode functions to first order in slow-roll are:

$$v_k(\tau) = \sqrt{\frac{\pi}{2}}(-\tau)^{1/2}H_\nu^{(1)}(-k\tau) \ , \tag{2.4.97}$$

To compute the power spectrum of curvature fluctuations, $P_\zeta = z^{-2}P_v$, we use $z \sim \tau^{\frac{1}{2}-\nu}$ (first order in SR)

$$P_\zeta \sim \frac{\pi}{2}(-\tau)^{2\nu}|H_\nu^{(1)}(-k\tau)|^2 \ . \tag{2.4.98}$$

In the superhorizon limit, $-k\tau \ll 1$, this reduces to

$$\Delta_s^2 \equiv \frac{k^3}{2\pi^2}P_\zeta \sim k^{3-2\nu} \ , \tag{2.4.99}$$

where we used

$$\lim_{k\tau\to 0} H_\nu^{(1)}(-k\tau) = \frac{i}{\pi}\Gamma(\nu)\left(\frac{-k\tau}{2}\right)^{-\nu} \ . \tag{2.4.100}$$

Finally, the scale-dependence of the scalar spectrum is

$$n_s - 1 \equiv \frac{d\ln\Delta_s^2}{d\ln k} \ = \ 3 - 2\nu \ , \tag{2.4.101}$$

or, in terms of the slow-roll parameters,

$$n_s - 1 \ = \ -2\varepsilon - \eta \ . \tag{2.4.102}$$

This shows that the spectrum is perfectly scale-invariant in de Sitter space, while slow-roll corrections to de Sitter led to percent-level deviations from $n_s = 1$.

2.5 Gravitational Waves from Inflation

One of the most robust and model-independent predictions of inflation is a stochastic background of gravitational waves with an amplitude given simply by the Hubble scale H during inflation. The simplicity of this prediction means that a measurement of primordial gravitational waves would give clean information about arguably the most important inflationary parameter, namely the energy scale of inflation. Most excitingly, inflationary gravitational waves lead to a unique signature in the polarization of the CMB. A large number of ground-based, balloon and satellite experiments are currently searching for this signal.

The formalism we introduced for the scalar fluctuations can easily be applied to compute the quantum generation of tensor perturbations (i.e. transverse and traceless perturbations to the spatial metric, $\delta g_{ij} = a^2 h_{ij}$). In fact, in this case, our job is considerably simpler due to the fact that first-order tensor perturbation are gauge-invariant and don't backreact on the inflationary background. Expansion of the Einstein-Hilbert action gives the second-order action for tensor fluctuations

$$S = \frac{M_{\rm pl}^2}{8}\int \mathrm{d}\tau\, \mathrm{d}^3\mathbf{x}\; a^2\left[(h_{ij}')^2 - (\nabla h_{ij})^2\right] \ . \tag{2.5.103}$$

Here, we have reintroduced the explicit factor of $M_{\rm pl}^2$ to make h_{ij} manifestly dimensionless. Up to the normalization factor of $\frac{M_{\rm pl}}{2}$ this is the same as the action for a massless scalar field in an FRW universe.

We define the standard Fourier representation for transverse, traceless tensors

$$h_{ij}(\tau,\mathbf{x}) = \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^{3/2}}\sum_{\gamma=+,\times}\epsilon_{ij}^{\gamma}(k)h_{\mathbf{k},\gamma}(\tau)\, e^{i\mathbf{k}\cdot\mathbf{x}} \ , \tag{2.5.104}$$

where $\epsilon_{ii}^{\gamma} = k^i\epsilon_{ij}^{\gamma} = 0$ and $\epsilon_{ij}^{\gamma}\epsilon_{ij}^{\gamma'} = 2\delta_{\gamma\gamma'}$. The fields $h_{\mathbf{k},\gamma}$ describe the two polarization modes of the gravitational waves (+ and ×). Eq. (2.5.103) then becomes

$$S = \sum_{\gamma}\int \mathrm{d}\tau\, \mathrm{d}^3\mathbf{k}\; \frac{a^2}{4}M_{\rm pl}^2\left[(h_{\mathbf{k},\gamma}')^2 - k^2(h_{\mathbf{k},\gamma})^2\right] \ . \tag{2.5.105}$$

For the canonically-normalized fields,

$$v_{\mathbf{k},\gamma} \equiv \frac{a}{2}M_{\rm pl}h_{\mathbf{k},\gamma} \ , \tag{2.5.106}$$

this reads

$$S = \sum_{\gamma}\frac{1}{2}\int \mathrm{d}\tau\, \mathrm{d}^3\mathbf{k}\left[(v_{\mathbf{k},\gamma}')^2 - \underbrace{\left(k^2 - \frac{a''}{a}\right)}_{\equiv\,\omega_k^2(\tau)}(v_{\mathbf{k},\gamma})^2\right] \ . \tag{2.5.107}$$

For a de Sitter background, we have

$$\frac{a''}{a} = \frac{2}{\tau^2} \ . \tag{2.5.108}$$

Eq. (2.5.107) should be recognized as essentially two copies of the action (2.2.20). Hence, we can jump straight to result in eq. (2.3.73):

$$P_v = \frac{1}{2k^3}(aH)^2 \ . \tag{2.5.109}$$

Defining the tensor power spectrum P_t as the sum of the power spectra for each polarization mode of h_{ij}, we find

$$P_t = 2 \cdot P_h = 2 \cdot \left(\frac{2}{aM_{\rm pl}}\right)^2 \cdot P_v = \frac{4}{k^3}\frac{H^2}{M_{\rm pl}^2} \,, \tag{2.5.110}$$

or the dimensionless spectrum

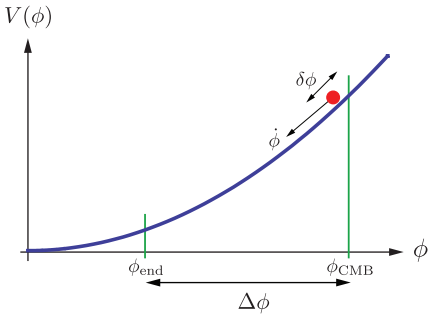
$$\Delta_t^2(k) = \left.\frac{2}{\pi^2}\frac{H^2}{M_{\rm pl}^2}\right|_{k=aH} \,. \tag{2.5.111}$$

This completes our treatment of the quantum generation of scalar and tensor fluctuations in inflation.

2.6 The Lyth Bound

A useful way of normalizing the gravitational wave amplitude is the tensor-to-scalar ratio

$$r \equiv \frac{\Delta_t^2}{\Delta_s^2} = 16\varepsilon = \frac{8}{M_{\rm pl}^2}\frac{\dot\phi^2}{H^2} \,. \tag{2.6.112}$$



It turns out that it is the size of r that determines whether inflationary gravitational waves are detectable in future CMB observations. Roughly, tensors will be observable in the not to distant future, if $r > 0.01$. Note that the tensor-to-scalar ratio relates directly to the evolution of the inflaton as a function of e-folds N

$$r = \frac{8}{M_{\rm pl}^2}\left(\frac{d\phi}{dN}\right)^2 \,. \tag{2.6.113}$$

The total field evolution between the time when CMB fluctuations exited the horizon at $N_{\rm cmb}$ and the end of inflation at $N_{\rm end}$ can therefore be written as the following integral

$$\frac{\Delta\phi}{M_{\rm pl}} = \int_{N_{\rm end}}^{N_{\rm cmb}} {\rm d}N \sqrt{\frac{r}{8}} \,. \tag{2.6.114}$$

During slow-roll evolution, $r(N)$ doesn't evolve much and one may obtain the following approximate relation, called the *Lyth bound*,

$$\frac{\Delta\phi}{M_{\rm pl}} = \mathcal{O}(1) \times \left(\frac{r}{0.01}\right)^{1/2} \,, \tag{2.6.115}$$

where $r \equiv r(N_{\rm cmb})$ is the tensor-to-scalar ratio on CMB scales. Large values of the tensor-to-scalar ratio, $r > 0.01$, therefore correlate with $\Delta\phi > M_{\rm pl}$ or *large-field inflation*. How to make sense of such super-Planckian field excursions in effective field theory remains a key theoretical challenge (see Chapter 6).

3 Contact with Observations

3.1 Introduction

So far, we have computed the power spectra of ζ and h at horizon exit. In this chapter, we show how to relate these results to observations of the cosmic microwave background (CMB) and large-scale structure (LSS). Making this correspondence explicit is crucial if the data is to be used to extract information about the early universe.

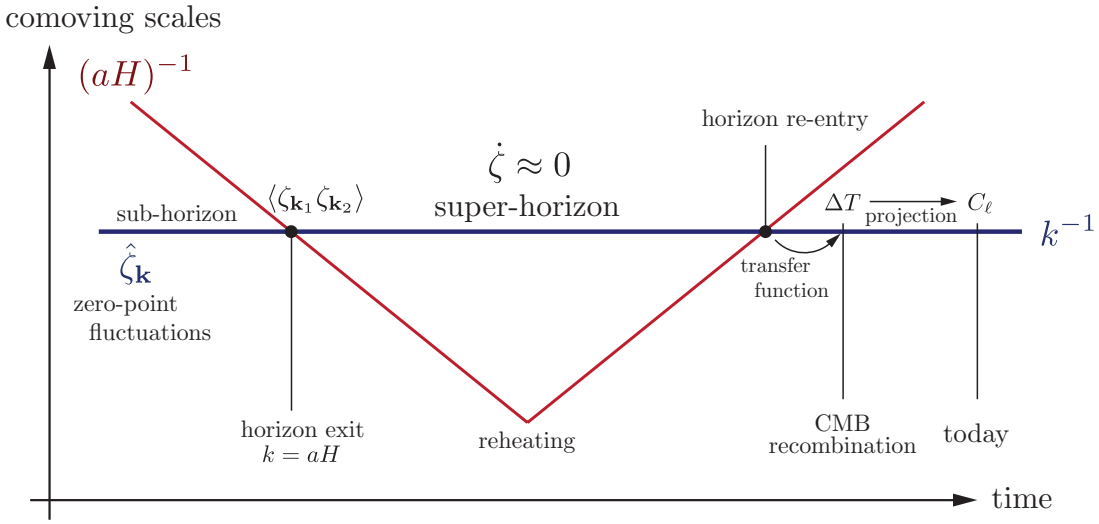


Figure 3.1: *From vacuum fluctuations to CMB anisotropies.*

The challenge is to relate the predictions made at horizon exit (high energies) to the observables after horizon re-entry (low energies). These times are separated by a time interval in which the physics is very uncertain. Not even the equations governing perturbations are well-known. How can we still make predictions? The only reason that we are able to connect late-time observables to inflationary theories is the fact that the wavelengths of the perturbations of interest we outside the horizon during the period from well before the end of inflation until the relatively near present (see fig. 3.1).

In the previous chapter, we showed that the curvature perturbation ζ freezes after horizon crossing for inflation driven by a single scalar field. However, this is not enough. After inflation, the universe becomes filled with matter and radiation, and we need establish under which conditions ζ remains conserved on superhorizon scales. In §3.2, we review two proofs of the conservation of ζ outside of the horizon for adiabatic matter perturbations. In §3.3, we

discuss how the subsequent subhorizon evolution processes the primordial perturbations until the fluctuations get imprinted in the CMB surface of last scattering. This evolution involves well-understood low-energy physics and is hence computable. We will present a simplified CMB code the evaluate the CMB transfer function. Finally, in §3.4 we make some brief comments about the relation between large-scale structure observables and the primordial perturbations.

3.2 Superhorizon (Non)-Evolution*

Establishing rigorously the conditions under which ζ doesn't evolve on superhorizon scales, is crucial for making reliable predictions about late time observables such as the CMB. In §3.2.1, we therefore review Weinberg's proof of the conservation of superhorizon curvature perturbations.[1] Beware, coming from Weinberg, the proof is long, convoluted, but precise. For a more easy-going (if less rigorous) proof, we sketch the 'separate universe' approach[2] in §3.2.2. Readers whose attention span has been sufficiently reduced by the internet[3] may skip to the next section without loss of continuity.

3.2.1 Weinberg's Proof

We start with the gauge-invariant curvature perturbation written in Newtonian gauge variables (see Appendix B)

$$\zeta \equiv -\Psi + H\delta u \ . \tag{3.2.1}$$

Here, Ψ is the scalar metric perturbation and δu is the velocity potential for the total energy-momentum tensor. During slow-roll inflation $\delta u = \delta\phi/\dot{\phi}$. The rate of change of ζ is[4]

$$\dot{\zeta} = X + \mathcal{O}(\tfrac{k^2}{a^2H^2}) \ , \quad \text{where} \quad X \equiv \frac{\dot{\bar{\rho}}\delta p - \dot{\bar{p}}\delta\rho}{3(\bar{\rho}+\bar{p})^2} \ . \tag{3.2.2}$$

Thus ζ is conserved in the limit of small wavenumber if and only if $X = 0$ in this limit.[5] We then proceed in two steps:

1. Following Weinberg, we prove that *whatever the constituents of the universe and the classical equations governing them may be*, these equations always have a physical solution for which $X \to 0$ and ζ approaches a non-zero constant ζ^o in the limit $k \to 0$. This is called the *adiabatic* mode.

2. We then prove that single-field inflation excites the adiabatic mode.

Together these two facts prove that the curvature perturbations set up during single-field inflation remain conserved after inflation.

[1] Weinberg (arXiv:astro-ph/0302326).
[2] Wands et al. (arXiv:astro-ph/0003278).
[3] Nicholas Carr, *The Shallows* (What the internet is doing to our brains).
[4] This follows from the conservation of the stress-tensor.
[5] We see that X vanishes when the pressure $\bar{p} + \delta p$ is a function only of the perturbed energy density $\bar{\rho} + \delta\rho$.

Adiabatic Modes in Cosmology

Weinberg's proof is based on the observation that in the special case of a spatially homogeneous universe, the field equations and dynamical equation for matter and radiation are invariant under coordinate transformations that are *not* symmetries of the unperturbed metric.[6]

His explicit proof was performed in Newtonian gauge whose properties are reviewed in Appendix B. For convenience we here collect some of its most relevant equations:

Newtonian gauge. Scalar perturbations to the metric in Newtonian gauge are

$$\mathrm{d}s^2 = (-1-2\Phi)\mathrm{d}t^2 + a^2(t)(1-2\Psi)\mathrm{d}\mathbf{x}^2 \equiv (\bar{g}_{\mu\nu}+\hat{g}_{\mu\nu})\mathrm{d}x^\mu \mathrm{d}x^\nu \ . \tag{3.2.3}$$

The Einstein field equations are

$$\begin{aligned}
\nabla^2\Psi - a^2\ddot{\Psi} - 6a\dot{a}\dot{\Psi} - a\dot{a}\dot{\Phi} - (4\dot{a}^2+2a\ddot{a})\Phi &= 4\pi G a^2\left[\delta\rho - \delta p - \nabla^2\pi\right] \ , && (3.2.4)\\
\partial_i\partial_j[\Psi-\Phi] &= 8\pi G a^2 \partial_i\partial_j\pi \ , && (3.2.5)\\
\partial_i[a\dot{\Psi}+\dot{a}\Phi] &= -4\pi G a(\bar{\rho}+\bar{p})\partial_i\delta u \ , && (3.2.6)\\
3a^2\ddot{\Psi} + 6a\dot{a}\dot{\Psi} + \nabla^2\Phi + 3a\dot{a}\dot{\Phi} + 6a\ddot{a}\Phi &= 4\pi G a^2\left[\delta\rho + \delta p + \nabla^2\pi\right] \ . && (3.2.7)
\end{aligned}$$

We will need to consider spacetime coordinate transformations of the form

$$x^\mu \to x^\mu + \epsilon^\mu(x) \ . \tag{3.2.8}$$

Metric perturbations transform as

$$\begin{aligned}
\Delta\hat{g}_{00} &= -2\partial_0\epsilon_0 \ , && (3.2.9)\\
\Delta\hat{g}_{i0} &= -\partial_0\epsilon_i - \partial_i\epsilon_0 + 2H\epsilon_i \ , && (3.2.10)\\
\Delta\hat{g}_{ij} &= -\partial_i\epsilon_j - \partial_j\epsilon_i + 2a\dot{a}\delta_{ij}\epsilon_0 \ . && (3.2.11)
\end{aligned}$$

Similarly, the perturbations of the stress-tensor transform as

$$\Delta\delta p = \dot{\bar{p}}\epsilon_0 \ , \quad \Delta\delta\rho = \dot{\bar{\rho}}\epsilon_0 \ , \quad \text{and} \quad \Delta\delta u = -\epsilon_0 \ . \tag{3.2.12}$$

In Newtonian gauge, general first-order spatially homogeneous scalar and tensor perturbations to the metric take the form

$$\hat{g}_{00} = -2\Phi(t) \ , \quad \hat{g}_{i0} = 0 \ , \quad \text{and} \quad \hat{g}_{ij} - 2a^2(t)\Psi(t)\delta_{ij} + a^2(t)h_{ij}(t) \ , \tag{3.2.13}$$

where h_{ij} are transverse and traceless tensor perturbations. We now want to find those gauge transformations (3.2.8) that preserve both the conditions of the Newtonian gauge and of spatial homogeneity. The field equations (not matter what they are!) will necessarily be invariant under those transformations. Eq. (3.2.9) shows that in order for $\hat{g}_{00}$ to remain spatially homogeneous, the transformation parameter ϵ_0 must be of the form

$$\epsilon_0(t,\mathbf{x}) = \epsilon(t) + \chi(\mathbf{x}) \ , \tag{3.2.14}$$

so that

$$\Delta\Phi = \dot{\epsilon} \ . \tag{3.2.15}$$

[6] In this respect, the theorem is analogous to the Goldstone theorem of QFT.

Eq. (3.2.10) then shows that in order for $\hat{g}_{i0}$ to remain equal to zero, ϵ_i must have the form

$$\epsilon_i(t,\mathbf{x}) = a^2(t) f_i(\mathbf{x}) - a^2(t)\partial_i\chi(\mathbf{x}) \int \frac{\mathrm{d}t}{a^2(t)} \ . \tag{3.2.16}$$

Eq. (3.2.11) then implies

$$\Delta\hat{g}_{ij} = -a^2(\partial_i f_j + \partial_j f_i) + 2\delta_{ij} a\dot{a}[\epsilon + \chi] - 2\partial_{ij}\chi \int \frac{\mathrm{d}t}{a^2} \ . \tag{3.2.17}$$

In order not to introduce any spatial dependence in $\hat{g}_{ij}$, the parameter χ must be a constant, in which case it can be set to zero simply by absorbing it into the definition of $\epsilon(t)$. We must also take f_i to be of the form $f_i = \omega_{ij}x^j$, where ω_{ij} is a constant matrix. Hence, we get

$$\Delta\hat{g}_{ij} = -a^2[\omega_{ij} + \omega_{ji}] + 2\delta_{ij} a\dot{a}\epsilon \ . \tag{3.2.18}$$

Comparing this to

$$\Delta\hat{g}_{ij} = -2a^2\Delta\Psi\delta_{ij} + a^2\Delta h_{ij} \ , \tag{3.2.19}$$

we find

$$\Delta\Psi = \tfrac{1}{3}\omega_{ii} - H\epsilon \ , \tag{3.2.20}$$

$$\Delta h_{ij} = -\omega_{ij} - \omega_{ji} + \tfrac{2}{3}\omega_{kk}\delta_{ij} \ . \tag{3.2.21}$$

The corresponding gauge transformations of quantities appearing in the stress-tensor are given in eq. (3.2.12). Since $[\hat{g}_{\mu\nu}, T_{\mu\nu}]$ and $[\hat{g}_{\mu\nu} + \Delta\hat{g}_{\mu\nu}, T_{\mu\nu} + \Delta T_{\mu\nu}]$ are both solutions of the field equations, their difference must also be a solution. We conclude that there is always a spatially homogeneous solution of the Newtonian gauge field equations, with scalar perturbations

$$\Psi = H\epsilon - \tfrac{1}{3}\omega_{kk} \ , \quad \Phi = -\dot{\epsilon} \ , \qquad \delta p = -\dot{\bar{p}}\epsilon \ , \quad \delta\rho = -\dot{\bar{\rho}}\epsilon \ , \quad \delta u = \epsilon \ , \quad \pi = 0 \ . \tag{3.2.22}$$

Exercise. Confirm that eq. (3.2.22) satisfies the Einstein equations (3.2.4)–(3.2.7).

Substituting eq. (3.2.22) into eq. (3.2.1), gives ζ the time-independent value

$$\zeta^o = \tfrac{1}{3}\omega_{kk} \ . \tag{3.2.23}$$

Similarly, there is a spatially homogeneous solution with a tensor perturbation

$$h_{ij} \propto \omega_{ij} - \tfrac{1}{3}\omega_{kk}\delta_{ij} \ , \quad \pi_{ij} = 0 \ . \tag{3.2.24}$$

So far, this may seem a bit like an empty statement: ϵ is an arbitrary function of time, and ω_{ij} is an unrelated arbitrary constant matrix. But for zero wavenumber, these are just gauge modes! To find physical modes, we have to extend these solutions to non-zero wavenumber. For the tensor modes, there is no subtlety: in this case, no field equations disappear for zero wavenumber, so the solution with h_{ij} time-independent automatically has an extension to a physical mode with non-zero wavenumber. For scalars, however, we need to work a bit more.

First, we note that the 'constraint'[7] equation (3.2.5) vanishes for zero wavenumber. To get a physical mode we must impose on the perturbations the condition

$$\Phi = \Psi \ . \tag{3.2.25}$$

[7] It is important that we don't have to use an evolution equation. The evolution equations are still completely arbitrary.

From eq. (3.2.22), we therefore get

$$\dot{\epsilon} = -H\epsilon + \zeta^o \ . \tag{3.2.26}$$

For $\zeta^o \neq 0$, this differential equation has the solution

$$\epsilon(t) = \frac{\zeta^o}{a(t)} \int_{\mathcal{T}}^{t} a(t')\mathrm{d}t' \ , \tag{3.2.27}$$

with the integration limit $\mathcal{T}$ arbitrary. This implies the following solutions for the long-wavelength metric perturbations

$$\Psi = \Phi = \zeta^o \left[-1 + \frac{H(t)}{a(t)} \int_{\mathcal{T}}^{t} a(t')\mathrm{d}t' \right] \ , \tag{3.2.28}$$

and matter perturbations

$$\frac{\delta p}{\dot{\bar{p}}} = \frac{\delta \rho}{\dot{\bar{\rho}}} = -\delta u = -\frac{\zeta^o}{a(t)} \int_{\mathcal{T}}^{t} a(t')\mathrm{d}t' \ . \tag{3.2.29}$$

Notice that this solution satisfies $X = 0$ and hence is called *adiabatic*.

Finally, we observe that two solutions of the form (3.2.27) with different values of $\mathcal{T}$, but the same ζ^o are still solutions of (3.2.26). The difference of these two solutions is also a solution, but with $\zeta^o = 0$. This solution is a decaying mode, so it can typically be ignored at late times, but its existence is important for the counting of the number of adiabatic solutions.

Adiabatic Modes in Single-Field Inflation

Weinberg proved that there are always *two* independent *physical* adiabatic solutions of the differential equations governing the scalar fluctuations. Hence, if these equations have no more than two independent solutions, then any perturbations must be adiabatic! This is the case for single-field inflation. We are done. The perturbations set up by single-field inflation are necessarily adiabatic and will therefore remain constant on superhorizon scales even after inflation ends.

3.2.2 Separate Universe Approach

A popular alternative to Weinberg's proof of the conservation of ζ is the separate universe approach. We here briefly sketch the basic idea.[8]

Consider different super-horizon sized regions of the universe to be evolving like separate FRW universes, where density and pressure may take different values, but that are locally homogeneous. After patching together the different regions, this can be used to follow the evolution of the curvature perturbation with time. In fig. 3.2 we show two locally homogeneous regions (a) and (b), separated by a coordinate distance λ on an initial hypersurface (e.g. uniform-density hypersurface) at time t_1. Note that ζ may be interpreted as a local rescaling of the background scale factor, $a(t, \vec{x}) = a(t)e^{\zeta(t,\vec{x})}$. Moreover, adiabatic fluctuations are completely determined by ζ alone. For adiabatic fluctuations the two regions (a) and (b) only differ by a shift in time, $a(t)e^{\zeta_a}$ and $a(t)e^{\zeta_b}$. In particular, for adiabatic fluctuations, the density and the pressure perturbations simply correspond to shifts forward and backwards in time along the background solution. The two regions therefore have identical, but slightly time-shifted,

[8] More details can be found in Wands et al. (arXiv:astro-ph/0003278).

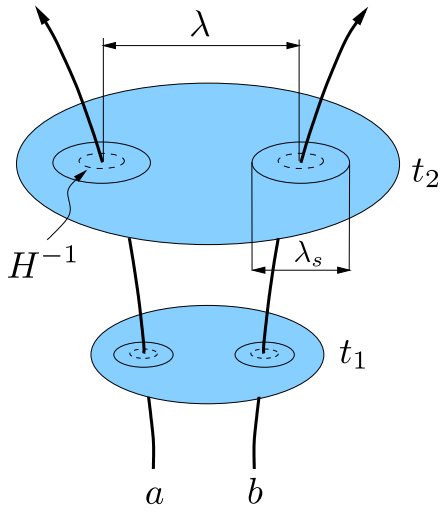


Figure 3.2: (Figure reproduced from Wands et al.) *Illustration of the separate universe approach.*

evolutions. Under these conditions, uniform-density hypersurfaces are separated by a uniform expansion and hence the curvature perturbation, ζ, remains constant.

For non-adiabatic perturbations it is no longer possible to define a simple shift to describe both the density and pressure perturbation. The existence of a non-zero pressure perturbation on uniform-density hypersurfaces changes the equation of state in different regions of the universe and hence leads to perturbations in the expansion along different worldlines between uniform-density hypersurfaces.

This is the gist of the separate universe argument for the conservation of ζ for adiabatic perturbations.

3.3 From Vacuum Fluctuations to CMB Anisotropies

Next, we discuss the evolution of perturbations after modes re-enter the horizon and eventually become the anisotropies we observe in the CMB. A complete discussion of CMB physics clearly would require a whole course of its own.[9] Here, we sketch schematically the different effects that have to be accounted for in order to related the observed CMB spectrum to the primordial fluctuations.

3.3.1 Statistics of Temperature Anisotropies

Inside of the horizon the curvature perturbations ζ lead to density fluctuations $\delta\rho$ in the primordial plasma. When the universe cools, neutral hydrogen forms and photons decouple. These photons become the CMB and the primordial density perturbations get imprinted in the CMB anisotropies. Fig. 3.3 shows a map of the measured CMB temperature fluctuations $\Delta T(\vec{n})$ relative to the background temperature $T_0 = 2.7$ K, where $\vec{n}$ denotes the direction in sky. The

[9]The formation of the CMB and its acoustic oscillations involves some fascinating physics. If you haven't studied this before, I strongly recommend reading Scott Dodelson's brilliant book (Dodelson, *Modern Cosmology.*) and/or listening to Matias Zaldarriaga's recent lectures at the PiTP summer school (http://video.ias.edu/pitp-2011).

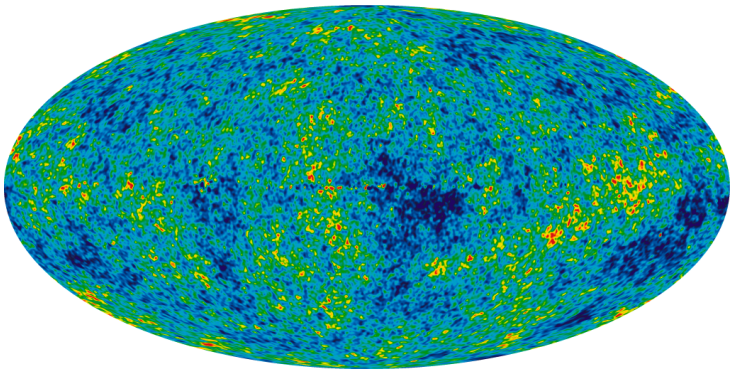


Figure 3.3: *Temperature fluctuations in the CMB. Blue spots represent directions on the sky where the CMB temperature is* $\sim 10^{-4}$ *below the mean,* $T_0 = 2.7$ *K. This corresponds to photons losing energy while climbing out of the gravitational potentials of overdense regions in the early universe. Yellow and red indicate hot (underdense) regions. The statistical properties of these fluctuations contain important information about both the background evolution and the initial conditions of the universe.*

harmonic expansion of this map is

$$\Theta(\vec{n}) \equiv \frac{\Delta T(\vec{n})}{T_0} = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\vec{n})\,, \tag{3.3.30}$$

where

$$a_{\ell m} = \int \mathrm{d}\Omega\, Y^*_{\ell m}(\vec{n})\Theta(\vec{n})\,. \tag{3.3.31}$$

Here, $Y_{\ell m}(\vec{n})$ are the standard spherical harmonics on a two-sphere with $\ell = 0$, $\ell = 1$ and $\ell = 2$ corresponding to the monopole, dipole and quadrupole, respectively. The magnetic quantum numbers satisfy $m = -\ell, \ldots, +\ell$. The multipole moments $a_{\ell m}$ may be combined into the rotationally-invariant angular power spectrum[10]

$$C_\ell^{TT} = \frac{1}{2\ell+1}\sum_m \langle a^*_{\ell m} a_{\ell m}\rangle\,, \qquad \text{or} \qquad \langle a^*_{\ell m} a_{\ell' m'}\rangle = C_\ell^{TT}\delta_{\ell\ell'}\delta_{mm'}\,. \tag{3.3.32}$$

The angular power spectrum is an important tool in the statistical analysis of the CMB. It describes the cosmological information contained in the millions of pixels of a CMB map in terms of a much more compact data representation. Fig. 3.4 shows recent measurements of the CMB angular power spectrum. The figure also shows a fit of the theoretical prediction for the CMB spectrum to the data. The theoretical curve depends both on the background cosmological parameters and on the spectrum of initial fluctuations. We hence can use the CMB as a probe of both.

3.3.2 Transfer Function and Projection Effects

The linear evolution which relates ζ and ΔT is summarized by the *transfer function* $\Delta_{T\ell}(k)$ appearing in the following integral over momentum modes

$$a_{\ell m} = 4\pi(-i)^\ell \int \frac{\mathrm{d}^3 k}{(2\pi)^3}\,\Delta_{T\ell}(k)\,\zeta_{\vec{k}}\,Y_{\ell m}(\hat{k})\,. \tag{3.3.33}$$

[10] This is the Legendre transform of the real space two-point correlation function $\langle \Delta T(\vec{n})\Delta T(\vec{n}')\rangle$.

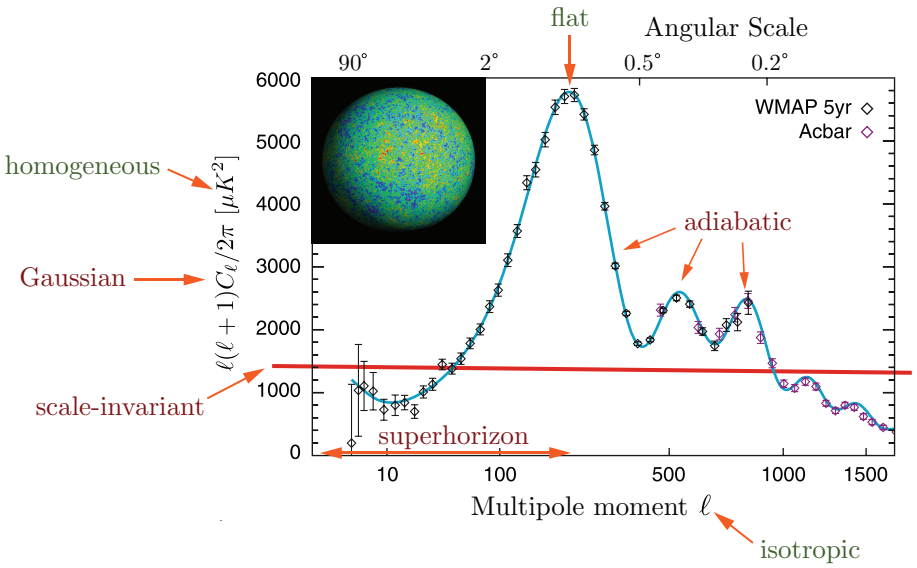


Figure 3.4: *Angular power spectrum of CMB temperature fluctuations.*

The transfer function may be written as the *line-of-sight integral* over physical source terms $S_T(k,\tau)$ and a geometric projection factor $P_{T\ell}(k[\tau_0-\tau])$ (combinations of Bessel functions)[11]

$$\Delta_{T\ell}(k) = \int_0^{\tau_0} \mathrm{d}\tau \underbrace{S_T(k,\tau)}_{\text{Sources}} \underbrace{P_{T\ell}(k[\tau_0-\tau])}_{\text{Projection}} , \tag{3.3.34}$$

where τ_0 is conformal time today. The transfer functions $\Delta_{T\ell}(k)$ generally have to be computed numerically using Boltzmann codes such as CMBFast or CAMB.[12] In the next section, I will present a simplified treatment (the so-called two fluid approximation) that captures the most relevant physics without having to solve the complete hierarchy of Boltzmann equations. You are then invited to write a simple Mathematica code to calculate the CMB transfer function yourself!

Substituting (3.3.78) into (3.3.32) and using the identity

$$\sum_{m=-\ell}^{\ell} Y_{\ell m}(\hat{k}) Y_{\ell m}(\hat{k}') = \frac{2\ell+1}{4\pi} P_\ell(\hat{k}\cdot\hat{k}') , \tag{3.3.35}$$

we find

$$C_\ell^{TT} = \frac{2}{\pi}\int k^2 \mathrm{d}k \underbrace{P_\zeta(k)}_{\text{Inflation}} \underbrace{\Delta_{T\ell}(k)\Delta_{T\ell}(k)}_{\text{Anisotropies}} . \tag{3.3.36}$$

This result shows how the primordial power spectrum $P_\zeta(k)$ gets processed into the observed CMB power spectrum C_ℓ^{TT}. To measure the primordial spectrum, C_ℓ^{TT} needs to be *deconvolved* by taking into account the appropriate transfer functions and projection effects, i.e. for a given background cosmology we can compute the evolution and projection effects in eq. (3.3.36) and therefore extract the inflationary initial conditions $P_\zeta(k)$. By this deconvolution procedure, the CMB provides a fascinating probe of the early universe.

[11] A derivation of the source terms and the projection factors is beyond the scope of this lecture, but may be found in Dodelson's book (see also the simplified treatment in the next section).

[12] http://camb.info/

3.3.3 CMBSimple*

High-energy theorists often consider CMBFast as a black-box that takes the primordial perturbations as an input and magically spits out the CMB power spectrum as an output. This need not be so. In this section, you will be guided to writing your own CMB code, let us call it CMBSimple.

Our code will implement the *two fluid approximation* of Seljak.[13] This approximation is based on the simple observation that before recombination photons and baryons (meaning electrons and protons) are coupled strongly to each other via Thomson scattering. We can therefore treat the photons and baryons a single fluid. Since the dark matter and the photon-baryon fluids

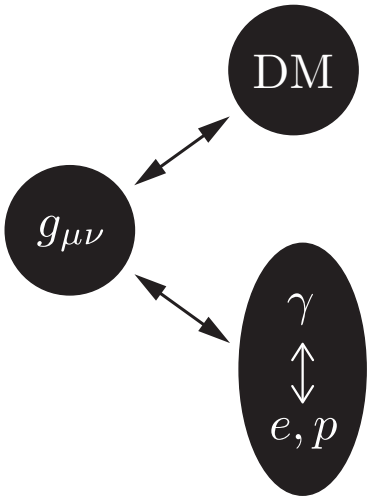


Figure 3.5: *Illustration of the two fluid approximation.*

don't couple directly to each other, their stress tensors are separately conserved,

$$\nabla_\mu T^{\mu\nu}_{(i)} = 0 \ , \qquad \text{for each fluid } i \ . \tag{3.3.37}$$

From this we get the fluid equations of motion. The dynamics of scalar perturbations is then governed by the continuity equation for density fluctuations $\delta = \delta\rho/\rho$ and the Euler equation for the divergence of the velocity field, $\theta = \vec{\nabla} \cdot \vec{v}$. For a single uncoupled fluid, with equation of state $w = p/\rho$ and sound speed $c_s^2 = \delta p/\delta\rho$, these are

$$\delta' = -(1+w)(\theta - 3\phi') - 3\mathcal{H}\left(c_s^2 - w\right)\delta \ , \tag{3.3.38}$$

$$\theta' = -\mathcal{H}(1-3w)\theta - \frac{w'}{1+w}\theta + \frac{c_s^2}{1+w}k^2\delta + k^2\phi \ , \tag{3.3.39}$$

where $\mathcal{H} = a'/a$ and ϕ is the gravitational potential, i.e. the metric perturbation in Newtonian gauge[14]

$$\mathrm{d}s^2 = a^2(\tau)\left[-(1+2\phi)\mathrm{d}\tau^2 + (1-2\phi)\mathrm{d}\vec{x}^2\right] \ . \tag{3.3.40}$$

We first determine the linear dynamics of a single Fourier mode $\phi_{\vec{k}}$ and then sum over all modes at the end. For any given perturbation mode with wavenumber $\vec{k} = k\hat{k}$, gravity sources

[13] Seljak (arXiv:astro-ph/9406050).

[14] Note that comoving gauge is not a good gauge inside the horizon.

only velocities in the density perturbations parallel to $\hat{k}$. We thus take $\vec{v}$ to be of the form $\vec{v} = -iv\hat{k}$, which implies

$$\theta = kv \ . \tag{3.3.41}$$

For cold dark matter, $w = c_s^2 = 0$, the fluid equations then become

$$\delta_c' = -kv_c + 3\phi' \ , \qquad v_c' = -\mathcal{H}v_c + k\phi \ . \tag{3.3.42}$$

The photons contribute a pressure $p_\gamma = \frac{1}{3}\rho_\gamma$ to the photon-baryon fluid. The effective equation of state and sound speed of the photon-baryon fluid therefore are

$$w = \frac{1}{3+4R} \ , \qquad c_s^2 = \frac{1}{3(1+R)} \ , \tag{3.3.43}$$

where

$$R = \frac{3}{4}\frac{\bar{\rho}_b}{\bar{\rho}_\gamma} \ . \tag{3.3.44}$$

In the *tight coupling approximation* we have $v_\gamma = v_b$ and $\delta_\gamma = \frac{4}{3}\delta_b$.[15] With this we get the photon evolution equations

$$\delta_\gamma' = -\tfrac{4}{3}kv_\gamma + 4\phi' \ , \qquad (1+R)v_\gamma' = -R\,v_\gamma + \tfrac{1}{4}k\delta_\gamma + (1+R)k\phi \ . \tag{3.3.45}$$

Exercise. Derive eq. (3.3.45) more formally following the treatment in Ma and Bertschinger:[16] i.e. expand the hierarchy of Boltzmann equations in the mean free path of the photons.

The fluid equations (3.3.42) and (3.3.45) have to be supplemented by the linearized Einstein equation for the gravitational potential ϕ,

$$k\left(\phi' + \mathcal{H}\phi\right) = 4\pi G a^2 \sum_i (\bar{\rho}_i + \bar{p}_i)v_i \ , \tag{3.3.46}$$

where the sum is over both fluids.

Finally, we have the Friedmann equation for the background

$$\mathcal{H}^2 = \left(\frac{a'}{a}\right)^2 = \frac{8\pi G a^2}{3}\sum_i \bar{\rho}_i \ . \tag{3.3.47}$$

Exercise. Show that eq. (3.3.47) has the following analytic solution

$$y \equiv \frac{a}{a_{\rm eq}} = (\alpha x)^2 + 2\alpha x \ , \qquad x \equiv \frac{\tau}{\tau_r} \ , \qquad \alpha^2 \equiv \frac{a_{\rm rec}}{a_{\rm eq}} \ , \tag{3.3.48}$$

where , $a_{\rm rec}^{-1} \approx 1100$, $a_{\rm eq}^{-1} \approx 2.4 \times 10^4\,\Omega_m h^2$, and τ_r is the would-be conformal time at recombination if the universe had always been matter-dominated after the Big Bang. Show that

$$\tau_r \equiv \left(\frac{4a_{\rm rec}}{\Omega_m H_0^2}\right)^{1/2} \ . \tag{3.3.49}$$

---

[15] Recall that $\delta n_\gamma = \delta n_b$, while $n_\gamma \propto T^3$ and $\rho_\gamma \propto T^4$.

[16] Ma and Bertschinger (arXiv:astro-ph/9506072).

For numerical convenience we rescale time, $x \equiv \tau/\tau_r$, and momentum, $\kappa \equiv k\tau_r$. From now on primes will always refer to derivatives with respect to x. We also define

$$\eta \equiv \frac{a'}{a} = \frac{1}{a}\frac{da}{dx} = \frac{2\alpha(\alpha x+1)}{(\alpha x)^2+2\alpha x} \equiv \tau_r \mathcal{H} \ . \tag{3.3.50}$$

The fluid equations of motion then become

$$\delta_c' = -\kappa v_c + 3\phi' \ , \tag{3.3.51}$$
$$v_c' = -\eta v_c + \kappa\phi \ , \tag{3.3.52}$$
$$\delta_\gamma' = -\tfrac{4}{3}\kappa v_\gamma + 4\phi' \ , \tag{3.3.53}$$
$$v_\gamma' = (1+\tfrac{3}{4}y_b)^{-1}(-\tfrac{3}{4}y_b\eta\, v_\gamma + \tfrac{1}{4}\kappa\delta_\gamma) + \kappa\phi \ , \tag{3.3.54}$$

with

$$\phi' = -\eta\phi + \frac{3\eta^2}{2\kappa}\,\frac{v_\gamma(\frac{4}{3}+y-y_c)+v_c y_c}{1+y} \ . \tag{3.3.55}$$

Here, we defined $y_{b,c} \equiv y\frac{\Omega_{b,c}}{\Omega_m}$. The task now is to solve these equations for ϕ, δ_γ, δ_c, v_γ, v_c, subject to the initial conditions at $x = x_i$,

$$\phi = 1 \ , \tag{3.3.56}$$
$$\delta_\gamma = -2\phi \ , \tag{3.3.57}$$
$$\delta_c = \tfrac{3}{4}\delta_\gamma \ , \tag{3.3.58}$$
$$v_\gamma = -\tfrac{1}{4}\tfrac{\kappa}{\eta}\delta_\gamma \ , \tag{3.3.59}$$
$$v_c = v_\gamma \ . \tag{3.3.60}$$

By setting the initial gravitational potential $\phi(x_i) \equiv 1$ we have extracted the primordial curvature perturbation ζ^o from the transfer function, cf. eq. (3.3.78).

Exercise. Write a short Mathematica notebook to solve this system of equations from $x_i = 10^{-3}$ to $x_{\rm rec} = (\sqrt{(\alpha^2+1)}-1)/\alpha$ (use NDSolve). Use $\omega_m \equiv \Omega_m h^2 = 0.13$ and $\omega_b \equiv \Omega_b h^2 = 0.02$. Evaluate the solutions at $x_{\rm rec}$ and tabulate $[\phi + \frac{1}{4}\delta_\gamma](x_{\rm rec})$ and $v_\gamma(x_{\rm rec})$ between $\kappa_{\rm min} = 0.01$ and $\kappa_{\rm max} = 300$.

Assuming *instantaneous recombination*, we have the following relation between these perturbations and the observed CMB anisotropies,

$$\Theta(k,\vec{n}) = [\phi + \tfrac{1}{4}\delta_\gamma](\tau_{\rm rec}) + \vec{n}\cdot\vec{v}_\gamma(\tau_{\rm rec}) + 2\int_{\tau_{\rm rec}}^{\tau_0}\phi'(\tau){\rm d}\tau \ , \tag{3.3.61}$$

where the quantities on the r.h.s. are implicitly functions of the wavenumber k.

Explaining this formula in detail would really take us too far off the main track of these lectures. We therefore contend ourselves with a few comments and refer the interested reader to Dodelson's excellent book for more details and a derivation of eq. (3.3.61): The first two terms are evaluated at recombination. This reflects the fact that the CMB is a snapshot of the last-scattering surface. The first (monopole) term includes a term associated with the intrinsic fluctuation in the photon density δ_γ and a gravitational redshift contribution arising from the Newtonian potential ϕ. Combined these terms lead to the famous Sachs-Wolfe (SW) effect. The second (dipole) term describes a Doppler shift. Besides being a source from temperature

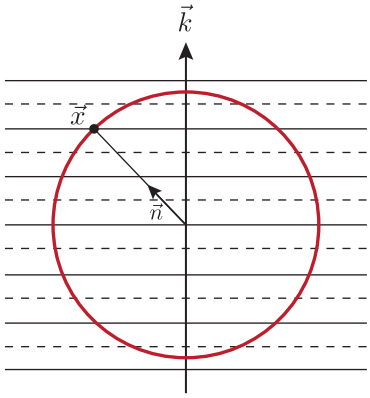


Figure 3.6: *Projection of a plane wave onto the surface of last scattering.*

anisotropies, it is important for the generation of CMB polarization (see below) and the correlation between polarization and temperature anisotropies. The last term leads to the Integrated Sachs-Wolfe (ISW) effect. It doesn't receive any contributions from the matter era, when $\phi' = 0$, but only from residual radiation at early times and dark energy at late times.

Equipped with the solution for a single Fourier mode, we can now sum over all modes, weigthed by the initial conditions $\zeta_{\vec{k}}$, and then compute the temperature anisotropies as a function of direction on the sky $\vec{n}$. To do this, we need to relate $\Theta(\vec{n})$ in eq. (3.3.30) to the Fourier space solution $\Theta(k)$ in eq. (3.3.61). First, we note that our assumption of instantaneous recombination implies

$$\Theta(\vec{n}) = \int \mathrm{d}r\, \Theta(\vec{x})\, \delta(r - r_\star) \ , \tag{3.3.62}$$

where the integral is over conformal distance and $r_\star$ is the angular diameter distance between us and the last-scattering surface. Here, $\Theta(\vec{x})$ is the real space temperature field, related to our solution $\Theta(k)$ via

$$\Theta(\vec{x}) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\, e^{i\vec{k}\cdot\vec{x}}\, \zeta_{\vec{k}}\, \Theta(k) \ . \tag{3.3.63}$$

The weighting by the initial condition $\zeta_{\vec{k}}$ was introduced because our solution $\Theta(k)$ was found for $\zeta_{\vec{k}} = 1$ (see eq. (3.3.56)). Hence, we get

$$\Theta(\vec{n}) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\, e^{i\vec{k}\cdot r_\star\vec{n}}\, \zeta_{\vec{k}}\, \Theta(k) \ . \tag{3.3.64}$$

Using the identity

$$e^{i\vec{k}\cdot r_\star\vec{n}} = 4\pi \sum_{lm} i^\ell j_\ell(kr_\star) Y^*_{\ell m}(\vec{k}) Y_{\ell m}(\vec{n}) \ , \tag{3.3.65}$$

we find that the CMB angular power spectrum takes the form of eq. (3.3.36),

$$C_\ell^{TT} = \frac{2}{\pi} \int k^2 \mathrm{d}k\, P_\zeta(k) \Delta^2_{T\ell}(k) \ , \tag{3.3.66}$$

with

$$\Delta_{T\ell}(k) = (\phi + \tfrac{1}{4}\delta_\gamma)\, j_\ell(k[\tau_0 - \tau_{\rm rec}]) + v_\gamma\, j'_\ell(k[\tau_0 - \tau_{\rm rec}]) + 2\int_{\tau_{\rm rec}}^{\tau_0} \mathrm{d}\tau\, j_\ell(k[\tau_0 - \tau]) \frac{\phi'(\tau)}{\phi'(\tau_0)} \ . \tag{3.3.67}$$

Here, the spherical Bessel functions j_ℓ are coming form the projection of the plane waves onto the last-scattering surface. The argument of the Bessel functions is related to the angular diameter distance between us and the last-scattering surface, which in flat space is $r_\star = \tau_0 - \tau_{\rm rec}$.

By assuming perfect tight coupling (mean free path = zero) and instantaneous recombination, our solution is still missing some important physics. Including the effects of a finite mean free path for the photons and a finite duration of recombination would lead to the damping of small scale fluctuations[17] (the former effect is sometimes called Silk damping). We could rectify this by adding viscosity directly in the fluid equations of motion.[18] Here, we follow Seljak and take a simpler, more phenomenological approach. Haters would call it a fudge. We introduce a high momentum cutoff in the integral (3.3.36), i.e. we define

$$C_\ell^{TT} \propto \int k^2 {\rm d}k\, P_\zeta(k) D(k) \Delta_{T\ell}^2(k) \ , \tag{3.3.68}$$

where

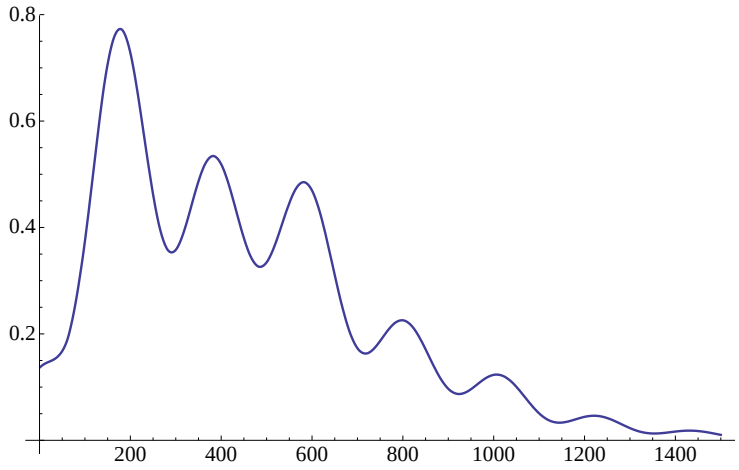
$$D(k) = e^{-(k/k_D)^2} \ . \tag{3.3.69}$$

The damping scale k_D can be related to Ω_m and Ω_b (see the exercise below). See Seljak's paper for more details.

Exercise. Ignore the ISW term in eq. (3.3.67). For the Silk damping scale in (3.3.69) use

$$\kappa_D^{-2} = 2x_s^2 + \sigma^2 x_{\rm rec}^2 \ , \tag{3.3.70}$$

with $x_s \equiv 0.6\, \omega_m^{1/4} \omega_b^{-1/2} a_{\rm rec}^{3/4}$ and $\sigma \approx 0.03$. Then use the tabulated solutions from the previous exercise to evaluate the integral (3.3.68) (use NIntegrate). Ask Mathematica to find an interpolating solution (use Interpolation). Plot $\ell(\ell+1)C_\ell^{TT}$. The result should look like this:



---

[17] Essentially, fluctuations on scales smaller than the mean free path are damped; just like you can't have sound waves with wavelengths smaller than the mean free path in air.

[18] See Mukhanov's book.

3.3.4 Coherent Phases and Superhorizon Fluctuations*

Let me digress briefly to describe arguably the most dramatic observational evidence that something like inflation must have occurred in the early universe. The following is a trivialization of arguments that have been explained beautifully in an article by Dodelson.[19]

The Peaks of the TT Spectrum

As we have shown in the previous chapter, inflation produces a nearly scale-invariant spectrum of perturbations, i.e. a particular Fourier mode is drawn from a distribution with zero mean and variance given by

$$\langle \zeta_{\vec{k}} \zeta_{\vec{k}'} \rangle = (2\pi)^3 \delta(\vec{k} + \vec{k}') P_\zeta(k) \,, \tag{3.3.71}$$

where $k^3 P_\zeta(k) \propto k^{n_s - 1}$ with $n_s \approx 1$. You might think then that the shape of the power spectrum can be measured in observations, and this is what convinces us that inflation is right. Of course, it is true that we can measure the power spectrum, both of the matter and of the radiation, and it is true that the observations agree with the theory. But, according to Dodelson, "this is not what tingles our spines when we look at the data". Rather, the truly striking aspect of perturbations generated during inflation is that *all Fourier modes have the same phase.*

Consider a Fourier mode of ζ with wavenumber k. In §3.2, we proved that $\zeta_{\vec{k}}$ is conserved outside the horizon, $k < aH$. Since the fluctuation amplitude was constant outside the horizon, $\dot{\zeta} \approx 0$ at horizon re-entry. If we think of each Fourier mode as a linear combination of a sine and a cosine mode, then inflation excited only the cosine modes (defining horizon re-entry as $t \equiv 0$).

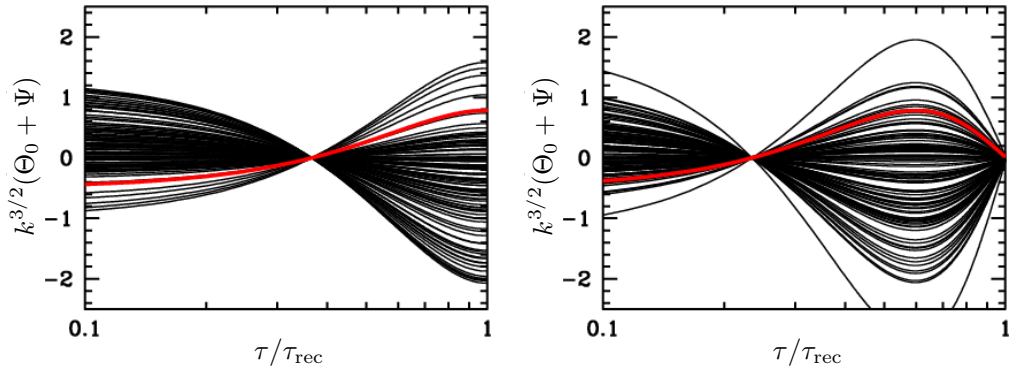


Figure 3.7: *Evolution of an infinite number of modes all with the same wavelength. Recombination is at $\tau = \tau_{\rm rec}$. (Left) Wavelength corresponding to the first peak in the CMB angular power spectrum. (Right) Wavelength corresponding to the first trough. Although the amplitudes of all these different modes differ from one another, since they start with the same phase, the ones on the left all reach maximum amplitude at recombination, the ones on the right all go to zero at recombination. This leads to the acoustic peaks of the CMB power spectrum.*

Once inside the horizon, the curvature perturbation ζ sources density fluctuations δ which evolve under the influence of gravity and pressure. In the previous section, we described this process in Newtonian gauge, where density fluctuations are sourced by the Newtonian potential ϕ. Inside the horizon we can think of ϕ and ζ interchangably. Combining the two photon

[19] Dodelson, *Coherent Phase Argument for Inflation* (arXiv:hep-ph/0309057).

evolution equations in (3.3.45), we find

$$\delta_\gamma'' + \frac{R'}{1+R}\delta_\gamma' + c_s^2 k^2 \delta = F_g[\phi]\,, \tag{3.3.72}$$

where c_s is the sound speed of the photon-baryon fluid and F_g is the gravitational source term

$$F_g = 4\left[\phi'' + \frac{R'}{1+R}\phi' - \frac{1}{3}k^2\phi\right]\,. \tag{3.3.73}$$

We see that the photon-baryon fluid can sustain acoustic oscillations, where the inertia is provided by the baryons, while the pressure is provided by the photons. Imagine that recombination happens instantaneously (as we saw in the previous section, this is not a terrible approximation). At last scattering, fluctuations with different wavelengths are then captured at different phases in their oscillations. Modes of a certain wavelength are captured at maximum or minimum amplitude, while others would be captured at zero amplitude. If all Fourier modes of a given wavelength have the same phases they interfere coherently (see fig. 3.7) and the spectrum of all Fourier produces a series of peaks and troughs in the CMB power spectrum as seen on the last-scattering surface. This is of course what we see in the data (see fig. 3.3). However, in order for the theory of initial fluctuations to explain this it needs to involve a mechanism that produces coherent initial phases for all Fourier modes. Inflation does precisely that! Because fluctuations freeze when they exit the horizon, the phases for the Fourier modes were set well before the modes of interest entered the horizon. When were are admiring the peak structure of the CMB power spectrum we are really admiring the ability of the primordial mechanism for generating fluctuations to coordinate the phases of all Fourier modes. Without this coherence, the CMB power spectrum would simply be white noise (see fig. 3.8) with no peaks and troughs (in fact, this is precisely why cosmic strings or topological defects are ruled out as the primary sources for the primordial fluctuations.).

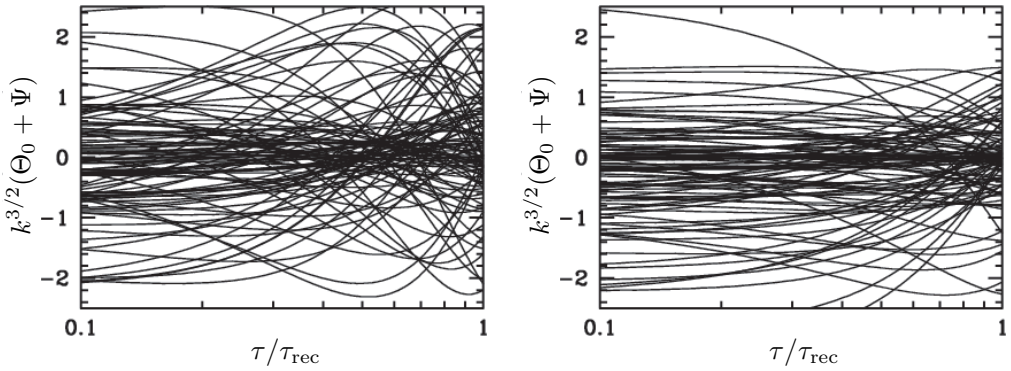


Figure 3.8: *Modes corresponding to the same two wavelengths as in fig. 3.7, but this time with random initial phases. The anisotropies at the angular scales corresponding to these wavelengths would have identical rms's if the phases were random, i.e. the angular peak structure of the CMB would be washed away.*

The Peaks of the TE Spectrum

The skeptic might not be convinced by the above argument. The peaks and troughs of the CMB temperature fluctuation spectrum are at $\ell > 200$, corresponding to angular scales $\theta < 1^\circ$. All

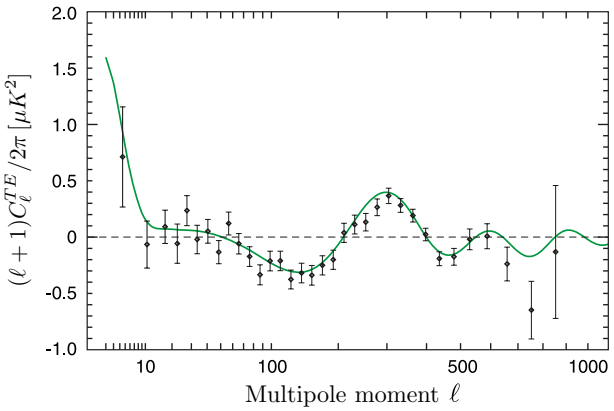


Figure 3.9: *Power spectrum of the cross-correlation between temperature and E-mode polarization anisotropies. The anti-correlation for $\ell = 50 - 200$ (corresponding to angular separations $5^\circ > \theta > 1^\circ$) is a distinctive signature of adiabatic fluctuations on superhorizon scales at the epoch of decoupling, confirming a fundamental prediction of the inflationary paradigm.*

of these scales were within the horizon at the time of recombination. Hence, it is in principle possible (and people have tried in the 90s) to engineer a theory of structure formation which obeys causality and still manages to produce only the 'cosine mode'. Such a theory would explain the CMB peaks without appealing to inflation. This doesn't sound like the most elegant thing in the world, but it can't be excluded as a logical possibility.

However, we now show that even these highly-tuned alternatives to inflation can be ruled out by considering CMB polarization. Looking at fig. 3.9 we see that the cross-correlation between CMB temperature fluctuations and the E-mode polarization has a negative peak around $100 < \ell < 200$. This anti-correlation signal is also the result of phase coherence, but now the scales involved were *not* within the horizon at recombination. Hence, there is *no* causal mechanism (after $\tau = 0$) that could have produced this signal. One is almost forced to consider something like inflation with its shrinking comoving horizon leading to horizon exit and re-entry.[20] As Dodelson explains

> At recombination, [the phase difference between the monopole (sourcing T) and the dipole (sourcing E) of the density field] causes the product of the two to be negative for $100 < \ell < 200$ and positive on smaller scales until $\ell \sim 400$. But this is precisely what WMAP has observed! We have clear evidence that the monopole and the dipole were out of phase with each other at recombination.
>
> This evidence is exciting for the small scale modes ($\ell > 200$). Just as the acoustic peaks bear testimony to coherent phases, the cross-correlation of polarization and temperatures speaks to the coherence of the dipole as well. It solidifies our picture of the plasma at recombination. The evidence from the larger scale modes ($\ell < 200$) though is positively stupendous. For, these modes were not within the horizon at recombination. So the *only* way they could have their phases aligned is if some

[20] It should be mentioned here that there are two ways to get a shrinking comoving Hubble radius, $1/(aH)$. During inflation H is nearly constant and the scale factor a grows exponentially. However, in a *contracting* spacetime a shrinking horizon can be achieved if H grows with time. This is the mechanism employed by ekpyrotic/cyclic cosmology. When viewed in terms of the evolution of the comoving Hubble scale inflation and ekpyrosis appear very similar, but there are important differences, e.g. in ekyprosis it is a challenge to match the contracting phase to our conventional Big Bang expansion.

primordial mechanism did the job, when they were in causal contact. Inflation is just such a mechanism.

3.3.5 CMB Polarization

In addition to anisotropies in the CMB temperature, we expect the CMB to become polarized via Thomson scattering of the photons from free electrons just before decoupling (see fig. 3.10). As we explain in this section, this polarization contains crucial information about the primordial fluctuations and hence about inflation.

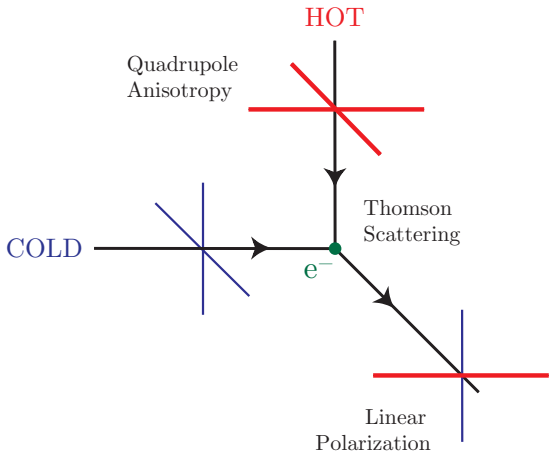


Figure 3.10: *Thomson scattering of radiation with a quadrupole anisotropy generates linear polarization. If a free electron 'sees' an incident radiation pattern that is isotropic, then the outgoing radiation remains unpolarized because orthogonal polarization directions cancel out. However, if the incoming radiation field has a quadrupole component, a net linear polarization is generated.*

The E/B Decomposition

The spin-1 polarization field can be decomposed spin-0 quantities, the so-called E- and B-modes.

Characterization of the radiation field.

The mathematical characterization of CMB polarization anisotropies is slightly more involved than that the description of temperature fluctuations because polarization is not a scalar field so the standard expansion in terms of spherical harmonics is not applicable.

The anisotropy field is defined in terms of a 2×2 intensity tensor $I_{ij}(\hat{n})$, where as before $\hat{n}$ denotes the direction on the sky. The components of I_{ij} are defined relative to two orthogonal basis vectors $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ perpendicular to $\hat{n}$. Linear polarization is then described by the Stokes parameters $Q = \frac{1}{4}(I_{11} - I_{22})$ and $U = \frac{1}{2}I_{12}$, while the temperature anisotropy is $T = \frac{1}{4}(I_{11} + I_{22})$. The polarization magnitude and angle are $P = \sqrt{Q^2 + U^2}$ and $\alpha = \frac{1}{2}\tan^{-1}(U/Q)$. The quantity T is invariant under a rotation in the plane perpendicular to $\hat{n}$ and hence may be expanded in terms of scalar (spin-0) spherical harmonics (3.3.30). The quantities Q and U, however, transform under rotation by an angle ψ as a spin-2 field $(Q \pm iU)(\hat{n}) \to e^{\mp 2i\psi}(Q \pm iU)(\hat{n})$. The harmonic analysis of $Q \pm iU$ therefore requires expansion on the sphere in terms of tensor (spin-2) spherical harmonics

$$(Q \pm iU)(\hat{n}) = \sum_{\ell,m} a_{\pm 2,\ell m}\, {}_{\pm 2}Y_{\ell m}(\hat{n}) \ . \tag{3.3.74}$$

For a description of the mathematical properties of these tensor spherical harmonics, ${}_{\pm 2}Y_{\ell m}$, the reader

should consult the original paper by Kamionkowski et al.[21] Instead of the moments $a_{\pm 2,\ell m}$ it is convenient to introduce the linear combinations

$$a_{E,\ell m} \equiv -\frac{1}{2}\left(a_{2,\ell m} + a_{-2,\ell m}\right) , \qquad a_{B,\ell m} \equiv -\frac{1}{2i}\left(a_{2,\ell m} - a_{-2,\ell m}\right) . \tag{3.3.75}$$

Then one can define two scalar (spin-0) fields instead of the spin-2 quantities Q and U

$$E(\hat{n}) = \sum_{\ell,m} a_{E,\ell m}\, Y_{\ell m}(\hat{n}) , \qquad B(\hat{n}) = \sum_{\ell,m} a_{B,\ell m}\, Y_{\ell m}(\hat{n}) . \tag{3.3.76}$$

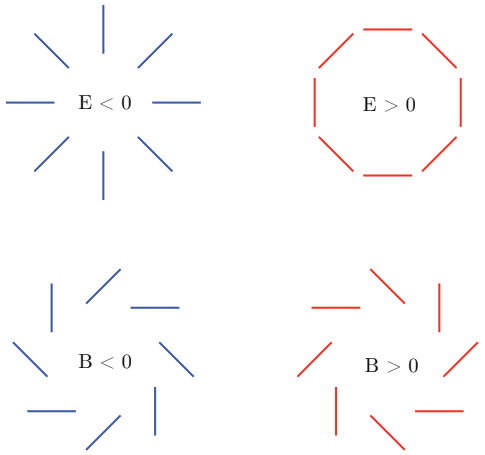


Figure 3.11: *Examples of E-mode and B-mode patterns of polarization. Note that if reflected across a line going through the center the E-mode patterns are unchanged, while the positive and negative B-mode patterns get interchanged.*

The scalar quantities E and B completely specify the linear polarization field. E-mode polarization is *curl-free* with polarization vectors that are radial around cold spots and tangential around hot spots on the sky (see fig. 3.11). In contrast, B-mode polarization is *divergence-free* but has a *curl*: its polarization vectors have vorticity around any given point on the sky.[22] The symmetries of temperature and polarization anisotropies allow four types of correlations: the autocorrelations of temperature fluctuations and of E- and B-modes denoted by TT, EE, and BB, respectively, as well as the cross-correlation between temperature fluctuations and E-modes: TE. All other correlations (TB and EB) vanish for symmetry reasons. Only the TE and EE correlations have been detected so far.

The angular power spectra are defined as before

$$C_\ell^{XY} \equiv \frac{1}{2\ell+1}\sum_m \langle a^*_{X,\ell m} a_{Y,\ell m}\rangle , \qquad X, Y = T, E, B , \tag{3.3.77}$$

where now both scalars ζ and tensors h can act as a source

$$a_{X,\ell m} = 4\pi(-i)^\ell \int \frac{\mathrm{d}^3k}{(2\pi)^3}\, \Delta_{X\ell}(k)\, \{\zeta_{\vec{k}}, h_{\vec{k}}\}\, Y_{\ell m}(\hat{k}) . \tag{3.3.78}$$

[21] Kamionkowski et al. (astro-ph/9611125).

[22] Evidently the E and B nomenclature reflects the properties familiar from electrostatics, $\nabla \times \mathbf{E} = 0$ and $\nabla \cdot \mathbf{B} = 0$.

Note that there will be distinct transfer functions $\Delta_{X\ell}(k)$ for temperature fluctuations and E- and B-mode polarization. In particular, the transfer functions for polarization will be more involved and can't be computed in a simple tight coupling approximation. The full Boltzmann machinery will be needed.

A Smocking Gun

The cosmological significance of the E/B decomposition of CMB polarization is related to the following remarkable facts:

i) scalar (density) perturbations create only E-modes and *no* B-modes.

ii) vector (vorticity) perturbations create mainly B-modes.[23]

iii) tensor (gravitational wave) perturbations create both E-modes and B-modes.

The angular power spectrum of CMB B-modes is related to the primordial tensor power spectrum $P_h(k)$ as follows

$$C_\ell^{BB} = (4\pi)^2 \int k^2 \mathrm{d}k \underbrace{P_h(k)}_{\text{Inflation}} \Delta^2_{B\ell}(k)\,, \tag{3.3.79}$$

where $\Delta_{B\ell}(k)$ is the transfer function for B-modes.

The fact that *scalars do not produce B-modes* while *tensors do* is the basis of the famous statement that a detection of B-modes is a smoking gun of tensor modes, and therefore of inflation.

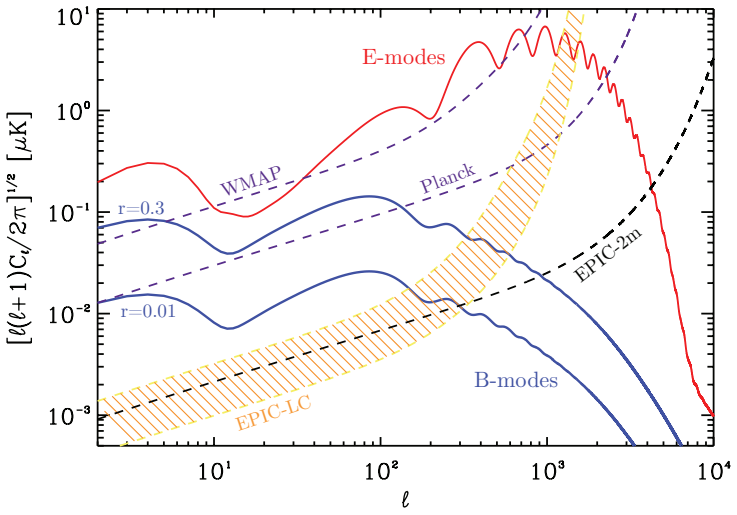


Figure 3.12: *E- and B-mode power spectra for a tensor-to-scalar ratio saturating current bounds, $r = 0.3$, and for $r = 0.01$. Shown are also the experimental sensitivities for WMAP, Planck and two different realizations of a future CMB satellite (CMBPol) (EPIC-LC and EPIC-2m).*

What precisely would we learn from a B-mode detection?

[23] However, vectors decay with the expansion of the universe and are therefore believed to be subdominant at recombination.

1. *"proof" that inflation occurred*

   No other early universe mechanism produces a stochastic background of tensor fluctuations that span superhorizon scales at recombination.

2. *energy scale of inflation*

   The tensor-to-scalar ratio is a direct measure of the energy scale of inflation

   $$V^{1/4} \sim \left(\frac{r}{0.01}\right)^{1/4} 10^{16}\,\text{GeV}\ . \tag{3.3.80}$$

   Large values of the tensor-to-scalar ratio, $r > 0.01$, correspond to inflation occurring at GUT scale energies.

3. *super-Planckian field-variation*

   We showed in the previous chapter that the tensor-to-scalar ratio relates to the field evolution during inflation,

   $$\frac{\Delta\phi}{M_{\rm pl}} = \mathcal{O}(1) \times \left(\frac{r}{0.01}\right)^{1/2}\ . \tag{3.3.81}$$

   Observable tensors, $r > 0.01$, therefore imply super-Planckian field vevs. One of the key challenges for string inflation is to make sense of such large-field models of inflation.

Primordial tensor modes have of course not yet been detected, but the hunt is on. A number of CMB experiments are now going after the elusive gravitational waves predicted by inflation. A detection could be imminent.

3.3.6 Non-Gaussianity

The primordial fluctuations are Gaussian to a high degree. However, as we describe in detail in Chapter 5, even a small amount of non-Gaussianity would encode a tremendous amount of information about the physics of inflation. The primary diagnostic of non-Gaussian statistics is the three-point function of inflationary fluctuations. In momentum space, the three-point correlation function (or bispectrum) is

$$\langle \zeta_{\mathbf{k}_1} \zeta_{\mathbf{k}_2} \zeta_{\mathbf{k}_3} \rangle = (2\pi)^3\, \delta(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3)\, f_{\rm NL}\, B(k_1, k_2, k_3)\,. \tag{3.3.82}$$

Here, $f_{\rm NL}$ is a dimensionless parameter defining the amplitude of non-Gaussianity, while the function $B(k_1, k_2, k_3)$ captures the momentum dependence. The amplitude and sign of $f_{\rm NL}$, as well as the shape and scale dependence of $B(k_1, k_2, k_3)$, depend on the details of the interaction generating the non-Gaussianity, making the three-point function a powerful discriminating tool for probing models of the early universe.

Current CMB data imply $|f_{\rm NL}| < \mathcal{O}(100)$ (with the precise value depending on the shape of non-Gaussianity that is tested for). To appreciate the agree of Gaussianity that this implies, we should notice that the better measure of non-Gaussianity is

$$\text{NG} \ \sim\ f_{\rm NL}\zeta \ \sim\ \mathcal{O}(0.1\%)\ . \tag{3.3.83}$$

Constraints on Gaussianity are therefore stronger than our constraints on the flatness of the universe. Nevertheless, many popular inflationary model suggest the possibility of non-Gaussian signals an order of magnitude below the current bounds. This important regime will soon be probed by the Planck satellite. Something else to look forward to in the near future.

3.4 From Vacuum Fluctuations to Large-Scale Structure

Large-scale structure observations are becoming a more and more important complementary probe of the early universe. While the CMB studies the largest scales (and hence earliest horizon exits during inflation), large-scale structure probes smaller scales (and later horizon exits).

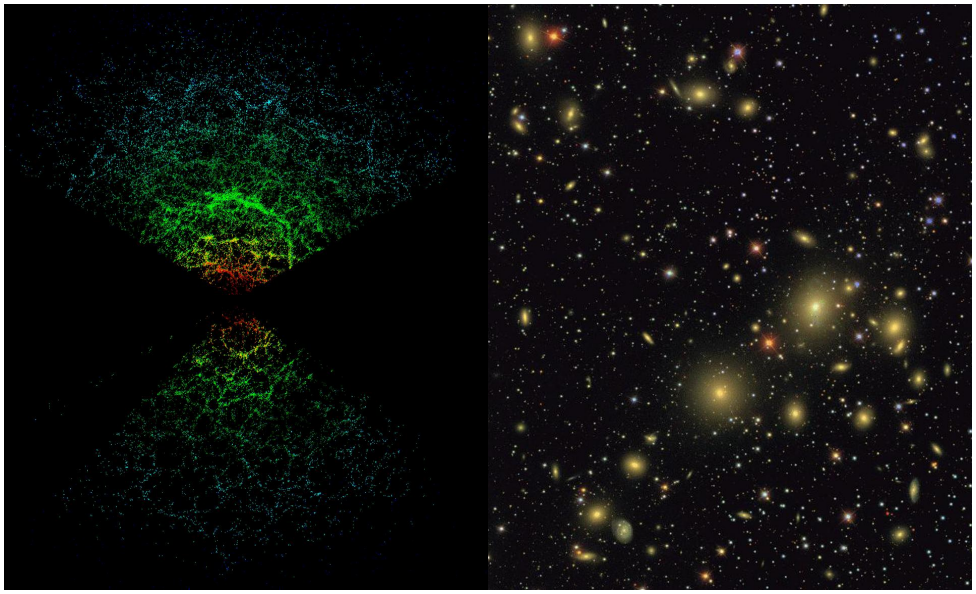


Figure 3.13: *Distribution of galaxies. The Sloan Digital Sky Survey (SDSS) has measured the positions and distances (redshifts) of nearly a million galaxies. Galaxies first identified on 2d images, like the one shown above on the right, have their distances measured to create the 3d map. The left image shows a slice of such a 3d map.*

3.4.1 Dark Matter Transfer Function

We derived the dark matter evolution equations in §3.3.3, cf. eq. (3.3.42). Together with the Einstein equation for the gravitational potential $\phi(\tau)$ (with initial conditions set up by inflation), we can therefore compute the evolution of the dark matter density. We allow us to relate the dark matter power spectrum to the power spectrum of primordial curvature perturbations,

$$P_\delta(k,z) \quad \Leftrightarrow \quad P_\zeta(k) \ , \tag{3.4.84}$$

where we have indicated that the observed dark matter power is a function of redshift z. This leads to the so-called dark matter transfer function $T_\delta(k,\tau)$, which is conventionally defined as follows

$$P_\delta(k,\tau) = \frac{4}{25}\left(\frac{k}{aH}\right)^4 T_\delta^2(k,\tau) P_\zeta(k) \ . \tag{3.4.85}$$

The numerical factor and the k-scaling that have been factored out from the transfer function are conventional. They can be rationalized if we recall that subhorizon modes during the matter era satisfy the following Poisson equation:

$$\delta_{\mathbf{k}}(\tau) = -\frac{2}{3}\frac{k^2}{(aH)^2}\,\phi_{\mathbf{k}} = -\frac{2}{5}\frac{k^2}{(aH)^2}\,\zeta_{\mathbf{k}} \ . \tag{3.4.86}$$

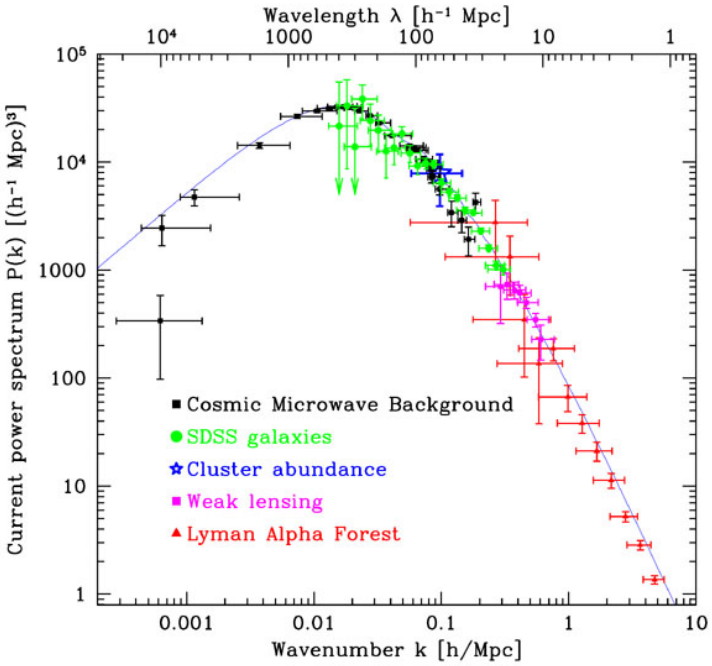


Figure 3.14: *The matter power spectrum. The break in the spectrum around $k_{\rm eq}$ reflects the change in the growth of density perturbations before and after matter-radiation equality, as well as the decay of the gravitational potential for modes that enter the horizon during radiation domination. For scale-invariant initial conditions, $k^3 P_\zeta(k) = const.$, the quantity $k^{-1} P_\delta(k)$ would be a constant for $k < k_{\rm eq}$ (i.e. for modes than didn't enter the horizon before matter-radiation equality).*

Let us sketch the physics of the transfer function: the qualitative shape of the matter power spectrum (see fig. 3.14) is easily understood if we recall some basic facts about the evolution of fluctuations after they enter the horizon. Density fluctuations evolve under the competing influence of pressure and gravity. During radiation domination the large radiation pressure prevents the rapid growth of fluctuations; the density contrast only grows logarithmically, $\delta_c \sim \ln a$. Moreover, the gravitational potential ϕ decays inside the horizon during radiation domination.[24]

[24] If the universe is dominated by a fluid with equation of state w and sound speed $c_s^2 = w$, the evolution equation for the gravitational potential can be written as

$$\phi_{\mathbf{k}}'' + \frac{6(1+w)}{1+3w}\frac{1}{\tau}\phi_{\mathbf{k}}' + wk^2\phi_{\rm k} = 0 \ . \tag{3.4.87}$$

This has the following exact solution

$$\phi_{\mathbf{k}}(\eta) = y^\alpha \left[C_1(k) J_\alpha(y) + C_2(k) Y_\alpha(y)\right] , \qquad y \equiv \sqrt{w}k\tau , \quad \alpha \equiv \frac{1}{2}\left(\frac{5+3w}{1+3w}\right) , \tag{3.4.88}$$

where J_α and Y_α are Bessel functions of order α. During the matter-dominated era, $w = 0$, this becomes

$$\phi_{\mathbf{k}}(\tau) = C_1(k) + \frac{C_2(k)}{y^5} , \tag{3.4.89}$$

whereas during the radiation-dominated era, $w = \frac{1}{3}$, we find

$$\phi_{\mathbf{k}}(\tau) = \frac{1}{y^2}\left[C_1(k)\Big(\frac{\sin y}{y} - \cos y\Big) + C_2(k)\Big(\frac{\cos y}{y} + \sin y\Big)\right] . \tag{3.4.90}$$

In both cases the decaying mode may be dropped by setting $C_2(k) \equiv 0$. For a radiation-dominated background, the Newtonian potential is time-independent on superhorizon scales, $\lim_{k\tau\to 0}\phi_{\mathbf{k}}(\tau) = C_1(k)$, but decays on

Modes that enter the horizon during the radiation era will therefore be suppressed. This explains the suppression of the matter power spectrum for $k > k_{\rm eq}$ (see fig. 3.14). In contrast, during matter domination the background pressure is negligible and gravitational collapse operates more effectively, $\delta_c \sim a$. Importantly, the gravitational potential is constant on all scales during the matter era. Under the simplifying assumption that there is no significant growth of density perturbations between the time of horizon entry and matter domination one therefore arrives at the following approximate transfer function

$$T_\delta(k) \approx \begin{cases} 1 & k < k_{\rm eq} \\ (k_{\rm eq}/k)^2 & k > k_{\rm eq} \end{cases} . \tag{3.4.92}$$

Although eq. (3.4.92) is intuitively appealing for understanding the qualitative shape of the spectrum (i.e. the break in the spectrum at $k \approx k_{\rm eq}$), it is not accurate enough for most quantiative applications. Exact transfer functions can of course be computed numerically with CMBFast or CAMB. A famous fitting function for the matter transfer function was given by Bardeen, Bond, Kaiser and Szalay (BBKS)[25]

$$T_\delta(q) = \frac{\ln(1+2.34q)}{2.34q}\left[1+3.89q+(1.61q)^2+(5.46q)^3+(6.71q)^4\right]^{-1/4} , \tag{3.4.93}$$

where $q = k/\Gamma h\,\mathrm{Mpc}^{-1}$ and we defined the shape parameter

$$\Gamma \equiv \Omega_m h \exp(-\Omega_b - \sqrt{2h}\Omega_b/\Omega_m) . \tag{3.4.94}$$

More accurate transfer functions may be found in a paper by Eisenstein and Hu.[26] For our purposes it is only important to note that (give the background cosmological parameters) the dark matter transfer function can be computed and used to relate the dark matter power spectrum $P_\delta(k,z)$ to the inflationary spectrum $P_\zeta(k)$.

3.4.2 Galaxy Biasing

With the exception of gravitational lensing, we unfortunately never observe the dark matter directly. What we observe (e.g. in galaxy surveys like the Sloan Digital Sky Survey (SDSS)) is luminous or baryonic matter. Let us call the density contrast for galaxies δ_g. On large scales the following phenomenological ansatz for relating the galaxy distribution and the dark matter has proven useful[27]

$$\delta_g = b\,\delta_c \qquad \text{or} \qquad P_{\delta_g} = b^2 P_\delta . \tag{3.4.95}$$

Here, b is the so-called the (linear) bias parameter. It may be viewed as a parameter describing the ill-understood physics of galaxy formation. The bias parameter b can be obtained by measuring the galaxy bispectrum B_{δ_g}.

subhorizon scales. In contrast, during matter-domination the growing mode linear gravitational potential is time-independent on all scales, with a spatial profile $\phi_{\mathbf{k}}$ given by the Poisson equation

$$-k^2\phi_{\mathbf{k}} = \frac{3}{2}\mathcal{H}^2\delta_{\mathbf{k}} + 3\mathcal{H}^2\phi_{\mathbf{k}} . \tag{3.4.91}$$

[25] Bardeen et al. (ApJ, 304, 15-61,1986).
[26] Eisenstein and Hu (astro-ph/9709112).
[27] This may be 'derived' via the Press-Schechter formalism.

Modulo these complications the galaxy power spectrum $P_{\delta_g}(k)$ is an additional probe of inflationary scalar fluctuations $P_\zeta(k)$. As it probes smaller scales it is complementary to observations of the CMB. Finally, in the presence of primordial non-Gaussianity, the galaxies bias can develop an interesting scale-dependence $b(k)$ on large scales. We will discuss this effect and its relevance for early universe cosmology in Chapter 5.

3.5 Future Prospects

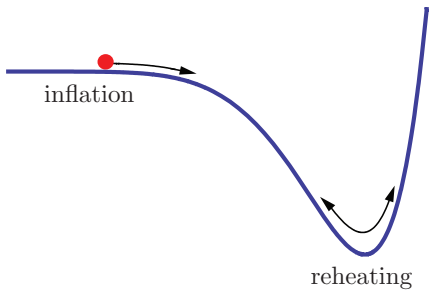
CMB and LSS experiments are starting to reach fantastic levels of precision. For the first time, we will be taking a serious stab at detecting the elusive B-modes and possibly signatures of non-Gaussianity.

4 Reheating after Inflation

If inflation is correct, then reheating is an important era in the history of our universe. It is the time when the known matter was created. It therefore may be surprising that courses on inflation rarely treat the reheating process as carefully as they should. The main reason for this is probably that reheating is a very complex phenomenon that almost requires a course on its own. However, it also contains a lot of interesting physics. In this lecture, I will try to summarize the most important effects as we now understand them. We will see that reheating can be understood as a nice application of quantum field theory in a time-dependent classical background. For further details, I refer to the classic reference by Kofman, Linde and Starobinsky.[1]

4.1 Introduction

Before drowning ourselves in technical details, it will be useful to give a qualitative overview about our current understanding of reheating. After inflation ends, the inflaton ϕ starts to oscillate around the global minimum of the potential $V(\phi)$. Most of the inflationary energy ρ_ϕ becomes kinetic energy $\frac{1}{2}\dot{\phi}^2$. To avoid an empty universe requires that this energy is converted into Standard Model degrees of freedom. Throughout this chapter, we will study a simple toy model that contains all the important physical effects



$$\mathcal{L} = \tfrac{1}{2}(\partial_\mu\phi)^2 - V(\phi) - g^2\phi^2\chi^2 \ . \tag{4.1.1}$$

Here, we have introduced a direct coupling between the inflaton and an additional boson[2] χ. The field χ is a proxy for Standard Model fields or any hidden sector fields that subsequently decay into Standard Model fields. We assume that the effective potential has a minimum at $\phi = \sigma$. Finite σ will serve as a toy model for spontantaneous symmetry breaking. Of course, this also includes the case without symmetry breaking if we set $\sigma = 0$. Expanding around the minimum we have

$$V(\phi) = \tfrac{1}{2}m^2(\phi - \sigma)^2 + \cdots \ . \tag{4.1.2}$$

[1] Kofman, Linde and Starobinsky, *Towards The Theory of Reheating After Inflation*, (arXiv:hep-ph/9704452).

[2] We could also include a Yukawa coupling to a fermion, $h\phi\psi\bar{\psi}$. However, it turns out that Bose condensation will be important for efficient reheating.

Shifting the field, $\phi - \sigma \Rightarrow \phi$, we get

$$V(\phi) + g^2\phi^2\chi^2 \quad \Rightarrow \quad \tfrac{1}{2}m^2\phi^2 + 2g^2\sigma\phi\chi^2 + g^2\phi^2\chi^2 + \cdots \ . \tag{4.1.3}$$

We have therefore generated two types of interactions between the inflaton and the additional scalar: $\phi\chi^2$ and $\phi^2\chi^2$. Both of these interactions can contribute to the decay of the inflationary energy.

In §4.2, we study reheating due to the elementary decays $\phi \to \chi\chi$ and $\phi\phi \to \chi\chi$, with decay rates $\Gamma_{\phi\to\chi\chi}$ and $\Gamma_{\phi\phi\to\chi\chi}$. These decays lead to a transfer of energy from the inflaton oscilations to the χ-field. In the equation of motion for the inflaton, this effect is captured by an additional friction term,

$$\ddot{\phi} + 3H\dot{\phi} + \Gamma\dot{\phi} + m^2\phi = 0 \ . \tag{4.1.4}$$

The expansion rate H decreases with time, and reheating completes when $\Gamma = H$.

This perturbative theory of reheating ignores many crucial effects. In particular, it doesn't include the backreaction of the *classical* inflaton oscillations on the *quantum mechanical* production of χ-particles. In particular, the ϕ-χ couplings in eq. (4.1.3) imply source terms in the equation of motion for χ. For example, for the $\phi^2\chi^2$ interaction we have

$$\ddot{\chi}_k + 3H\dot{\chi}_k + \left(\frac{k^2}{a^2} + g^2\phi^2(t)\right)\chi_k = 0 \ . \tag{4.1.5}$$

In §4.3, we show that this can lead to resonance phenomena that can enormously enhance the reheating efficiency. Depending on the size of the coupling g and the oscillation amplitude ϕ, the resonance will be either *narrow* or *broad*. In the former case, a small range of momenta ($k = k_\star \pm \Delta k$) participate in the resonance,[3] while the latter case allows resonance for a wide range of momenta ($k \leq k_\star$). It is now believed that the first stages of reheating most likely occur in a regime of broad parametric resonance. To distinguish this stage from the subsequent stages of slow reheating and thermalization, this phenomenon is called *preheating*. However, reheating never completes at this stage of parametric resonance. Eventually the resonance becomes narrow and inefficient, and the final stages of the decay of the inflation and thermalization of its decay products can be described by the elementary theory of reheating. Thus, the elementary theory of reheating proves to be useful even in theories that begin with preheating. However, it should be applied not to the original coherently oscillating inflaton field, but to its decay products, as well as to the parts of the inflationary energy that survived preheating.

4.2 Elementary Theory of Reheating

We begin with a description the inflaton dynamics after inflation, as the field oscillates around a minimum of the potential. Ignoring the coupling to χ, the inflaton satisfies the Klein-Gordon equation,

$$\ddot{\phi} + 3H\dot{\phi} + m^2\phi = 0 \ , \qquad \text{where} \qquad H^2 = \tfrac{1}{3M_{\rm pl}^2}\left(\tfrac{1}{2}\dot{\phi}^2 + \tfrac{1}{2}m^2\phi^2\right) \ . \tag{4.2.6}$$

This has the following solution

$$\phi(t) \approx \Phi(t)\sin(mt) \ , \qquad \text{where} \qquad \Phi(t) \sim \frac{M_{\rm pl}}{mt} \ . \tag{4.2.7}$$

[3] As we will see, this regime can be understood as the amplification of the inflaton decay due to Bose condensation of the produced χ-particles.

Averaged over many oscillations the scale factor grows as $a \sim t^{2/3}$ and the energy density is

$$\rho_\phi = \tfrac{1}{2}\langle\dot\phi^2\rangle + \tfrac{1}{2}m^2\langle\phi^2\rangle \simeq \tfrac{1}{2}m^2\Phi^2 \sim a^{-3} \ . \tag{4.2.8}$$

This reproduces the well-known result that a scalar field oscillating in a quadratic potential behaves as pressureless dust.

Exercise. Show that a scalar field oscillating in a quartic potential $V(\phi) = \frac{1}{4}\lambda\phi^4$ behaves like radiation: $\rho_\phi \sim a^{-4}$.

A homogeneous scalar field oscillating with frequency $\omega = m$ can be considered as a coherent wave of ϕ-particles with zero momenta and particle density

$$n_\phi = \frac{\rho_\phi}{m} = \frac{1}{2m}\left(\langle\dot\phi^2\rangle + m^2\langle\phi^2\rangle\right) \simeq \tfrac{1}{2}m\Phi^2 \ . \tag{4.2.9}$$

So far, this has ignored the coupling to χ and the associated particle production. Let us now include this effect perturbatively. For concreteness, we consider the coupling $2g^2\sigma\phi\chi^2$. The decay rate of the process $\phi \to \chi\chi$ is computed via standard field theory methods[4]

$$\Gamma_{\phi\to\chi\chi} = \frac{g^4\sigma^2}{8\pi m} \ . \tag{4.2.10}$$

Taking into account the expansion of the universe, the time-evolution equations for the number densities of the ϕ and χ-particles can be written as

$$\frac{1}{a^3}\frac{d(a^3 n_\phi)}{dt} = -\Gamma n_\phi \qquad \text{and} \qquad \frac{1}{a^3}\frac{d(a^3 n_\chi)}{dt} = 2\,\Gamma n_\phi \ , \tag{4.2.11}$$

where the factor of two in the second equation arises because one ϕ-particle decays into two χ-particles. Using eq. (4.2.9), this can be written as

$$\ddot\phi + 3H\dot\phi + \Gamma\dot\phi + m^2\phi = 0 \ . \tag{4.2.12}$$

We see that the particle decay can be described by an additional friction term in the Klein-Gordon equation. Reheating completes when the Hubble expansion rate $H \sim \frac{2}{3t}$ drops below the decay rate Γ. At this time the energy density is

$$\rho(t_r) = 3\Gamma^2 M_{\rm pl}^2 \ . \tag{4.2.13}$$

Assuming that thermodynamic equilibrium is reached quickly after that, we can relate this to the reheating temperature

$$\rho(t_r) = 3\Gamma^2 M_{\rm pl}^2 = \frac{\pi^2}{30} g_\star T_r^4 \ , \tag{4.2.14}$$

where $g_\star \sim 100$ is the number of relativistic degrees of freedom at that time. We find

$$T_r \sim 0.1\sqrt{\Gamma M_{\rm pl}} \ . \tag{4.2.15}$$

Exercise. Repeat the analysis for the coupling $g^2\phi^2\chi^2$. First, show that

$$\Gamma_{\phi\phi\to\chi\chi} = \frac{g^4\Phi^2}{8\pi m} \ .$$

Notice that $\Gamma \propto \Phi^2 \sim 1/t^2$ decreases faster than $H \sim 1/t$. When/how does reheating end?

---

[4] Peskin and Schroeder, *Introduction to Quantum Field Theory.*

4.3 Parametric Resonance and Preheating

The perturbative analysis above has an important short coming: it ignored the coherent nature of the oscillating inflaton field. In reality, the beginning of reheating is not well-described by a superposition of free asymptotic single inflaton states. Instead the inflaton is a coherently oscillating homogeneous field. When we take this into account we are led to the possibility that the time-dependent classical inflaton background ϕ induces the quantum mechanical production of matter particles χ. In this section, we will see that this can completely change our conception of reheating.

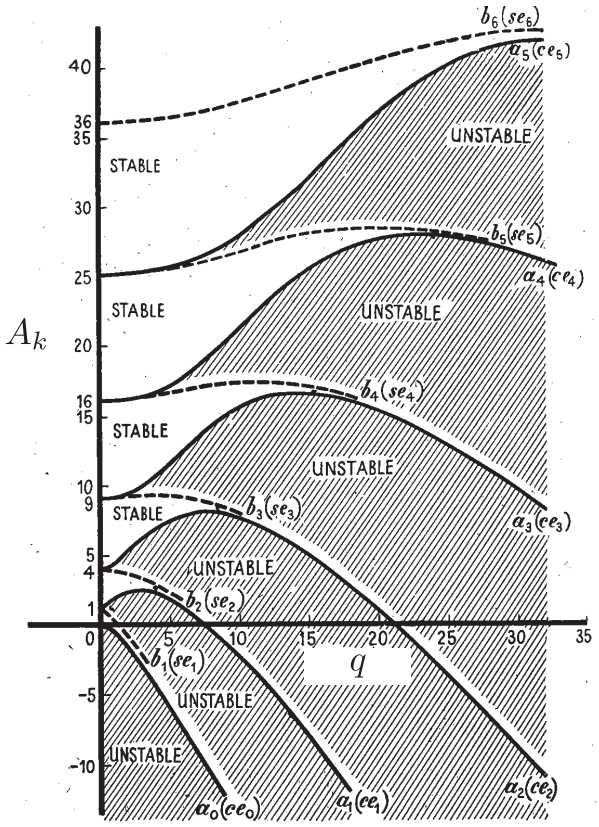


Figure 4.1: *Instability bands of the Mathieu equation.*

4.3.1 QFT in a Time-Dependent Background

Consider the *quantum* field $\hat{\chi}$ in the *classical* background $\phi(t)$,

$$\hat{\chi}(t,\mathbf{x}) = \int \frac{\mathrm{d}^3k}{(2\pi)^{3/2}} \left(\hat{a}_{\mathbf{k}}\, \chi_k(t) e^{-i\mathbf{k}\cdot\mathbf{x}} + \hat{a}^\dagger_{\mathbf{k}}\, \chi^*_k(t) e^{i\mathbf{k}\cdot\mathbf{x}} \right) \ , \tag{4.3.16}$$

where $\hat{a}^\dagger_{\mathbf{k}}$ and $\hat{a}_{\mathbf{k}}$ are creation and annihilation operators, respectively, and the mode functions satisfy[5]

$$\ddot{\chi}_k + 3H\dot{\chi}_k + \left(\frac{k^2}{a^2} + g^2\phi^2(t) \right) \chi_k = 0 \ . \tag{4.3.17}$$

[5]Here, we have ignored the possibility of an explicit mass for the field χ, i.e. we have set $m_\chi \equiv 0$. It would be straightforward to include finite m_χ, but since it doesn't lead to qualitatively new features we won't do so.

Ignoring the expansion of the universe, eq. (4.3.17) becomes

$$\ddot{\chi}_k + \left(k^2 + g^2\Phi^2 \sin^2(mt)\right) \chi_k = 0 \ . \tag{4.3.18}$$

Defining time as $z \equiv mt$, this becomes the *Mathieu equation*

$$\boxed{\chi_k'' + (A_k - 2q\cos 2z)\,\chi_k = 0} \ , \tag{4.3.19}$$

where

$$A_k \equiv \frac{k^2}{m^2} + 2q \quad \text{and} \quad q \equiv \frac{g^2\Phi^2}{4m^2} \ . \tag{4.3.20}$$

The properties of the solutions to the Mathieu equation have been classified in so-called stability/instability charts (see fig. 4.1). The main characteristic of the solutions is exponential instabilities within certain resonance bands Δk,

$$\chi_k \propto \exp\left(\mu_k z\right) \ , \tag{4.3.21}$$

where μ_k are called Floquet exponents. This corresponds to the exponential growth of occupation numbers (i.e. particle production)

$$n_k = |\chi_k|^2 \propto \exp\left(2\mu_k z\right) \ . \tag{4.3.22}$$

We will study the solutions of the Mathieu equation in two important regimes: $q \ll 1$ (*narrow resonance*) and $q > 1$ (*broad resonance*).

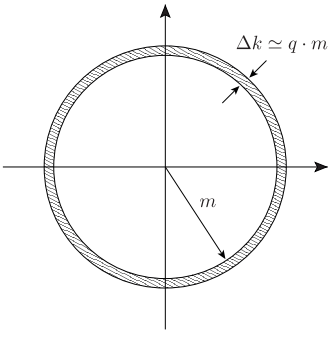
4.3.2 Narrow Resonance

We first discuss the solutions to the Mathieu equation in the regime $q \ll 1$.

Narrow Resonance in Minkowski Space

The stabilitiy/instability chart of the Mathieu equation shows that, for $q \ll 1$, resonances occur near $A_k^{(n)} \approx n^2$, where $n \in \mathbb{Z}$ (see fig. 4.1). The widths of the resonance bands is $\Delta k^{(n)} \sim m\, q^n$. For $q < 1$, the first band is the most important:

$$k = m\left(1 \pm \tfrac{1}{2}q\right) \ . \tag{4.3.23}$$



The instability parameter in this band is

$$\mu_k = \sqrt{\left(\frac{q}{2}\right)^2 - \left(\frac{k}{m} - 1\right)^2} \ . \tag{4.3.24}$$

It vanishes at the edges of the band and is maximal at its center

$$\mu_{\max} = \mu_{k=m} = \frac{q}{2} = \frac{g^2\Phi^2}{8m^2} \ . \tag{4.3.25}$$

This looks similar in spirit to the elementary theory of reheating where two ϕ-particles of mass m decay into two χ-particles with momenta, $k = m$. However, as we will see, narrow resonance is a completely different effect, relying crucially on Bose condensation of the produced particles. It dominates over elementary decays when $qm > 3H + \Gamma$.

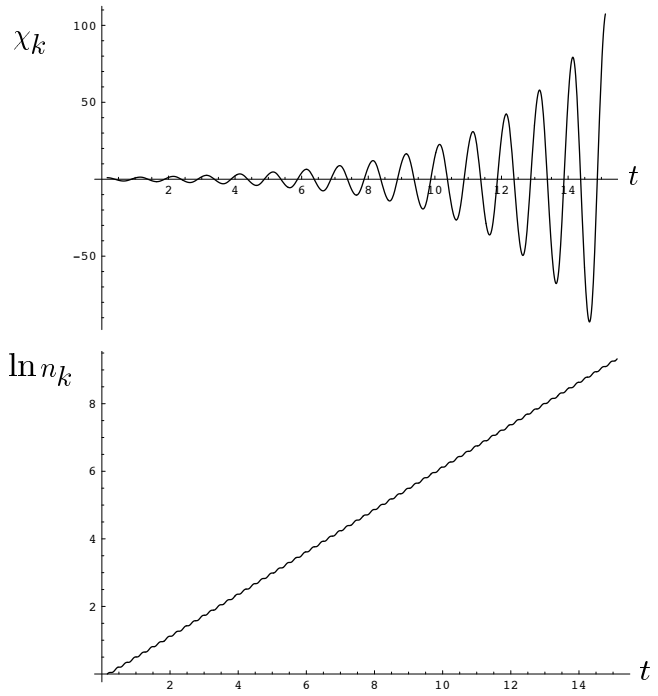


Figure 4.2: (Reproduced from Kofman et al.) *Narrow parametric resonance for the field χ in the theory $\frac{1}{2}m^2\phi^2$ in Minkowski space for $q \sim 0.1$. Here, time is plotted in units of $[m/2\pi]^{-1}$.*

Exercise. Repeat the analysis for the coupling $2g^2\sigma\phi\chi^2$. First, show that the problem can be mapped to a mathematically equivalent one. Then match the parameters of the answers.

Fig. 4.2 show a numerical simulation of the regime of narrow parametric resonance. For each oscillation of the field $\phi(t)$ the growing mode of the field χ oscillates one time. The upper figure shows the growth of the mode χ_k for the momentum k corresponding to the maximal speed of growth. The lower figure shows the logarithm of the occupation number of particles n_k in this mode. As we see, the number of particles grows exponentially, and $\ln n_k$ in the narrow resonance regime looks like a straight line with a constant slope. This slope divided by 4π gives the value of the parameter μ_k. In this particular case $\mu_k \sim 0.05$, exactly as it should be in accordance with the relation $\mu_k \sim \frac{1}{2}q$ for this model.

Narrow Resonance as Bose Condensation*

The narrow resonance effect can also be understood as the Bose condensation of χ-particles.[6] Let us sketch the basic reasoning for the case $2g^2\sigma\phi\chi^2$: In this case, a single ϕ decays into two χ. In the rest frame of the ϕ-particle, the momenta of the two produced χ-particles have the same magnitude k, but opposite directions. If the corresponding states in phase space are already occupied, then the inflaton decay is enhanced by a Bose factor. The rate of the $\phi \to \chi\chi$ process is proportional to

$$\Gamma_{\phi\to\chi\chi} \;\propto\; \Big|\langle \mathrm{n}_\phi - 1, \mathrm{n}_{\mathbf{k}} + 1, \mathrm{n}_{-\mathbf{k}} + 1|\;\; \hat{a}^+_{\mathbf{k}}\hat{a}^+_{-\mathbf{k}}\hat{a}^-_\phi \;\; |\mathrm{n}_\phi, \mathrm{n}_{\mathbf{k}}, \mathrm{n}_{-\mathbf{k}}\rangle\Big|^2 \;=\; (\mathrm{n}_{\mathbf{k}}{+}1)(\mathrm{n}_{-\mathbf{k}}{+}1)\mathrm{n}_\phi\,. \quad (4.3.26)$$

[6] Mukhanov, *Physical Principles of Cosmology.*

The inverse decay $\chi\chi \to \phi$ can also take place. Its rate is

$$\Gamma_{\chi\chi\to\phi} \;\propto\; \left|\langle \mathrm{n}_\phi + 1, \mathrm{n}_{\mathbf{k}} - 1, \mathrm{n}_{-\mathbf{k}} - 1|\;\hat{a}^+_\phi \hat{a}^-_{\mathbf{k}} \hat{a}^-_{-\mathbf{k}}\;|\mathrm{n}_\phi, \mathrm{n}_{\mathbf{k}}, \mathrm{n}_{-\mathbf{k}}\rangle\right|^2 \;=\; \mathrm{n}_{\mathbf{k}}\mathrm{n}_{-\mathbf{k}}(\mathrm{n}_\phi + 1)\;. \tag{4.3.27}$$

In this section, pay careful attention to the fonts to distinguish occupation numbers (n) from number densities (n). Taking into accoung that $\mathrm{n}_{\mathbf{k}} = \mathrm{n}_{-\mathbf{k}} \equiv \mathrm{n}_k$ and $\mathrm{n}_\phi \gg 1$, we find that the effective decay rate in (4.2.11) is

$$\Gamma \simeq \Gamma_\chi(1 + 2\,\mathrm{n}_k)\;. \tag{4.3.28}$$

It remains to determine n_k in terms of the number density n_χ. In the rest frame of ϕ, both χ-particles have energy $\frac{1}{2}m$. The three-momentum of the produced χ-particles is therefore

$$k = \left(\left(\frac{m}{2}\right)^2 - 4g^2\sigma\phi(t)\right)^{1/2}\;, \tag{4.3.29}$$

where we assume $4g^2\sigma\phi \ll m^2$. The oscillating term, $g^2\sigma\phi \simeq g^2\sigma\Phi\sin(mt)$, leads to a scattering of the momenta in phase space. If $g^2\sigma\Phi \ll \frac{1}{16}m^2$, then all particles are created in a thin shell of width

$$\Delta k \simeq m \cdot \frac{8g^2\sigma\Phi}{m^2} \ll m\;, \tag{4.3.30}$$

centered around $k_\star \simeq \frac{1}{2}m$. Hence,

$$\mathrm{n}_{k=\frac{1}{2}m} \simeq \frac{n_\chi}{(4\pi k_\star^2\Delta k)/(2\pi)^3} \simeq \frac{\pi^2 n_\chi}{m g^2\sigma\Phi} = \frac{\pi^2\Phi}{2g^2\sigma}\frac{n_\chi}{n_\phi}\;. \tag{4.3.31}$$

Bose condensation is essential when $\mathrm{n}_k \gg 1$, or

$$n_\chi > \frac{2g^2\sigma}{\pi^2\Phi} n_\phi\;. \tag{4.3.32}$$

For $\Phi \sim M_{\rm pl}$, the occupation number therefore exceeds unity as soon as a fraction $g^2\sigma/M_{\rm pl}$ of the inflaton energy is converted into χ-particles. Since the above derivation required $g^2\sigma/M_{\rm pl} \ll \frac{1}{6}m^2/M_{\rm pl}^2 \sim 10^{-10}$, only a tiny fraction of the inflaton energy is transferred in the regime of the elementary theory of reheating, $\mathrm{n}_k < 1$. Bose condensation effects take over almost immediately.

Substituting (4.3.31) in (4.3.28), gives

$$\Gamma \simeq \frac{g^4\sigma^2}{8\pi m}\left(1 + \frac{2\pi^2\Phi}{g}\frac{n_\chi}{n_\phi}\right)\;, \tag{4.3.33}$$

where we used (4.2.10) for Γ_χ. Substituting this into the second equation in (4.2.11), leads to

$$\frac{1}{a^3}\frac{d(a^3 n_\chi)}{dN} = \frac{g^4\sigma^2}{2m^2}\left(1 + \frac{\pi^2\Phi}{g^2\sigma}\frac{n_\chi}{n_\phi}\right) n_\phi\;, \tag{4.3.34}$$

where $N \equiv mt/2\pi$ measures the number of inflaton oscillations. For simplicity, let us ignore the expansion of the universe and neglect the decrease of the inflaton amplitude due to particle production, i.e. $\Phi \approx const.$ In the regime $\mathrm{n}_k \gg 1$, cf. eq. (4.3.32), the differential equation (4.3.34) can easily be integrated

$$n_\chi \propto \exp\left(\frac{2\pi^2 g^2\sigma\Phi}{m^2}N\right) \equiv e^{2\pi\mu N}\;. \tag{4.3.35}$$

As before, we find exponential growth.

Exercise. Repeat the analysis for the coupling $g^2\phi^2\chi^2$.

Expansion and Rescattering

So far, this has ignored the expansion of the universe and the rescattering of the produced χ-particles. Both effects reduce the efficiency of the resonance. First, we see that expansion narrows the width of the resonance band, $\Delta k \propto \Phi(t) \propto 1/t$. Moreover, due to the expansion, modes redshift out of the resonance. The rescattering also moves modes out of the resonance band. This shows that narrow parametric resonance is quite delicate and requires detailed numerical simulations including all relevant effects to decide if it really occurs.

4.3.3 Broad Resonance

Next, we consider the Mathieu equation (4.3.19) in the regime $q \gg 1$. Fig. 4.1 shows that instabilities now occur for much broader ranges of k. Moreover, we anticipate that the instability coefficients μ_k will be larger and reheating will be very efficient. In this section, we discuss broad parameteric resonance both numerically and analytically.

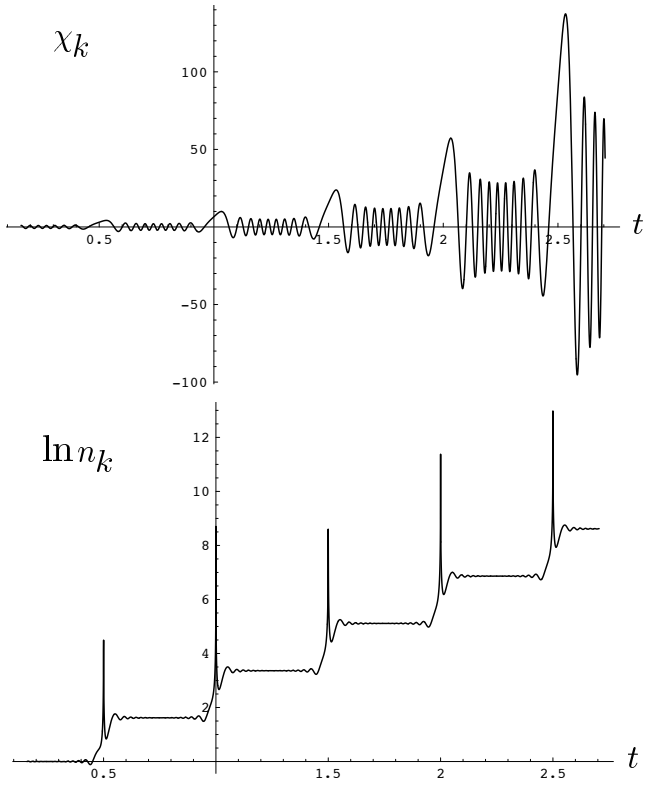


Figure 4.3: (Reproduced from Kofman et al.) *Broad parametric resonance for the field χ in Minkowski space for $q \sim 2 \times 10^2$ in the theory $\frac{1}{2}m^2\phi^2$.*

Numerical Simulations

A numerical solution in the broad resonance regime in a Minkowski background is shown in fig. 4.3. For each oscillation of the field $\phi(t)$ the field χ_k oscillates many times (since, for $q \gg 1$, the effective mass for χ is much larger than the inflaton mass, $m_\chi \equiv g\Phi \gg m$). Each peak in the amplitude of the oscillations of the field χ corresponds to a place where $\phi(t) = 0$. At

this time the occupation number n_k is not well-defined, but soon after that time it stabilizes at a new, higher level, and remains constant until the next jump. A comparison of the two parts of this figure demonstrates the importance of using proper variables for the description of preheating. Both χ_k and the integrated dispersion $\langle\chi^2\rangle$ behave erratically in the process of parametric resonance. Meanwhile n_k is an adiabatic invariant. Therefore, the behavior of n_k is relatively simple and predictable everywhere except during the short intervals of time when $\phi(t)$ is very small and the particle production occurs.

Analytic Interpretation

The structure in the plot for $\ln n_k(t)$ suggests that a simple analytic treatment should be possible. Away from $\phi(t) = 0$, the frequency of χ changes adiabatically, $|\dot{\omega}| \ll \omega^2$, and n_k is conserved. Particle production occurs when the adiabatic condition is violated

$$R \equiv \frac{|\dot{\omega}|}{\omega^2} > 1 \ , \qquad \text{where} \qquad \omega(t) = \sqrt{k^2 + g^2\phi^2(t)} \ . \tag{4.3.36}$$

For small wavenumbers, $k \to 0$, we find

$$R \simeq \frac{\dot{\phi}}{g\phi^2} \ . \tag{4.3.37}$$

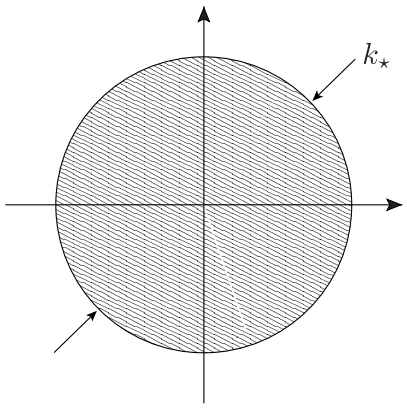
The key point about this relation is that R diverges whenever $\phi \to 0$, i.e. twice every oscillation. At these points we expect explosive particle production. For finite k, adiabaticity is violated if

$$R = \frac{g^2\phi\dot{\phi}}{(k^2 + g^2\phi^2)^{3/2}} \sim \frac{g^2\phi m\Phi}{(k^2 + g^2\phi^2)^{3/2}} > 1 \ , \tag{4.3.38}$$

where in the second equality we used $\dot{\phi} \simeq m\Phi\cos(mt) \sim m\Phi$, which is valid near the origin $\phi = 0$. Eq. (4.3.38) implies

$$k^2 \lesssim (g^2\phi m\Phi)^{2/3} - g^2\phi^2 \equiv f(\phi) \ . \tag{4.3.39}$$

We see that this starts to be satisfied, for small k, when the field $\phi(t)$ becomes smaller than $\phi_{\rm max} \equiv \sqrt{\frac{m\Phi}{g}}$, where $f(\phi_{\rm max}) = 0$. The maximum range of wavenumbers satisfy eq. (4.3.39) at $\phi_\star$, where $f'(\phi_\star) = 0$. A rough estimate is $\phi_\star \sim \frac{1}{2}\phi_{\rm max}$. The effective range of k participating in the resonance is roughly,



$$k^2 \lesssim k_\star^2 \equiv gm\Phi \ . \tag{4.3.40}$$

This condition changes in a simple way if we take into account the expansion of the universe,

$$\frac{k^2}{a^2(t)} \lesssim k_\star^2(t) \equiv gm\Phi(t) \ . \tag{4.3.41}$$

We see that the expansion makes broad resonance *more* effective, since more k-modes are redshifted into the instability band as time proceeds.

Exercise. Consider chaotic inflation with $V(\phi) = \frac{1}{2}m^2\phi^2$ and the coupling $g^2\phi^2\chi^2$. Broad resonance occurs for $g > \frac{2m}{\Phi} \sim 10^{-6}$, where we used $\Phi \sim M_{\rm pl}$ and $m \sim 10^{-6}M_{\rm pl}$ (from COBE normalization). Is such a large coupling consistent with naturalness of the inflationary potential? Consider the one-loop correction to the inflaton mass

$$\delta m^2 = \frac{g^2}{16\pi^2}\Lambda_{\rm uv}^2 \sim \frac{g^2}{16\pi^2}M_{\rm pl}^2 \ .$$

Naturalness requires $\delta m < m \sim 10^{-6}M_{\rm pl}$, or $g < 10^{-5}$. This seems to disallow the regime of broad parametric resonance for chaotic inflation. The reheating of this model then has to occur predominantly through the elementary decays of ϕ and narrow resonances in χ.

Broad Resonance in an Expanding Universe

Let us study broad resonance more quantitatively and include the expanding of the universe. It is convenient to remove the Hubble friction from the equation of motion, by defining $X_k(t) \equiv a^{3/2}(t)\chi_k(t)$. The mode equation then becomes

$$\ddot{X}_k + \omega_k^2 X_k = 0 \ , \tag{4.3.42}$$

where

$$\omega_k^2 \equiv \frac{k^2}{a^2} + g^2\Phi^2\sin^2(mt) + \Delta \ , \qquad \Delta \equiv -\tfrac{3}{4}(3H^2 + 2\dot{H}) \ . \tag{4.3.43}$$

Since $3H^2 = -2\dot{H}$ during matter domination, we can set $\Delta \equiv 0$. We also define the *comoving* occupation number of particles

$$n_k = \frac{\omega_k}{2}\left(\frac{|\dot{X}_k|^2}{\omega_k^2} + |X_k|^2\right) - \frac{1}{2} \ . \tag{4.3.44}$$

Fig. 4.4 show a simulation of broad parametric resonance in an expanding universe (plotting X_k rather than χ_k illustrates the growth of χ relative to the decay of $\Phi \propto 1/t$.) Note that the number of particles n_k in this process typically increases, but it may occasionally decrease as well. This is a distinctive feature of *stochastic resonance* in an expanding universe. A decrease in the number of particles is a purely quantum mechanical effect which would be impossible if these particles were in a state of thermal equilibrium.

The stochastic resonance effect can be understood by inspecting the behavior of the *phases* of the functions χ_k near $\phi(t) = 0$: here, Minkowski space and an expanding universe are qualitatively different. In Minkowski space, near $\phi(t) = 0$ the phases of χ_k are equal (see fig. 4.3). In an expanding universe, this is not the case (see fig. 4.4): the phases of χ_k at successive moments when $\phi(t) = 0$ are practically uncorrelated. The reason is easy to understand. The frequency of oscillations of χ_k is proportional to Φ, which in an expanding universe is a function of time. In the broad resonance regime, this implies that the frequency of oscillations of χ_k changes a lot during each oscillation of the inflaton ϕ. Interestingly, this doesn't completely destroy the resonance effect. As we will show analytically in the next section (and as is confirmed in the simulations), although in an expanding universe with $q \gg 1$ the phases of χ_k are practically stochastic, in 75% of all events the amplitude of χ_k grows after passing through $\phi(t) = 0$. Over a long time period, the occupation number of χ-particles therefore grows (see fig. 4.5). However, the parameter $q = \frac{g^2\Phi^2}{4m^2} \propto t^{-2}$ decreases with time, making the resonance more and more narrow. Eventually, the resonance ceases to exist and n_k stabilizes at a constant value.

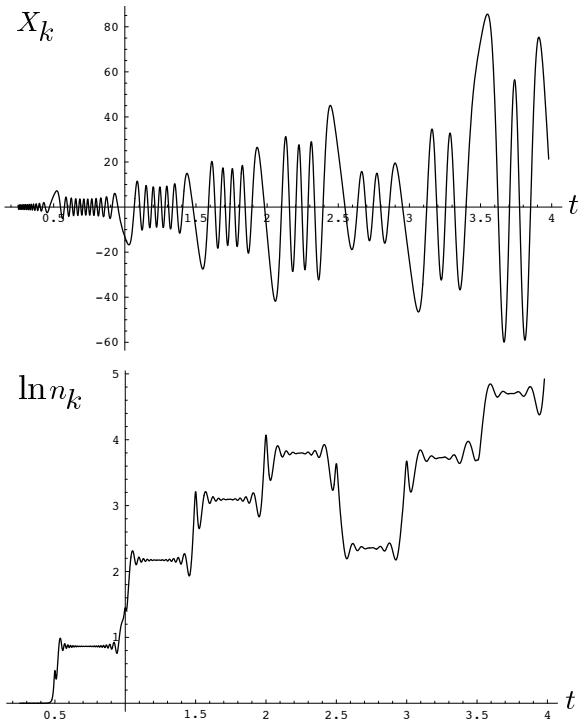


Figure 4.4: (Reproduced from Kofman et al.) *Early stages of parametric resonance in the theory* $\frac{1}{2}m^2\phi^2$ *in an expanding universe with scale factor* $a \sim t^{2/3}$ *for* $g = 5 \times 10^{-4}$, $m = 10^{-6} M_{\rm pl}$. *The initial value of the parameter* q *in this process is* $q_0 \sim 3 \times 10^3$.

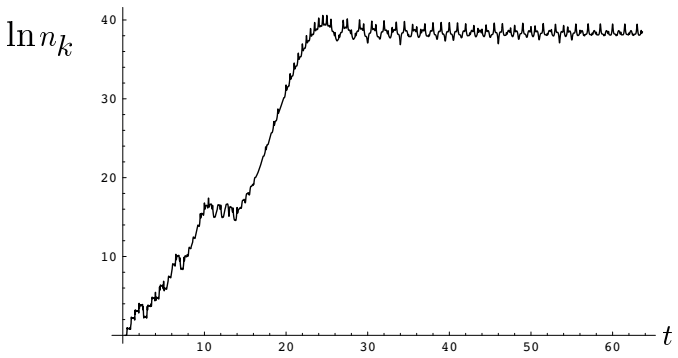


Figure 4.5: (Reproduced from Kofman et al.) *Same process as in fig. 4.4, but now followed over a longer period of time.*

Broad Resonance as Schrödinger Scattering*

We would like to have a more analytic understanding of broad resonance in an expanding universe. The structures we have seen in the numerical simulations suggest that this is feasible.

Let us label with t_j, $j \in \mathbb{Z}$, the times when $\phi(t_j) = 0$. Away from those points the modes evolve adiabatically, $e^{\pm i \int \omega \, \mathrm{d}t}$. Consider incoming waves

$$X_k^j(t) = \frac{\alpha_k^j}{\sqrt{2\omega}} e^{-i \int_0^t \omega \, \mathrm{d}t} + \frac{\beta_k^j}{\sqrt{2\omega}} e^{+i \int_0^t \omega \, \mathrm{d}t} \ , \tag{4.3.45}$$

where the coefficients α_k^j and β_k^j are constant for $t_{j-1} < t < t_j$. At $t = t_j$ the waves scatter, and the outgoing waves are

$$X_k^{j+1}(t) = \frac{\alpha_k^{j+1}}{\sqrt{2\omega}} e^{-i\int_0^t \omega\,\mathrm{d}t} + \frac{\beta_k^{j+1}}{\sqrt{2\omega}} e^{+i\int_0^t \omega\,\mathrm{d}t} \ , \tag{4.3.46}$$

where the coefficients α_k^{j+1} and β_k^{j+1} are constant for $t_j < t < t_{j+1}$. The outgoing amplitudes α_k^{j+1} and β_k^{j+1} can be expressed in terms of the incoming amplitudes α_k^j and β_k^j and reflection and transmission amplitudes R_k and T_k,

$$\begin{pmatrix} \alpha_k^{j+1}\, e^{-i\theta_k^j} \\ \beta_k^{j+1}\, e^{+i\theta_k^j} \end{pmatrix} = \begin{pmatrix} \frac{1}{T_k} & \frac{R_k^*}{T_k^*} \\ \frac{R_k}{T_k} & \frac{1}{T_k^*} \end{pmatrix} \begin{pmatrix} \alpha_k^j\, e^{-i\theta_k^j} \\ \beta_k^j\, e^{+i\theta_k^j} \end{pmatrix} \ . \tag{4.3.47}$$

Here, $\theta_k^j = \int_0^{t_j} \mathrm{d}t\, \omega(t)$ is the phase accumulated by the time t_j. To compute R_k and T_k, we approximate the coupling in the vicinity of t_j as $g^2\phi^2(t) \approx g^2\Phi^2 m^2 (t-t_j)^2 \equiv k_\star^4 (t-t_j)^2$. The mode equation then is

$$\frac{d^2 X_k}{dt^2} + \left(\frac{k^2}{a^2} + k_\star^4 (t-t_j)^2 \right) X_k = 0 \ . \tag{4.3.48}$$

Defining a rescaled momentum $\kappa \equiv k/(a k_\star)$ and new time variable $\tau \equiv k_\star (t - t_j)$, this becomes

$$\boxed{\frac{d^2 X_k}{d\tau^2} + \left(\kappa^2 + \tau^2\right) X_k = 0} \ . \tag{4.3.49}$$

This equation may be interpreted as a *one-dimensional Schrödinger equation for particle scattering through an inverted parabolic potential*, $V = -\tau^2$. The problem has therefore been reduced to a standard quantum mechanics problem. The solution is

$$R_k = -\frac{i e^{i\varphi_k}}{\sqrt{1 + e^{\pi\kappa^2}}} \qquad \text{and} \qquad T_k = -\frac{i e^{i\varphi_k}}{\sqrt{1 + e^{-\pi\kappa^2}}} \ , \tag{4.3.50}$$

where

$$\varphi_k \equiv \arg\Gamma\left(\frac{1 + i\kappa^2}{2}\right) + \frac{\kappa^2}{2}\left(1 + \ln\frac{2}{\kappa^2}\right) \ . \tag{4.3.51}$$

Notice that $R_k = -iT_k e^{-\frac{\pi}{2}\kappa^2}$ and $|R_k|^2 + |T_k|^2 = 1$. Substituting (4.3.50) into (4.3.47), we get

$$\begin{pmatrix} \alpha_k^{j+1} \\ \beta_k^{j+1} \end{pmatrix} = \begin{pmatrix} \sqrt{1 + e^{-\pi\kappa^2}}\, e^{i\varphi_k} & i e^{-\frac{\pi}{2}\kappa^2 + 2i\theta_k^j} \\ -i e^{-\frac{\pi}{2}\kappa^2 - 2i\theta_k^j} & \sqrt{1 + e^{-\pi\kappa^2}}\, e^{-i\varphi_k} \end{pmatrix} \begin{pmatrix} \alpha_k^j \\ \beta_k^j \end{pmatrix} \ . \tag{4.3.52}$$

We can use this to express the number density of the outgoing particles $n_k^{j+1} = |\beta_k^{j+1}|^2$ in terms of the number density of the incoming particles $n_k^j = |\beta_k^j|^2$ and the parameters of the scattering potential,

$$n_k^{j+1} = e^{-\pi\kappa^2} + (1 + 2e^{-\pi\kappa^2}) n_k^j - 2e^{-\frac{\pi}{2}\kappa^2}\sqrt{1 + e^{-\pi\kappa^2}}\sqrt{n_k^j (n_k^j + 1)}\, \sin\theta_{\rm tot}^j \ , \tag{4.3.53}$$

where $\theta_{\rm tot}^j \equiv 2\theta_k^j - \varphi_k + \arg\beta_k^j - \arg\alpha_k^j$.

Digression. Let us digress briefly to understand part of the answer more intuitively. Consider eq. (4.3.49). This has two solutions, χ_k^{in} and χ_k^{out}, associated with vacuum states in the far past and the far future. These two sets of modes are related by a Bogoliubov transformation

$$\chi_k^{in} = \alpha_k \chi_k^{out} + \beta_k \chi_k^{out*} \ . \tag{4.3.54}$$

If we start in the state with no particles in the far past, the number density of particles in the far future is

$$n_k = |\beta_k|^2 \ . \tag{4.3.55}$$

Part of the incoming wave χ_k^{in} will be transmitted, $T_k\chi_k^{in}$, and part of it will be reflected $R_k\chi_k^{in*}$. The Bogoliubov coefficient in (4.3.54) can therefore be expressed in terms of the transmission and reflection coefficients

$$\beta_k = \frac{R_k^*}{T_k^*} \ . \tag{4.3.56}$$

We use a trick from quantum mechanics to relate R_k and T_k. Moving along the *real* time axis, the WKB form of the solution $\chi_k^{in}(\tau)$ will be violated at small τ. However, if we take τ to be *complex*, then we can move from $\tau = -\infty$ to $\tau = +\infty$ along a complex contour in such a way that the WKB approximation remain valid throughout,

$$\chi_k^{in}(\tau) \sim \frac{1}{\sqrt{2\sqrt{\kappa^2+\tau^2}}}\, e^{-i\int^\tau \sqrt{\kappa^2+(\tau')^2}\,\mathrm{d}\tau'} \ . \tag{4.3.57}$$

Here, the integral $\int^\tau \mathrm{d}\tau'$ becomes a contour integral along a semi-circle of large radius in the lower complex τ-plane. For large $|\tau|$, we can estimate the phase integral in (4.3.57) by expanding

$$\sqrt{\kappa^2+\tau^2} \sim \tau + \frac{\kappa^2}{2\tau} \ . \tag{4.3.58}$$

Going around the semi-circle, this gives

$$(e^{-i\pi})^{-i\kappa^2} \sim i e^{-\pi\kappa^2} \ . \tag{4.3.59}$$

This is exactly the ratio between R_k^* and T_k^*, so we find

$$n_k = |\beta_k| = e^{-\pi\kappa^2} \ , \tag{4.3.60}$$

in agreement with eq. (4.3.53) for $n_k^j = 0$.

From eq. (4.3.53) we see that particle creation is significant only for $\kappa^2 \lesssim 1$, or

$$\frac{k^2}{a^2} \lesssim k_\star^2 = gm\Phi \ . \tag{4.3.61}$$

Compare this to our previous estimate (4.3.41). For large occupation numbers, $n_k^j \gg 1$, eq. (4.3.53) reduces to

$$n_k^{j+1} = n_k^j\, e^{2\pi\mu_k^j} \ , \tag{4.3.62}$$

where

$$\mu_k^j \equiv \frac{1}{2\pi} \ln\left(1 + 2e^{-\pi\kappa^2} - 2e^{-\frac{\pi}{2}\kappa^2} \sin\theta_{\rm tot}^j \sqrt{1+e^{-\pi\kappa^2}}\right) \ . \tag{4.3.63}$$

The first two terms correspond to spontaneous particle creations which always increase the particle number ($\mu_k > 0$). The last term, however, corresponds to induced particle creation, which can either increase or decrease the number of particles (depending on the sign of $\sin\theta_{\rm tot}^j$). We see that the particle creation depends crucially on the interference of the wave functions, i.e. the phase (anti)correlation between successive scatterings. For the fastest growing mode ($k = 0$), $n_k^{j+1} > n_k^j$ if

$$-\tfrac{\pi}{4} < \theta_{\rm tot}^j < \tfrac{5\pi}{4} \ . \tag{4.3.64}$$

Treating $\theta_{\rm tot}^j$ as a random variable, the range of phases in (4.3.64) implies that the solution grows about 75% of the time.

Finally, let us determine the net particle creation after a number of oscillations of the inflaton and many scatterings. We start in the vacuum state $\alpha_k^0 = 1$, $\beta_k^0 = n_k^0 = 0$ and random initial phase θ_k^0. After a number of oscillations, the occupation number of χ-particles is

$$n_k(t) = \tfrac{1}{2} e^{2\pi \sum_j \mu_k^j} \equiv \tfrac{1}{2} e^{2(m\Delta t)\mu_k} \ , \tag{4.3.65}$$

where Δt is the total duration of the resonance and we defined the "average" Floquet exponent,

$$\mu_k \equiv \frac{\pi}{m\Delta t} \sum_j \mu_k^j \ . \tag{4.3.66}$$

This parameter is typically of order unity,[7] which shows that broad parametric resonance in an expanding background can be very efficient and can convert a substantial fraction of the inflaton energy density into matter in less than a Hubble time. The number density of created particles is

$$n_\chi(t) = \frac{1}{a^3} \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \, n_k(t) = \frac{1}{4\pi^2 a^3} \int \mathrm{d}k \, k^2 e^{2(mt)\mu_k} \ . \tag{4.3.68}$$

This can be estimated by the method of steepest decent. We first note that the function μ_k has a maximum $\mu_{\max} \equiv \mu$ at some $k_{\max}$. In practice, $k_{\max} \sim \frac{1}{2} k_\star$ and

$$n_\chi(t) \sim \frac{k_\star^3}{64\pi^2 a^3 \sqrt{\pi \mu (mt)}} \, e^{2(mt)\mu} \ . \tag{4.3.69}$$

This result determines how fast the energy is drained from the inflaton field.

4.3.4 Termination of Preheating

So far, we haven't taken into account the backreaction of the produced χ-particles on the dynamics of preheating. Ultimately, this backreaction terminates preheating.

The backreaction effect on the inflaton mass can be estimated as follows

$$\Delta m^2 = g^2 \langle \chi^2 \rangle \ , \tag{4.3.70}$$

where $\langle \chi^2 \rangle$ is a quantum expectation value

$$\langle \chi^2 \rangle = \frac{1}{2\pi^2 a^3} \int k^2 \mathrm{d}k \, |X_k(t)|^2 \ . \tag{4.3.71}$$

Inserting the mode expansion of X_k, we find

$$\Delta m^2(t) \approx \frac{g n_\chi(t)}{\phi(t)} \ . \tag{4.3.72}$$

This expression looks ill-defined when ϕ crosses zero. However, the equation of motion is well-behaved, so we can estimate the size of the backreaction by replacing ϕ by its amplitude Φ. Backreaction becomes important when $\Delta m^2 > m^2$, or

$$n_\chi(t) > \frac{m^2 \Phi(t)}{g} \ . \tag{4.3.73}$$

[7] With eq. (4.3.63), we find

$$\mu_k \approx \frac{1}{2\pi} \ln 3 - \mathcal{O}(\kappa^2) \ . \tag{4.3.67}$$

At this time the broad resonance is destroyed and preheaing ends. Using eq. (4.2.7) for $\Phi(t)$ and eq. (4.3.69) for $n_\chi(t)$, we find

$$\Delta t \sim (\mu m)^{-1} \ . \tag{4.3.74}$$

This time interval is typically short compared to the Hubble time. A more rigorous analysis of backreaction effects and the end of preheating requires numerical analysis.

4.3.5 Gravitational Waves from Preheating

Although reheating is a fundamental part of any complete theory of inflation, its imprints on cosmological observables are unfortunately rather limited. With some luck reheating could produce relics like cosmic strings or other topological defects that then survive in the late universe. Another interesting possibility is that inhomogeneities in χ source gravitational waves. The basic idea is that χ-modes (and also ϕ-modes from inverse scattering) contribute a source term to the equations of motion for tensor fluctuations,

$$h_{ij}'' + 2\frac{a'}{a}h_{ij}' - \nabla^2 h_{ij} = 16\pi G \left(T_{ij}\right)^{\mathrm{TT}} \ , \tag{4.3.75}$$

where

$$T_{ij} = \partial_i\chi\partial_j\chi + \cdots \ . \tag{4.3.76}$$

Details of the computation of the gravitational wave spectrum can be found in a nice paper by Kofman and collaborators.[8] For high-scale inflation the signal peaks at very high frequencies. In fact, the peak frequencies for GUT scale inflation are much too high to be observable by future direct detection experiments. However, if inflation is very low scale, then the gravitational signal could peak around 1 Hz and hence potentially be observable.

4.4 Conclusions

This chapter has given a sketch of some of the non-trivial physical processes that we expect to be relevant during reheating. Despite our best efforts the treatment is still terribly incomplete: For example, I didn't discuss *tachyonic preheating*, which is relevant for understanding reheating in hybrid inflation. Moreover, I didn't describe the problem of *thermalization* of the decay products. This involves non-perturbative rescattering of the decay products and turbulence. However, I hope that I have given you enough background to study these important topics in your own time.

[8] Dufaux et al., *Theory and Numerics of Gravitational Waves from Preheating after Inflation*, (arXiv:0707.0875).

5 Primordial Non-Gaussianity

5.1 Why Non-Gaussianity?

The CMB power spectrum analysis reduces the WMAP data from about 10^6 pixels to 10^3 multipole moments. This enormous data compression is only justified if the primordial perturbations were drawn from a Gaussian distribution with random phases. In principle, there can be a wealth of information that is contained in deviations from the perfectly Gaussian distribution. So far experiments haven't had the sensitivity to extract this information from the data. However, this is about to change. The Planck satellite will provide accurate measurements of higher-order CMB correlations, or *non-Gaussianity*. The will allow the study of primordial quantum fields to move beyond the free field limit and start to constrain (or measure!) interactions.

Measurements of primordial non-Gaussianity are a powerful way to bring us closer to the ultimate goal of particle physics, which is to determine the action (i.e. the fields, symmetries and couplings) as a function of energy scale. At low energies, $E < 1$ TeV, physics is completely described by the Standard Model of particle physics.[1] Various lines of arguments suggest that new physics should appear close to the TeV scale. The particle physics community is eagerly awaiting experimental results from the LHC to elucidate the physics of the TeV scale. However, it seems unlikely that this new physics will also explain the inflationary era in the early universe (or be relevant for alternatives to inflation). To probe the physics at energy scales far exceeding the TeV scale, we are likely to require cosmological data. In these notes I will explain how non-Gaussianity will help in this quest. In the process, I will give the readers two essential tools – the *in-in* formalism and the δN formalism – that will allow them to perform their own computations of non-Gaussianities in models of the early universe.

The outline of these chapter is as follows: In §5.2 I describe the Gaussian and non-Gaussian statistics of the CMB from a phenomenological perspective. I introduce the bispectrum as the primary diagnostic for non-Gaussianity. I discuss how the momentum dependence of the bispectrum encodes invaluable physical information. In the rest of the notes I present different mechanism to produce observable CMB bispectra: In §5.3 I describe *quantum* mechanically generated non-Gaussianities. I introduce the *in-in* formalism to compute higher-order correlation functions and apply it to a variety of examples. In §5.4 I discuss *classical* non-Gaussianities generated after horizon-exit. I explain the δN formalism and apply it to models of inflation with additional light fields. I conclude, in §5.6, with a few remarks about future prospects for both the theory and observations of primordial non-Gaussianity.

[1] An exception to this statement are neutrino masses.

For further reading I highly recommend the two classic papers by Maldacena[2] and Weinberg[3], as well as the wonderfully clear reviews by Chen[4], Lim[5] and Komatsu[6].

5.2 Gaussian and Non-Gaussian Statistics

5.2.1 Statistics of CMB Anisotropies

CMB experiments measure temperature fluctuations ΔT as a function of the position $\hat{\mathbf{n}}$ on the sky. Depending on the resolution of the experiment, the CMB maps have $N_{\rm pix}$ number of pixels, $\hat{\mathbf{n}}_i$, with $\Delta T_i \equiv \Delta T(\hat{\mathbf{n}}_i)$. The temperature anisotropy is Gaussian when its probability density function (PDF) is

$$P_g(\Delta T) = \frac{1}{(2\pi)^{N_{\rm pix}/2}|\xi|^{1/2}} \exp\left[-\frac{1}{2}\sum_{ij}\Delta T_i(\xi^{-1})_{ij}\Delta T_j\right] , \tag{5.2.1}$$

where $\xi_{ij} \equiv \langle \Delta T_i \Delta T_j \rangle$ is the covariance matrix (or two-point correlation function) of the temperature anisotropy and $|\xi|$ is its determinant. It is common practice to expand ΔT in spherical harmonics, $\Delta T(\hat{\mathbf{n}}) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\hat{\mathbf{n}})$. The Gaussian PDF for the $a_{\ell m}$'s is

$$P_g(a) = \frac{1}{(2\pi)^{N_{\rm harm}/2}|C|^{1/2}} \exp\left[-\frac{1}{2}\sum_{\ell m}\sum_{\ell' m'} a^*_{\ell m}(C^{-1})_{\ell m,\ell' m'} a_{\ell' m'}\right] , \tag{5.2.2}$$

where $C_{\ell m,\ell' m'} \equiv \langle a^*_{\ell m} a_{\ell m}\rangle$ and $N_{\rm harm}$ is the number of ℓ and m that is summed over. For a Gaussian CMB the covariance matrix, $C_{\ell m,\ell' m'}$, provides a full description of the data. All higher correlations either vanish, $\langle a_{\ell m} a_{\ell' m'} a_{\ell' m'}\rangle = 0$, or can be expressed in terms of $C_{\ell m,\ell' m'}$. When the CMB is statistically homogeneous and isotropic (i.e. invariant under translations and rotations on the sky), then[7]

$$C_{\ell m,\ell' m'} = C_\ell \, \delta_{\ell\ell'}\delta_{mm'} , \tag{5.2.3}$$

and (5.2.2) reduces to

$$P_g(a) = \prod_{\ell m} \frac{e^{-|a_{\ell m}|^2/(2C_\ell)}}{\sqrt{2\pi C_\ell}} . \tag{5.2.4}$$

How do we describe a non-Gaussian CMB? Naively, we have a problem. There is only one way for the CMB to be Gaussian but an infinite number of ways of being non-Gaussian. So which non-Gaussian PDF do we pick? Fortunately, we know from observations that the CMB is very close to the Gaussian distribution (5.2.2). It therefore makes sense to "Taylor expand"[8] the probability distribution around a Gaussian distribution

$$P(a) = \Big[1 - \frac{1}{6}\sum_{\text{all }\ell_i m_j} \langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3}\rangle \frac{\partial}{\partial a_{\ell_1 m_1}}\frac{\partial}{\partial a_{\ell_2 m_2}}\frac{\partial}{\partial a_{\ell_3 m_3}} + \cdots\Big] \times P_g(a) . \tag{5.2.5}$$

[2] Maldacena, (arXiv:astro-ph/0210603).
[3] Weinberg, (arXiv:hep-th/0506236).
[4] Chen, (arXiv:1002.1416).
[5] Lim, (Part III, Advanced Cosmology).
[6] Komatsu, (arXiv:1003.6097).
[7] This is equivalent to $\xi_{ij} = \langle \Delta T_i \Delta T_j\rangle = \xi(|\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_j|)$.
[8] Strictly speaking this is a "Gram-Charlier expansion".

Evaluating the derivatives results in

$$P(a) \;=\; P_g(a) \times \Big[1 + \frac{1}{6} \sum_{\text{all } \ell_i m_j} \langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} \rangle \Big\{ (C^{-1}a)_{\ell_1 m_1} (C^{-1}a)_{\ell_2 m_2} (C^{-1}a)_{\ell_3 m_3} \\ - \; 3 (C^{-1})_{\ell_1 m_1, \ell_2 m_2} (C^{-1}a)_{\ell_3 m_3} \Big\} \Big] \; . \qquad (5.2.6)$$

This formula tells us, as expected, that the leading deviation from the Gaussian PDF is proportional to the angular bispectrum $\langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} \rangle$. The formula is also used to estimate the angular bispectrum from data by maximizing this PDF. For the remainder of these notes it suffices to know that this can be done. We will instead focus our attention on describing possible sources for a non-zero bispectrum.

5.2.2 Sources of Non-Gaussianity

We distinguish the following sources for a non-Gaussian CMB:

1. *Primordial non-Gaussianity*

   Non-Gaussianity in the primordial curvature perturbation ζ produced in the very early universe by inflation (or an alternative).

2. *Second-order non-Gaussianity*

   Non-Gaussianity arising from non-linearities in the transfer function relating ζ to the CMB temperature anisotropy ΔT at recombination.

3. *Secondary non-Gaussianity*

   Non-Gaussianity generated by 'late' time effects after recombination (e.g. lensing).

4. *Foreground non-Gaussianity*

   Non-Gaussianity created by Galactic and extra-Galactic sources.

All of these sources contribute to the observed signal and it is important to understand them both theoretically and empirically. Only if we understand the secondary non-Gaussianity well enough can we hope to extract the primordial signal reliably. Having said that, for the remainder of these notes we will be high-energy chauvinists and focus entirely on the microphysical origin of primordial non-Gaussianity.

5.2.3 Primordial Bispectrum

The leading non-Gaussian signature is the three-point correlation function, or its Fourier-equivalent, the bispectrum

$$\langle \zeta_{\mathbf{k}_1} \zeta_{\mathbf{k}_2} \zeta_{\mathbf{k}_3} \rangle = B_\zeta(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) \; . \qquad (5.2.7)$$

For perturbations around an FRW background, the momentum dependence of the bispectrum simplifies considerably: Because of homogeneity, or translation invariance, the bispectrum is proportional to a delta function of the sum of the momenta, $B_\zeta(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) \propto \delta(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3)$, i.e. the sum of the momentum 3-vectors must form a closed triangle. Because of isotropy, or

rotational invariance, the bispectrum only dependence on the magnitudes of the momentum vectors, but not on their orientations,

$$B_\zeta(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) = (2\pi)^3 \delta(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) B_\zeta(k_1, k_2, k_3) \ . \tag{5.2.8}$$

If the power spectrum is scale-invariant, the shape of the bispectrum only depends on two ratios of k_i's, say $x_2 \equiv k_2/k_1$ and $x_3 \equiv k_3/k_1$,

$$B_\zeta(k_1, k_2, k_3) = k_1^{-6} B_\zeta(1, x_2, x_3) \ . \tag{5.2.9}$$

Most of our examples will be of the scale-invariant form, but for the moment we will keep an open mind and not restrict to that case.

5.2.4 Shape, Running and Amplitude

The bispectrum is often written as

$$B_\zeta(k_1, k_2, k_3) = \frac{S(k_1, k_2, k_3)}{(k_1 k_2 k_3)^2} \cdot \Delta^2_\zeta(k_\star) \ , \tag{5.2.10}$$

where $\Delta^2_\zeta(k_\star) = k_\star^3 P_\zeta(k_\star)$ is the dimensionless power spectrum evaluated at a fiducial momentum scale $k_\star$. The function S is dimensionless and, for scale-invariant bispectra, invariant under rescaling of all momenta. Moreover, it are the functions S that appear in the integrals of the CMB bispectra. We distinguish two types of momentum dependence of S:

- The *shape* of the bispectrum refers to the dependence of S on the momentum ratios k_2/k_1 and k_3/k_1, while fixing the overall momentum scale $K \equiv \frac{1}{3}(k_1 + k_2 + k_3)$.
- The *running* of the bispectrum refers to the dependence of the bispectrum on the overall momentum K, while keeping the ratios k_2/k_1 and k_3/k_1 fixed.

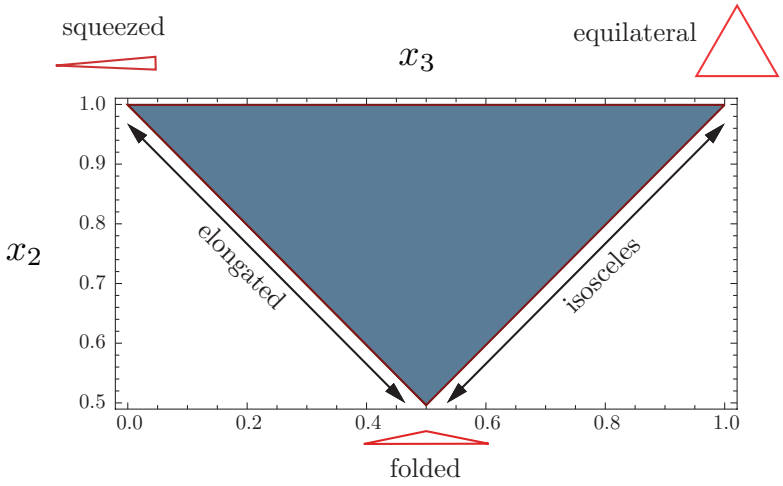


Figure 5.1: *Momentum configurations of the bispectrum.*

Finally, it is conventional to define the *amplitude* of non-Gaussianity as the size of the bispectrum in the equilateral momentum configuration,[9]

$$f_{\rm NL}(K) = \frac{5}{18} S(K, K, K) \ , \tag{5.2.11}$$

[9]The factor of $\frac{5}{18}$ is a historical convention (see §5.2.5).

where we have indicated that $f_{\rm NL}$ can depend on the overall momentum. However, for scale-invariant bispectra, $f_{\rm NL}$ is a constant and it is convenient to extract it explicitly from the shape function. In this case we write the bispectrum as

$$B_\zeta(k_1,k_2,k_3) = \frac{18}{5}\, f_{\rm NL}\cdot \frac{\mathcal{S}(k_1,k_2,k_3)}{(k_1k_2k_3)^2}\cdot \Delta_\zeta^2\ , \tag{5.2.12}$$

where now the shape function is normalized as $\mathcal{S}(K,K,K)\equiv 1$. We will use this definition of $\mathcal{S}$ in the remainder.

5.2.5 Shapes of Non-Gaussianity

Local Non-Gaussianity

One of the first ways to parameterize non-Gaussianity phenomenologically was via a non-linear correction to a Gaussian perturbation ζ_g,.[10]

$$\zeta(\mathbf{x}) = \zeta_g(\mathbf{x}) + \frac{3}{5} f_{\rm NL}^{\rm loc.}\ \left[\zeta_g(\mathbf{x})^2 - \langle \zeta_g(\mathbf{x})^2\rangle\right]\ . \tag{5.2.13}$$

This definition is local in real space and therefore called *local non-Gaussianity*. The bispectrum of local non-Gaussianity is

$$\begin{aligned} B_\zeta(k_1,k_2,k_3) &= \frac{6}{5} f_{\rm NL}^{\rm loc.}\times \left[P_\zeta(k_1)P_\zeta(k_2)+P_\zeta(k_2)P_\zeta(k_3)+P_\zeta(k_3)P_\zeta(k_1)\right]\ , &(5.2.14)\\ &= \frac{6}{5} f_{\rm NL}^{\rm loc.}\frac{\Delta_\zeta^2}{(k_1k_2k_3)^3}\left(\frac{k_1^2}{k_2k_3}+\frac{k_2^2}{k_1k_3}+\frac{k_3^2}{k_1k_2}\right)\ , &(5.2.15)\end{aligned}$$

where in the second line we assumed a scale-invariant spectrum, $P_\zeta(k)=\Delta_\zeta k^{-3}$. We can read off the local shape function as

$$\mathcal{S}_{\rm loc.}(k_1,k_2,k_3) = \frac{1}{3}\left(\frac{k_3^2}{k_1k_2}+2\,\text{perms.}\right)\ . \tag{5.2.16}$$

Without loss of generality, let us order the momenta such that $k_1\le k_2\le k_3$. The bispectrum for local non-Gaussianity is then largest when the smallest k (i.e. k_1) is very small, $k_1\ll k_2\sim k_3$. By momentum conservation, the other two momenta are then nearly equal. In this *squeezed* limit, the bispectrum for local non-Gaussianity becomes

$$\lim_{k_1\ll k_2\sim k_3}\mathcal{S}_{\rm loc.}(k_1,k_2,k_3) = \frac{2}{3}\frac{k_2}{k_1}\,. \tag{5.2.17}$$

We will later prove a powerful theorem that states that non-Gaussianity with a large squeezed limit *cannot* arise in single-field inflation, independent of the details of the inflationary action.

Equilateral Non-Gaussianity

We will see below that higher-derivative corrections during inflation can lead to large non-Gaussianities. A key characteristic of derivative interactions is that they are suppressed when any individual mode is far outside the horizon. This suggests that the bispectrum is maximal

[10] The factor of 3/5 in eq. (5.2.13) is conventional since non-Gaussianity was first defined in terms of the Newtonian potential, $\Phi(\mathbf{x})=\Phi_g(\mathbf{x})+f_{\rm NL}^{\rm loc.}\ \left[\Phi_g(\mathbf{x})^2-\langle\Phi_g(\mathbf{x})^2\rangle\right]$, which during the matter era is related to ζ by a factor of 3/5.

when all three moves have wavelengths equal to the horizon size. The bispectrum therefore has a shape that peaks in the equilateral configuration, $k_1 = k_2 = k_3$. The CMB analysis that searches for these signals uses the following template for the shape function

$$\mathcal{S}_{\text{equil.}}(k_1, k_2, k_3) = \left(\frac{k_1}{k_2} + 5\,\text{perms.}\right) - \left(\frac{k_1^2}{k_2 k_3} + 2\,\text{perms.}\right) - 2\,. \tag{5.2.18}$$

This template approximates the shape that we will compute for higher-derivative theories (such as DBI inflation) in §5.3.2.

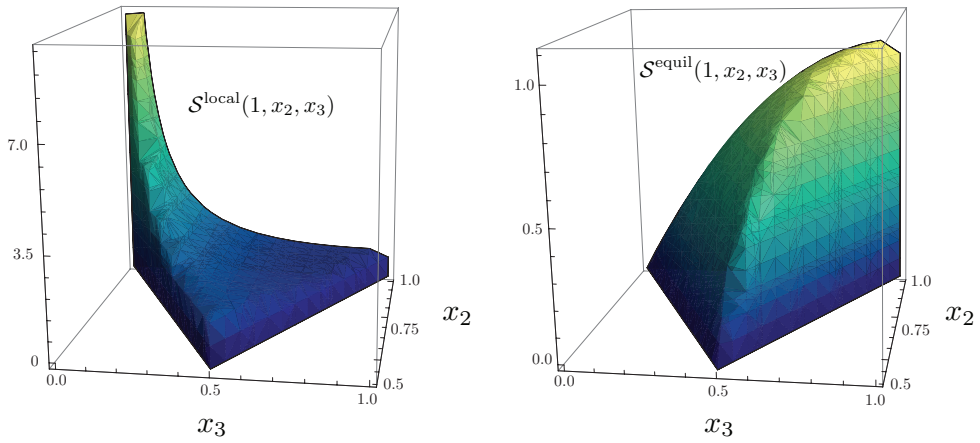


Figure 5.2: *Shape functions of local and equilateral non-Gaussianity.*

The Cosine

What if the real CMB was truly non-Gaussian, but we searched for the non-Gaussianity with the wrong template? Can we still detect a signal?

To answer this question, we digress briefly and sketch how non-Gaussianity is measured in practice. Importantly, the signal is much too small for a mode-by-mode measurement of the bispectrum. Instead, we have to start with a theoretically motivated momentum dependence for the bispectrum and fit for the overall amplitude. Concretely, imagine that we measure $\zeta_{\mathbf{k}}$ in a three-dimensional survey.[11] The three-point function is of the form

$$\langle \zeta_{\mathbf{k}_1} \zeta_{\mathbf{k}_2} \zeta_{\mathbf{k}_3} \rangle = f_{\text{NL}} \cdot (2\pi)^3 \delta(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3)\, B(k_1, k_2, k_3)\ . \tag{5.2.19}$$

We would like to test for some particular shape function $B(k_1, k_2, k_3)$ and use the data to measure the overall amplitude f_{NL}. In the limit of small non-Gaussianity, it can be shown that the best estimator for f_{NL} is

$$\widehat{f}_{\text{NL}} = \frac{\sum_{\mathbf{k}_i} \zeta_{\mathbf{k}_1} \zeta_{\mathbf{k}_2} \zeta_{\mathbf{k}_3} \frac{B(k_1,k_2,k_3)}{P_{k_1} P_{k_2} P_{k_3}}}{\sum_{\mathbf{k}_i} \frac{B(k_1,k_2,k_3)^2}{P_{k_1} P_{k_2} P_{k_3}}}\ , \tag{5.2.20}$$

where $P_{k_i} \equiv P_\zeta(k_i)$ is the power spectrum and the sums are over all physical triangles in momentum space. Eq. (5.2.20) naturally leads us to define a 'scalar product' of two bispectra B_1 and B_2,

$$B_1 \cdot B_2 \equiv \sum_{\mathbf{k}_i} \frac{B_1(k_1, k_2, k_3) B_2(k_1, k_2, k_3)}{P_{k_1} P_{k_2} P_{k_3}}\ . \tag{5.2.21}$$

[11] Of course, the CMB measurements are two-dimensional, but the principle will be the same.

If two shapes have a small scalar product, the optimal estimator (5.2.20) for one shape will be very bad in detecting non-Gaussianities with the other shape and vice versa. As we explained above, assuming isotropy, the shape function depends only on two momentum ratios, say $x_2 = k_2/k_1$ and $x_3 = k_3/k_1$. The definition of the scalar product (5.2.21) contains a factor of $x_2^3 x_3^3$ from the power spectra in the denominator (assuming scale-invariance). Furthermore, we get a measure factor $x_2 x_3$ when going from a three-dimensional sum over modes to one-dimensional integrals over x_2 and x_3. This suggests that we define the scalar product of two shapes as

$$\mathcal{S}_1 \cdot \mathcal{S}_2 \equiv \int_{\mathcal{V}} \mathcal{S}_1(x_2, x_3)\mathcal{S}_2(x_2, x_3)\, \mathrm{d}x_2 \mathrm{d}x_3 \ , \tag{5.2.22}$$

where $\mathcal{S}_i$ are the shape functions defined in (5.2.12) and the integrals are only over physical momenta satisfying the triangle inequality: $0 \leq x_2 \leq 1$ and $1 - x_2 \leq x_3 \leq 1$. An important quantity is the normalized scalar product, or cosine, of two shapes,

$$\mathcal{C}(\mathcal{S}_1, \mathcal{S}_2) = \frac{\mathcal{S}_1 \cdot \mathcal{S}_2}{(\mathcal{S}_1 \cdot \mathcal{S}_1)^{1/2}(\mathcal{S}_2 \cdot \mathcal{S}_2)^{1/2}} \ . \tag{5.2.23}$$

The cosine provides an observationally motivated answer to the questions: "How different are two bispectrum shapes?"

Orthogonal Non-Gaussianity

Using the cosine we can define a phenomenological shape that is orthogonal to both the local and equilateral templates, i.e. $\mathcal{S}_{\text{ortho.}} \cdot \mathcal{S}_{\text{loc.}} = \mathcal{S}_{\text{ortho.}} \cdot \mathcal{S}_{\text{equil.}} \equiv 0$,

$$\mathcal{S}_{\text{ortho.}}(k_1, k_2, k_3) = -3.84 \left(\frac{k_1^2}{k_2 k_3} + 2\,\text{perms} \right) + 3.94 \left(\frac{k_1}{k_2} + 5\,\text{perms.} \right) - 11.10 \ . \tag{5.2.24}$$

Below we will see how this shape arises in inflationary theories with higher-derivative interactions.

5.3 Quantum Non-Gaussianities

Let us start to do some physics. There are two classes of mechanisms that generate non-Gaussianity during inflation:

i) quantum-mechanical effects can generate non-Gaussianity at or before horizon exit;

ii) classical non-linear evolution can generate non-Gaussianity after horizon exit.

In this section we introduce the Schwinger-Keldysh *in-in* formalism[12] to compute quantum non-Gaussianities during inflation. In the next section we present the δN formalism to compute the classical effects.

[12] After pioneering work by Calzetta and Hu and Jordan the application of the *in-in* formalism to cosmological problems was revived by Maldacena and Weinberg.

5.3.1 The *in-in* Formalism

The problem of computing correlation functions in cosmology differs in important ways from the corresponding analysis of quantum field theory applied to particle physics. In particle physics the central object is the S-matrix describing the transition probability for a state $|in\rangle$ in the far past to become some state $|out\rangle$ in the far future,

$$\langle out|\, S|in\rangle = \langle out(+\infty)|in(-\infty)\rangle \ . \tag{5.3.25}$$

Imposing asymptotic conditions at very early *and* very late times makes sense in this case, since in Minkowski space, states are assumed to be non-interacting in the far past and the far future when the scattering particles are far from the interaction region. The asymptotic states relevant for particle physics experiments are therefore taken to be vacuum states of the free Hamiltonian H_0. In cosmology, however, we are interested in expectation values of products of operators *at a fixed time*. Special care has to be taken to define the time-dependence of operators in the interacting theory. Boundary conditions are *not* imposed on the fields at both very early and very late times, but only at very early times, when their wavelengths are much smaller than the horizon. In this limit the interaction picture fields should have the same form as in Minkowski space[13]. This leads to the definition of the Bunch-Davies vacuum—the free vacuum in Minkowski space. In this section we describe the *in-in* formalism to compute cosmological correlation functions as expectation values of two $|in\rangle$ states.

Preview. In presentations of the *in-in* formalism it is easy not to see the forest for the trees. Let me therefore give a brief summary of the qualitative ideas behind the formalism, before jumping into the technical details:

Our goal is to compute n-point functions of cosmological perturbations such as the primordial curvature perturbation ζ or the gravitational wave polarization modes $h^\times$ and h^+. We collectively denote these fluctuations by the field $\psi = \{\zeta, h^\times, h^+\}$ and consider expectation values of operators such as $Q = \psi_{\mathbf{k}_1}\psi_{\mathbf{k}_2}\cdots\psi_{\mathbf{k}_n}$,

$$\langle Q\rangle = \langle in|\, Q(t)\, |in\rangle \ . \tag{5.3.26}$$

Here $|in\rangle$ is the vacuum of the interacting theory at some moment t_i in the far past and $t > t_i$ is some later time such as horizon crossing or the end of inflation. To compute the matrix element in (5.3.26) we evolve $Q(t)$ back to t_i, using the perturbed Hamiltonian δH. Computing this time-evolution is complicated by the interactions inside of $\delta H = H_0 + H_{\rm int}$ (these lead to non-linear equations of motion). We therefore introduce the *interaction picture* in which the leading time-dependence of the fields is determined by the quadratic Hamiltonian H_0 (or linear equations of motion). Corrections arising from the interactions are then treated as a power series in $H_{\rm int}$.

This leads to the following important result

$$\langle Q\rangle \ = \ \langle 0|\, \bar{T} e^{i\int_{-\infty(1-i\epsilon)}^{t} H^I_{\rm int}(t')\mathrm{d}t'}\, Q^I(t)\, T e^{-i\int_{-\infty(1+i\epsilon)}^{t} H^I_{\rm int}(t'')\mathrm{d}t''}\, |0\rangle \ , \tag{5.3.27}$$

where T ($\bar{T}$) is the (anti-)time-ordering symbol. Note that both Q^I and $H^I_{\rm int}$ are evaluated using interaction picture operators. The standard $i\epsilon$ prescription has been used to effectively turn off the interaction in the far past and project the interacting $|in\rangle$ state onto the free vacuum $|0\rangle$.

[13] This is a consequence of the equivalence principle.

By expanding the exponential we can compute the correlation function perturbatively in $H_{\rm int}$. For example, at leading order (tree-level), we find

$$\langle Q(t)\rangle \;=\; -i\int_{-\infty}^{t}\mathrm{d}t'\;\langle 0|\,[Q^I(t),H^I_{\rm int}(t')]\,|0\rangle\;. \tag{5.3.28}$$

We can use Feynman diagrams to organize the power series, drawing interaction vertices for every power of $H_{\rm int}$. Notice that in the *in-in* formalism there is *no time flow*: each vertex insertion is associated not just with momentum conservation, but also a time integral.

In the rest of this section I will derive the *in-in* master formula (5.3.27). Readers more interesting in applications of the result, may jump to §5.3.2.

Time evolution. In the Heisenberg picture, the time-dependence of the operator $Q(t)$ is determined by the Hamiltonian

$$H[\psi(t),p_\psi(t)]=\int\mathrm{d}^3x\;\mathcal{H}[\psi(t,\mathbf{x}),p_\psi(t,\mathbf{x})]\;, \tag{5.3.29}$$

where $\mathcal{H}$ is the Hamiltonian density. As usual, the field ψ and its conjugate momentum p_ψ satisfy the equal-time commutation relation

$$\big[\psi(t,\mathbf{x}),p_\psi(t,\mathbf{y})\big]=i\delta(\mathbf{x}-\mathbf{y})\;, \tag{5.3.30}$$

and all other commutators vanish. The Heisenberg equations of motion are

$$\frac{d}{dt}\psi=i\big[H,\psi\big]\qquad\text{and}\qquad\frac{d}{dt}p_\psi=i\big[H,p_\psi\big]\;. \tag{5.3.31}$$

We split the fields into a time-dependent background $\{\bar\psi(t),\bar p_\psi(t)\}$ and spacetime-dependent fluctuations $\{\delta\psi(x),\delta p_\psi(x)\}$. Our goal will be to "quantize" the fluctuations on the "classical" background. The evolution of the background is determined by the classical equations of motion

$$\dot{\bar\psi}=\frac{\partial\mathcal{H}}{\partial\bar p_\psi}\qquad\text{and}\qquad\dot{\bar p}_\psi=-\frac{\partial\mathcal{H}}{\partial\bar\psi}\;. \tag{5.3.32}$$

Since $\bar\psi$ and $\bar p_\psi$ are complex numbers they commute with everything and the fluctuations satisfy the commutation relation

$$\big[\delta\psi(t,\mathbf{x}),\delta p_\psi(t,\mathbf{y})\big]=i\delta(\mathbf{x}-\mathbf{y})\;. \tag{5.3.33}$$

We expand the Hamiltonian into a background $\bar H[\bar\psi,\bar p_\psi]$ and fluctuations $\delta H[\delta\psi,\delta p_\psi;t]$. The time-dependence of the perturbation Hamiltonian arises both from the time-dependence of the fluctuations $\{\delta\psi(t),\delta p_\psi(t)\}$ and the explicit time-dependence of the background fields $a(t)$, $\bar\psi(t)$, and $\bar p_\psi(t)$. We denote this source of time-dependence by the "$;t$" in the argument of δH. Terms linear in fluctuations cancel on using background equations of motion. The perturbation Hamiltonian δH therefore starts quadratic in fluctuations. It determines the time-dependence of the fluctuations

$$\frac{d}{dt}\delta\psi=i\big[\delta H,\delta\psi\big]\qquad\text{and}\qquad\frac{d}{dt}\delta p_\psi=i\big[\delta H,\delta p_\psi\big]\;. \tag{5.3.34}$$

We see that the perturbations are evolved using the perturbed Hamiltonian. Like in standard quantum field theory, these equations are solved by

$$\delta\psi(t,\mathbf{x})=U^{-1}(t,t_i)\delta\psi(t_i,\mathbf{x})U(t,t_i)\;, \tag{5.3.35}$$

where the unitary operator U satisfies[14]

$$\frac{d}{dt}U(t,t_i) = -i\delta H[\delta\psi(t),\delta p_\psi(t);t]\,U(t,t_i)\ , \tag{5.3.37}$$

with initial condition $U(t_i,t_i)\equiv 1$.

Time evolution in the interaction picture. To describe the time evolution of the perturbations in the presence of interactions, we split the perturbed Hamiltonian into a quadratic part H_0 and an interacting part $H_{\rm int}$,

$$\delta H = H_0 + H_{\rm int}\,. \tag{5.3.38}$$

The free-field piece H_0 will then determine the leading time-evolution of the *interaction picture* fields $\{\delta\psi^I,\delta p_\psi^I\}$,

$$\frac{d}{dt}\delta\psi^I = i\big[H_0[\delta\psi^I(t),\delta p_\psi^I(t);t],\delta\psi^I(t)\big]\ , \tag{5.3.39}$$

and similarly for δp_ψ^I. The initial conditions of the interaction picture fields are chosen to coincide with those of the "full theory", $\delta\psi^I(t_i,\mathbf{x}) = \delta\psi(t_i,\mathbf{x})$ (which for inflation corresponds to the Bunch-Davies initial conditions). It follows from (5.3.39) that in evaluating $H_0[\delta\psi^I,\delta p_\psi^I;t]$ we can take the time argument of $\delta\psi^I$ and δp_ψ^I to have any value, and in particular we can take it to be t_i, so that

$$H_0[\delta\psi^I(t),\delta p_\psi^I(t);t]\ \to\ H_0[\delta\psi(t_i),\delta p_\psi(t_i);t]\ . \tag{5.3.40}$$

Notice that H_0 still depends on time through the time-evolution of the background $(;t)$. The solution of (5.3.39) can again be written as a unitary transformation

$$\delta\psi^I(t,\mathbf{x}) = U_0^{-1}(t,t_i)\delta\psi(t_i,\mathbf{x})U_0(t,t_i)\ , \tag{5.3.41}$$

where U_0 satisfies[15]

$$\frac{d}{dt}U_0(t,t_i) = -iH_0[\delta\psi(t_i),\delta p_\psi(t_i);t]U_0(t,t_i)\ , \tag{5.3.43}$$

with $U_0(t_i,t_i)\equiv 1$.

Expectation value. We now return to our goal: $\langle Q(t)\rangle = \langle in|\,Q(t)\,|in\rangle$. We first use the operator $U(t,t_i)$ to evolve $Q(t)$ back to $Q(t_i)$

$$\langle Q(t)\rangle = \langle in|\,Q[\delta\psi(t),\delta p_\psi(t)]\,|in\rangle = \langle in|\,U^{-1}(t,t_i)Q[\delta\psi(t_i),\delta p_\psi(t_i)]U(t,t_i)\,|in\rangle\ . \tag{5.3.44}$$

[14] We recall from standard quantum field theory that eq. (5.3.35) has the following formal solution

$$U(t,t_i) = T\exp\left(-i\int_{t_i}^{t}\delta H(t)\mathrm{d}t\right)\ , \tag{5.3.36}$$

where T is the time-ordering operator (as required for time-dependent Hamiltonians). However, this form of the solution is not very useful, since it is hard to calculate $\delta\psi(t,\mathbf{x})$ in the presence of interactions (since the equations of motion would be non-linear). In the following we will develop a more useful perturbative scheme to compute U as a power series in $H_{\rm int}$. For this purpose we introduce the interaction picture.

[15] Eq. (5.3.43) has the following solution

$$U_0(t,t_i) = T\exp\left(-i\int_{t_i}^{t}H_0(t)\mathrm{d}t\right)\ . \tag{5.3.42}$$

This operator is used to evolve the interaction picture fields. All of this is just a highbrow way of saying that the interaction picture fields are calculated using the free-field Hamiltonian. In inflation, this means that the linear solutions given by the Mukhanov-Sasaki equation *are* the interaction picture fields.

We then insert the identity operator $\mathbf{1} = U_0(t,t_i)U_0^{-1}(t,t_i)$ in two places (i.e. we do nothing) and write

$$\langle Q(t)\,\rangle = \langle in|\, F^{-1}(t,t_i)U_0^{-1}(t,t_i)Q[\delta\psi(t_i),\delta p_\psi(t_i)]U_0(t,t_i)F(t,t_i)\,|in\rangle\ , \tag{5.3.45}$$

where we defined the new operator

$$F(t,t_i) \equiv U_0^{-1}(t,t_i)U(t,t_i)\ . \tag{5.3.46}$$

Since $\delta\psi(t_i) = \delta\psi^I(t_i)$ and $U_0(t,t_i)$ determines the time-evolution of the interaction picture fields, this becomes

$$\langle in|\, Q(t)\,|in\rangle = \langle in|\, F^{-1}(t,t_i)Q[\delta\psi^I(t),\delta p_\psi^I(t)]F(t,t_i)\,|in\rangle\ . \tag{5.3.47}$$

Using eqs. (5.3.37) and (5.3.43), it is an easy exercise to show that $F(t,t_i)$ satisfies

$$\frac{d}{dt}F(t,t_i) = -iH_{\rm int}[\delta\psi^I(t),\delta p_\psi^I(t);t]\,F(t,t_i)\ , \tag{5.3.48}$$

with $F(t_i,t_i)\equiv 1$. Hence, $F(t,t_i)$ is the unitary evolution operator associated with $H_{\rm int}$, with the interaction Hamiltonian constructed out of interaction picture fields. Eq. (5.3.48) has the following solution

$$F(t,t_i) = T\exp\left(-i\int_{t_i}^{t} H_{\rm int}(t)\mathrm{d}t\right)\ , \tag{5.3.49}$$

where T is the standard time-ordering operator (as required for time-dependent Hamiltonians).

Projection onto the free vacuum. Letting $t_i \to -\infty^+ \equiv -\infty(1+i\epsilon)$, the *in-in* expectation value in (8.2.15) becomes

$$\langle Q\rangle = \langle in|\left(Te^{-i\int_{-\infty^+}^{t} H_{\rm int}(t')\mathrm{d}t'}\right)^\dagger Q^I(t)\,\left(Te^{-i\int_{-\infty^+}^{t} H_{\rm int}(t'')\mathrm{d}t''}\right)|in\rangle\ . \tag{5.3.50}$$

As in the standard *in-out* treatment of quantum field theory in Minkowski space, we have used the $i\epsilon$ prescription to effectively turn off the interaction $H_{\rm int}$ in the infinite past. We can therefore identify the *in* vacuum with the free-field vacuum, $|in\rangle \to |0\rangle$. However, in the *in-in* formalism there is an additional subtlety: the conjugation operator in (5.3.50) acts on the integration limit, leading to an important sign flip

$$\boxed{\langle Q\rangle = \langle 0|\,\bar{T}e^{i\int_{-\infty^-}^{t} H_{\rm int}(t')\mathrm{d}t'}\,Q^I(t)\;Te^{-i\int_{-\infty^+}^{t} H_{\rm int}(t'')\mathrm{d}t''}\,|0\rangle}\ , \tag{5.3.51}$$

where we defined $-\infty^\pm = -\infty(1\mp i\epsilon)$ and $\bar{T}$ denotes anti-time-ordering. The integration contour goes from $-\infty(1-i\epsilon)$ to t (where the correlation function is evaluated) and back to $-\infty(1+i\epsilon)$. Notice that the contour doesn't close. This will imply that the contraction of operators lead to real-valued *Wightman* Green's functions, rather than complex-valued *Feynman* Green's functions.

Perturbative expansion. Finally, correlation functions are then computed perturbatively in $H_{\rm int}$. For instance, to leading order (5.3.51) can be written as

$$\boxed{\langle Q(t)\rangle = -i\int_{-\infty}^{t}\mathrm{d}t'\ \langle 0|\,[Q^I(t),H_{\rm int}(t')]\,|0\rangle}\ . \tag{5.3.52}$$

This key equation allows us to compute bispectra at tree-level. Higher orders in $H_{\rm int}$ can be written as nested commutators

$$i^n\int_{-\infty}^{t}\mathrm{d}t_1\int_{-\infty}^{t_1}\mathrm{d}t_2\cdots\int_{-\infty}^{t_{n-1}}\mathrm{d}t_n\ \langle[H_{\rm int}(t_n),[H_{\rm int}(t_{n-1}),\ldots,[H_{\rm int}(t_1),Q^I(t)]\cdots]]\rangle\ . \tag{5.3.53}$$

5.3.2 Single-Field Inflation

Let us now apply the *in-in* formalism to a few examples in single-field inflation. This means finding the interaction Hamiltonian for the curvature perturbation in a specific inflationary model and applying the master equation (5.3.51).

Slow-Roll Inflation

We start with the action of a scalar field with canonical kinetic term $X \equiv -\frac{1}{2}(\partial_\mu\phi)^2$ and potential $V(\phi)$, minimally coupled to Einstein gravity,

$$S = \int \mathrm{d}^4x\sqrt{-g}\left[\frac{1}{2}R + X - V(\phi)\right] \ , \tag{5.3.54}$$

where $M_{\rm pl} \equiv 1$. Non-Gaussianity arises both from inflaton self-interactions and gravitational non-linearities.[16] We parameterize metric perturbations in the ADM formalism

$$ds^2 = -N^2dt^2 + h_{ij}(dx^i + N^idt)(dx^j + N^jdt) \ . \tag{5.3.55}$$

In these variables the action becomes

$$S = \int \mathrm{d}t\mathrm{d}^3x\sqrt{h}N\left(\frac{1}{2}R_{(3)} + X - V + \frac{1}{2N^2}(E_{ij}E^{ij} + E^2)\right) \ , \tag{5.3.56}$$

where $R_{(3)}[h_{ij}]$ is the three-dimensional Ricci scalar, E_{ij} is the extrinsic curvature,

$$E_{ij} = \frac{1}{2}(\dot{h}_{ij} - \nabla_iN_j - \nabla_jN_i) \ , \tag{5.3.57}$$

and $E \equiv E_{ij}h^{ij}$ is its trace. We work in comoving gauge, $\delta\phi \equiv 0$, and define the curvature perturbation as

$$h_{ij} = a^2e^{2\zeta}\delta_{ij} \ . \tag{5.3.58}$$

It is a straightforward, but tedious task to solve the constraint equations for the Lagrangian multipliers $N = 1 + \delta N$ and $N_i = \partial_i\psi$,

$$\delta N = \frac{\dot{\zeta}}{H} \qquad \text{and} \qquad \psi = -\frac{\zeta}{a^2H} + \chi \ , \quad \text{where} \quad \partial^2\chi = \varepsilon\dot{\zeta} \ . \tag{5.3.59}$$

It is sufficient to solve N and N_i to first order in ζ, since the second-order and third-order perturbations will multiply the first-order and zeroth-order constraint equation, respectively. Substituting the Lagrange multipliers into the action, Maldacena found

$$\mathcal{L}_2 = \varepsilon(\partial_\mu\zeta)^2 \ , \tag{5.3.60}$$

and

$$\mathcal{L}_3 = \varepsilon^2\zeta\dot{\zeta}^2 + \varepsilon^2\zeta(\partial_i\zeta)^2 - 2\varepsilon\dot{\zeta}(\partial_i\zeta)(\partial_i\chi) + 2f(\zeta)\frac{\delta L_2}{\delta\zeta} + \mathcal{O}(\varepsilon^3) \ , \tag{5.3.61}$$

where

$$f(\zeta) \equiv \frac{\eta}{4}\zeta^2 + \cdots \tag{5.3.62}$$

[16] Of course, this is a gauge-dependent statement. For example, in unitary gauge ($\delta\phi = 0$), the inflaton fluctuations are eaten by the metric and the non-Gaussianity is purely in the gravity sector.

and $\frac{\delta L_2}{\delta\zeta}$ is the variation of the quadratic action $L_2 = a^3\mathcal{L}_2$ with respect to ζ. The dots in (5.3.62) represent a large number of terms with derivatives acting on ζ. These terms vanish outside of the horizon and hence don't contribute to the bispectrum. Maldacena showed that the term proportional to $f(\zeta)$ can be removed by a field redefinition,

$$\zeta \to \zeta_n + f(\zeta_n) \ . \tag{5.3.63}$$

This field redefinition has the following effect on the correlation function

$$\langle\zeta(\mathbf{x}_1)\zeta(\mathbf{x}_2)\zeta(\mathbf{x}_3)\rangle = \langle\zeta_n(\mathbf{x}_1)\zeta_n(\mathbf{x}_2)\zeta_n(\mathbf{x}_3)\rangle + \frac{\eta}{2}\left(\langle\zeta_n(\mathbf{x}_1)\zeta_n(\mathbf{x}_2)\rangle\langle\zeta(\mathbf{x}_1)\zeta_n(\mathbf{x}_3)\rangle + \text{cyclic}\right) + \cdots \tag{5.3.64}$$

Expanding the in-in master formula (5.3.51) to first order in $H_{\rm int} = -L_3 + \mathcal{O}(\zeta^4)$, we get

$$\langle\zeta_n^3\rangle = -i\int_{-\infty}^{0} \mathrm{d}t\ \langle 0|[\zeta_n(\mathbf{k}_1)\zeta_n(\mathbf{k}_2)\zeta_n(\mathbf{k}_3), H_{\rm int}(t)]|0\rangle \ . \tag{5.3.65}$$

Before embarking on a lengthy calculation of the bispectrum, it is often advisable to perform an order-of-magnitude estimate of the expected size of the signal, i.e. to do a quick and dirty way to estimate (5.3.65) without explicitly performing the integral. For example, the first term in (5.3.61), can be written as $\int \mathrm{d}t\, H_{\rm int}(t) \subset -\int \mathrm{d}^3x\mathrm{d}\tau a^2\varepsilon^2\zeta(\zeta')^2$. We only need to keep track of factors of H and ε. Any time- and momentum-dependence will work itself out and only contributes to the shape function. Using $a \propto H^{-1}$ and $\zeta \propto \zeta' \propto \sqrt{\Delta_\zeta} \sim H/\sqrt{\varepsilon}$, we estimate that the contribution from the three-point vertex is $\sim H\sqrt{\varepsilon}$. Combining this with estimates for the size of the three external legs, $\zeta^3 \sim H^3\varepsilon^{-3/2}$, we find[17]

$$\langle\zeta^3\rangle = -i\int \mathrm{d}t\,\langle[\zeta^3, H_{\rm int}(t)]\rangle \propto \frac{H^4}{\varepsilon} \propto \mathcal{O}(\varepsilon)\Delta_\zeta^2 \sim f_{\rm NL}\Delta_\zeta^2 \ . \tag{5.3.67}$$

Similar results are obtained for the other two interactions in (5.3.61), $f_{\rm NL} \sim \mathcal{O}(\varepsilon)$. Moreover, it is easy to see that the contribution from the field redefinition in (5.3.64) is $f_{\rm NL} \sim \mathcal{O}(\eta)$. We conclude that the non-Gaussianity in slow-roll inflation is slow-roll surpressed,

$$f_{\rm NL} \sim \mathcal{O}(\varepsilon, \eta)\,. \tag{5.3.68}$$

This small amount of non-Gaussianity will never be observable in the CMB.

To get the full momentum-dependence of the bispectrum we actually have to do some real work and compute the integral in (5.3.65) using the free-field mode functions for ζ. Here, we just cite the final answer,[18]

$$S_{\rm s.r.} = \frac{11}{2}\varepsilon\,\mathcal{S}_\varepsilon + \frac{3}{2}\eta\,\mathcal{S}_\eta \ , \tag{5.3.69}$$

where we have separated the bispectrum into a contribution proportional to ε and a contribution proportional to η,

$$\mathcal{S}_\varepsilon = \frac{1}{11}\left[-\left(\frac{k_1^2}{k_2k_3} + 2\,\text{perms.}\right) + \left(\frac{k_1}{k_2} + 5\,\text{perms.}\right) + \frac{8}{3K}\left(\frac{k_1k_2}{k_3} + 2\,\text{perms.}\right)\right] \ , \tag{5.3.70}$$

$$\mathcal{S}_\eta = \frac{1}{3}\left(\frac{k_1^2}{k_2k_3} + 2\,\text{perms.}\right) \ , \tag{5.3.71}$$

[17] A simple way to remember this back-of-the-envelope technique for estimating non-Gaussianity is

$$\frac{\mathcal{L}_3}{\mathcal{L}_2} \sim f_{\rm NL}\zeta \ . \tag{5.3.66}$$

[18] In §5.3.2 we present a similar calculation in more detail. In that case, the result may actually be observable, so our efforts will be of more direct observational relevance.

where $K \equiv \frac{1}{3}(k_1 + k_2 + k_3)$. This slow-roll shape is well approximated by a superposition of the local and equilateral shapes

$$S_{\text{s.r.}} \propto (6\varepsilon - 2\eta)\, \mathcal{S}_{\text{loc.}} + \frac{5}{3}\varepsilon\, \mathcal{S}_{\text{equil.}} \ . \tag{5.3.72}$$

Like the local shape the slow-roll shape therefore peaks in the squeezed limit. However, the smallness of the slow-roll parameters makes the signal unobservable.

Small Sound Speed

So far this might seem a bit like Sisyphus work since we realized quickly that the non-Gaussianity from slow-roll inflation will never be observable. What are mechanisms that could produce large non-Gaussianity during inflation? This question will concern us in the remainder of these notes.

Higher Derivatives and Radiative Stability

What kind of high-energy effects could deform slow-roll inflation in such a way as to produce large non-Gaussianity without disrupting the inflationary background solution? In effective field theory the effects of high-energy physics are encoded in high-dimension operators for the inflaton lagrangian.[19] Non-derivative operators such as ϕ^n/Λ^{n-4} form part of the inflaton potential and are therefore strongly constrained by the background (see previous section). In other words, the existence of a slow-roll phase requires the non-Gaussianity associated with these operators to be small.[20] This naturally leads us to consider higher-derivative operators of the form $(\partial_\mu\phi)^{2n}/\Lambda^{4n-4}$. These operators don't affect the background, but in principle they could lead to strong interactions. Let us consider the leading correction to the slow-roll lagrangian

$$\mathcal{L} = \mathcal{L}_{\text{s.r.}} + \frac{(\partial_\mu\phi)^4}{8\Lambda^4} \ . \tag{5.3.73}$$

We split the inflaton field into background $\bar{\phi}(t)$ and fluctuations $\varphi(\mathbf{x}, t)$. For $\dot{\bar{\phi}} \ll \Lambda^2$, we can ignore the correction to the quadratic lagrangian for φ,

$$\mathcal{L}_2 \approx -\frac{1}{2}(\partial_\mu\varphi)^2 \ . \tag{5.3.74}$$

We get the cubic lagrangian for φ by evaluating one of the legs of the interaction $(\partial\phi)^4$ on the background $\dot{\bar{\phi}}$,

$$\mathcal{L}_3 = -\frac{\dot{\bar{\phi}}}{2\Lambda^4}\,\dot{\varphi}(\partial_\mu\varphi)^2 \ . \tag{5.3.75}$$

As before, we estimate the size of the non-Gaussianity in the quick and dirty way,

$$f_{\text{NL}} \sim \frac{1}{\zeta}\frac{\mathcal{L}_3}{\mathcal{L}_2} \sim \frac{1}{\zeta}\frac{\dot{\bar{\phi}}\,\dot{\varphi}}{\Lambda^4} \ . \tag{5.3.76}$$

[19] In addition, there could be high-energy modifications to the vacuum state. These effects require a separate discussion.

[20] A possibility to get large non-Gaussianities is to have additional light fields (i.e. fields with mass smaller than H) during inflation. These new degrees of freedom are not constrained by the slow-roll requirements and if their fluctuations are somehow converted into curvature perturbations, these can be much less Gaussian than in single-field slow-roll inflation. We discuss this possibility in the next section.

Using $\dot{\varphi} \sim H\varphi$ and $\zeta = \frac{H}{\dot{\bar{\phi}}}\varphi$, we get

$$f_{\rm NL} \sim \frac{\dot{\bar{\phi}}^2}{\Lambda^4} \ . \tag{5.3.77}$$

We therefore find that we only get significant non-Gaussianity when $\dot{\bar{\phi}} > \Lambda^2$, in which case we can't trust our derivative expansion. In other words, for $\dot{\bar{\phi}} > \Lambda^2$ there is no reason to truncate the expansion at finite order as we did in (7.4.14). Instead, operators of arbitrary dimensions become important in this limit,

$$P(X,\phi) = \sum c_n(\phi)\frac{X^n}{\Lambda^{4n-4}} \ , \qquad \text{where} \quad X \equiv -\frac{1}{2}(\partial_\mu\phi)^2 \ . \tag{5.3.78}$$

As an effective-field theory, eq. (7.4.19) makes little sense when $X > \Lambda^4$. All of the coefficients c_n are radiatively unstable. Hence, if we want to use a theory like (7.4.19) to generate large non-Gaussianity, we require a UV-completion. Interestingly, an example for such a UV-completion exists in string theory. In Dirac-Born-Infeld inflation,

$$P(X,\phi) = \frac{\Lambda^4}{f(\phi)}\sqrt{1 - f(\phi)\frac{X}{\Lambda^4}} - V(\phi) \ , \tag{5.3.79}$$

the form of the action is protected by a higher-dimensional boost symmetry. This symmetry protects eq. (5.3.79) from radiative corrections and allows a predictive inflationary model with large non-Gaussianity. It would be interesting to explore if there are other examples of $P(X)$ theories that are radiatively stable. In the next subsection, we will allow ourselves a bit of artistic freedom and study the phenomenology of general $P(X)$ theories, while ignoring the serious issues they face in explaining radiative stability. However, I emphasize that the problem of radiative stability should not be treated lightly. It is an important requirement of any satisfactory theory.

Non-Gaussianity in $P(X)$ Theories

We consider theories whose action can be written in the following form

$$S = \int d^4x\sqrt{-g}\left[\frac{1}{2}R + P(X,\phi)\right] \ . \tag{5.3.80}$$

We expand both the inflaton and the metric in small fluctuations, i.e. $\phi(\mathbf{x},t) = \bar{\phi}(t) + \varphi(\mathbf{x},t)$ and $g_{\mu\nu} = \bar{g}_{\mu\nu} + \delta g_{\mu\nu}$. We could now go ahead and solve the metric fluctuations in the Einstein constraint equations in terms of the matter fluctuations φ. After some hard work, we would arrive at an action for a single scalar degree of freedom. However, this is unnecessarily hard work in the case we are interested in. Large non-Gaussianities only arise from the higher-derivative inflaton self-interactions, while the effects of mixing with gravity are subdominant. At leading order, we will therefore arrive at the correct result by simply *ignoring the mixing with metric fluctuations* and computing the action for inflaton perturbations φ in an *unperturbed* spacetime $\bar{g}_{\mu\nu}$. In this *decoupling limit*, $g_{\mu\nu} \to \bar{g}_{\mu\nu}$, we find $X \equiv \bar{X} + \delta X$, where

$$\bar{X} = \frac{1}{2}\dot{\bar{\phi}}^2 \qquad \text{and} \qquad \delta X = \dot{\bar{\phi}}\dot{\varphi} - \frac{1}{2}(\partial_\mu\varphi)^2 \ . \tag{5.3.81}$$

Defining $\varphi \equiv \dot{\bar{\phi}}\pi$ and ignoring (slow-roll suppressed) derivatives of $\dot{\bar{\phi}}$, we get

$$\delta X = 2\bar{X}\left[\dot{\pi} - \frac{1}{2}(\partial_\mu\pi)^2\right] \ . \tag{5.3.82}$$

We used this in a Taylor expansion of $P(X)$,

$$P(X) = P(\bar{X}) + P_{,\bar{X}}\delta X + \frac{1}{2}P_{,\bar{X}\bar{X}}(\delta X)^2 + \frac{1}{6}P_{,\bar{X}\bar{X}\bar{X}}(\delta X)^3 + \cdots \tag{5.3.83}$$

In terms of the (rescaled) inflaton fluctuations π this becomes

$$P = \bar{X}P_{,\bar{X}}(\partial_\mu\pi)^2 + 2\bar{X}^2P_{,\bar{X}\bar{X}}\left[\dot{\pi}^2 - \dot{\pi}(\partial_\mu\pi)^2 + \cdots\right] + \frac{4}{3}\bar{X}^3P_{,\bar{X}\bar{X}\bar{X}}\left[\dot{\pi}^3 + \cdots\right] \tag{5.3.84}$$

The background equations of motion relate the first coefficient to the Hubble expansion parameter, $\bar{X}P_{,\bar{X}} = M_{\rm pl}^2\dot{H}$, while the remaining parameters are free, $M_n^4 \equiv \bar{X}^n\frac{d^nP}{d\bar{X}^n}$. Up to cubic order we hence get the following lagrangian

$$\boxed{\mathcal{L} = M_{\rm pl}^2\dot{H}(\partial_\mu\pi)^2 + 2M_2^2\left[\dot{\pi}^2 - \dot{\pi}(\partial_\mu\pi)^2\right] + \frac{4}{3}M_3^4\,\dot{\pi}^3} \,. \tag{5.3.85}$$

A few comments are in order:

1. The M_2 operator leads to a correction of the kinetic term $\dot{\pi}^2$, but *not* of the gradient term $(\partial_i\pi)^2$. Lorentz invariance is broken, space and time are treated differently and the theory for the fluctuations develops a non-trivial sound speed,

$$\mathcal{L}_2 = \frac{M_{\rm pl}^2\dot{H}}{c_s^2}\left[\dot{\pi}^2 - c_s^2(\partial_i\pi)^2\right] \,, \tag{5.3.86}$$

where

$$\frac{1}{c_s^2} = 1 - \frac{2M_2^4}{M_{\rm pl}^2\dot{H}} \,. \tag{5.3.87}$$

For $M_2^4 \gg M_{\rm pl}^2|\dot{H}|$ the sound speed is significantly smaller than the speed of light, $c_s \ll 1$. The free-field mode function derived from (5.3.86) is,

$$\pi_k(\tau) = \pi_k^{(o)}(1 + ikc_s\tau)e^{-ikc_s\tau} \,, \tag{5.3.88}$$

where $\pi_k^{(o)}$ is its superhorizon value

$$\pi_k^{(o)} \equiv \frac{i}{2M_{\rm pl}\sqrt{\epsilon k^3 c_s}} \,. \tag{5.3.89}$$

Hence, the power spectrum for $\zeta = -H\pi$ is

$$P_\zeta = H^2|\pi_k^{(o)}|^2 = \frac{H^2}{4\epsilon c_s M_{\rm pl}^2}\frac{1}{k^3} \,. \tag{5.3.90}$$

2. A non-linearly realized symmetry in (5.3.85) relates a small sound speed (large M_2) to large interactions. In this limit we expect large non-Gaussianities. Our standard estimate confirms this

$$f_{\rm NL} \sim \frac{1}{\zeta}\frac{\mathcal{L}_3}{\mathcal{L}_2} \sim \frac{1}{H\pi}\frac{\dot{\pi}(\partial_i\pi)^2}{\dot{\pi}^2} \sim \frac{1}{c_s^2} \gg 1 \,. \tag{5.3.91}$$

Let us use the *in-in* formalism (with $\psi \to \pi$) to compute the non-Gaussianity in the limit $c_s \ll 1$. At tree level, we use the *in-in* master formula to compute the bispectrum after horizon-crossing ($\tau \to 0$):

$$\lim_{\tau\to 0}\langle \pi^3\rangle = -i\int_{-\infty^+}^{0} \mathrm{d}\tau \langle [\pi^3(0), H_{\rm int}(\tau)]\rangle = (2\pi)^3\delta(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3)B_\pi(k_1,k_2,k_3)\ , \tag{5.3.92}$$

where, at leading order in the interactions, $H_{\rm int} = -\int \mathrm{d}^3x\, a^4 \mathcal{L}_{\rm int}$.

We first consider the $\dot\pi(\partial_i\pi)^2$ interaction. Performing the standard Wick contractions, we find

$$B_{\dot\pi(\partial_i\pi)^2} = 2M_2^4\cdot \pi^{(o)}_{k_1}\pi^{(o)}_{k_2}\pi^{(o)}_{k_3}\int_{-\infty^+}^{0}\frac{\mathrm{d}\tau}{H\tau}(\pi^*_{k_1})'\pi^*_{k_2}\pi^*_{k_3}(\mathbf{k}_2\cdot\mathbf{k}_3) + \text{perms.} + c.c.\ , \tag{5.3.93}$$

where $\mathbf{k}_2\cdot\mathbf{k}_3 = \frac{1}{2}(k_3^2-k_1^2-k_2^2)$. We inserting (5.3.88) for the wavefunctions. The resulting integral converges[21] due to the $i\epsilon$ in the integration limit $-\infty^+ = -\infty(1-i\epsilon)$. The bispectrum for the curvature fluctuation is obtained by a simple rescaling: $B_{\dot\zeta(\partial_i\zeta)^2} = H^3 B_{\dot\pi(\partial_i\pi)^2}$. Using eq. (5.2.12) to extract the amplitude and the shape of the non-Gaussianity, we find

$$f_{\rm NL}^{\dot\zeta(\partial_i\zeta)^2} = \frac{85}{324}\left(1-\frac{1}{c_s^2}\right)\ , \tag{5.3.94}$$

and

$$\mathcal{S}_{\dot\zeta(\partial_i\zeta)^2} \propto \left[\frac{(k_1^2-k_2^2-k_3^2)K}{k_1k_2k_3}\left(-1+\frac{1}{9}\sum_{i>j}\frac{k_ik_j}{K^2}+\frac{1}{27}\frac{k_1k_2k_3}{K^3}\right)+\text{perms.}\right]\ , \tag{5.3.95}$$

where $K\equiv\frac{1}{3}(k_1+k_2+k_3)$.

Similarly, the bispectrum associated with the $\dot\pi^3$ interaction is

$$B_{\dot\pi^3}(k_1,k_2,k_3) = 2M_3^4\cdot\pi^{(o)}_{k_1}\pi^{(o)}_{k_2}\pi^{(o)}_{k_3}\int_{-\infty^+}^{0}\frac{\mathrm{d}\tau}{H\tau}(\pi^*_{k_1})'(\pi^*_{k_2})'(\pi^*_{k_3})' + \text{perms.} + c.c. \tag{5.3.96}$$

Inserting the wavefunctions and performing the integral, we find the following amplitude

$$f_{\rm NL}^{\dot\zeta^3} = \frac{10}{243}\left(1-\frac{1}{c_s^2}\right)\left(\tilde c_3+\frac{3}{2}c_s^2\right)\ , \tag{5.3.97}$$

and the shape

$$\mathcal{S}_{\dot\zeta^3} = \frac{k_1k_2k_3}{K^3}\ . \tag{5.3.98}$$

The parameter $\tilde c_3$ in eq. (5.3.97) is defined via

$$M_3^4 \equiv \tilde c_3\cdot\frac{M_2^4}{c_s^2}\ . \tag{5.3.99}$$

In DBI inflation there is a particular relation between M_2 and M_3, corresponding to

$$\tilde c_3^{\rm DBI} = \frac{3}{2}(1-c_s^2) \approx \frac{3}{2}\ . \tag{5.3.100}$$

Both $\mathcal{S}_{\dot\zeta(\partial_i\zeta)^2}$ and $\mathcal{S}_{\dot\zeta^3}$ are equilateral shapes. However, the shapes are not identical. In fact, for $\tilde c_3 \approx -5.4$ the combined shape is nearly orthogonal to the equilateral shape (see §5.2.5).

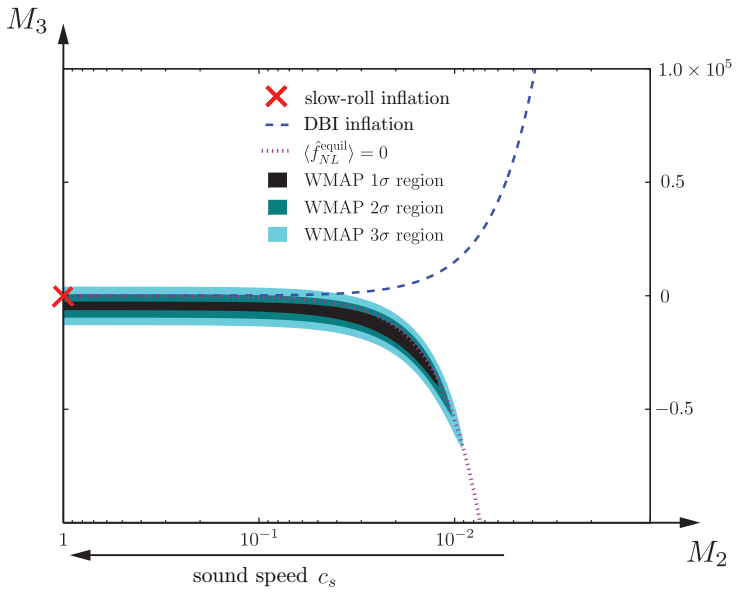


Figure 5.3: *CMB Precision Tests. CMB data constrains the parameters M_2 and M_3 in the effective lagrangian: $\mathcal{L} = M_{\rm pl}^2 \dot{H}(\partial_\mu \pi)^2 + 2M_2^4(\dot{\pi}^2 - \dot{\pi}(\partial_\mu \pi)^2) + \frac{4}{3}M_3^4 \dot{\pi}^3$.*

CMB Precision Tests

The lagrangian (5.3.85) is the minimal deformation of standard single-field slow-roll inflation. Smith et al.[22] used CMB data to constrain its parameters (see fig. 5.3).

A Consistency Relation

We conclude our discussion of non-Gaussianity in single-field inflation with a powerful theorem. Under the assumption of a single field, but making absolutely *no* other assumptions about the inflationary action, Creminelli and Zaldarriaga proved that the following has to hold in the squeezed limit:

$$\lim_{k_1 \to 0} \langle \zeta_{\mathbf{k}_1} \zeta_{\mathbf{k}_2} \zeta_{\mathbf{k}_3} \rangle = (2\pi)^3 \delta(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) \underline{\underline{(1 - n_{\rm s})}} P_\zeta(k_1) P_\zeta(k_3) \,, \tag{5.3.101}$$

i.e. for single-field inflation, the squeezed limit of the three-point function is suppressed by $(1 - n_{\rm s})$ and vanishes for perfectly scale-invariant perturbations. A detection of non-Gaussianity in the squeezed limit can therefore rule out all models of single-field inflation! In particular, this statement is independent of: the form of the potential, the form of the kinetic term (or sound speed) and the initial vacuum state.

Proof:
The squeezed triangle correlates one long-wavelength mode, $k_{\rm L} = k_1$ to two short-wavelength modes, $k_{\rm S} = k_2 \approx k_3$,

$$\langle \zeta_{\mathbf{k}_1} \zeta_{\mathbf{k}_2} \zeta_{\mathbf{k}_3} \rangle \;\to\; \langle (\zeta_{\mathbf{k}_{\rm S}})^2 \zeta_{\mathbf{k}_{\rm L}} \rangle \,. \tag{5.3.102}$$

Modes with longer wavelengths freeze earlier. Therefore, $k_{\rm L}$ will be already frozen outside the horizon when the two smaller modes freeze and acts as a background field for the two short-wavelength modes.

[21] If you Wick rotate, $\tau \to -i\tau$, Mathematica will do the integral.
[22] Smith, Senatore, and Zaldarriaga, (arXiv:0901.2572).

Why should $(\zeta_{\mathbf{k}_S})^2$ be correlated with $\zeta_{\mathbf{k}_L}$? The theorem says that it isn't correlated if ζ_k is precisely scale-invariant, but that the short scale power does get modified by the long-wavelength mode if $n_s \neq 1$. Let's see why. We decompose the evaluation of (5.3.102) into two steps:

i) we calculate the power spectrum of short fluctuations $\langle\zeta_S^2\rangle_{\zeta_L}$ in the presence of a long mode ζ_L;

ii) we then calculate the correlation $\langle(\zeta_{\mathbf{k}_S})^2\zeta_{\mathbf{k}_L}\rangle$.

The calculating is simplest in real-space: The long-wavelength curvature perturbation $\zeta_{\mathbf{k}_L}$ rescales the spatial coordinates within a given Hubble patch (recall that $ds^2 = -dt^2 + a(t)^2 e^{2\zeta(\mathbf{x},t)}d\mathbf{x}^2$). The two-point function $\langle\zeta_S^2\rangle$ will depend on the value of the background fluctuations ζ_L already frozen outside the horizon. In position space the variation of the two-point function given by the long-wavelength fluctuations ζ_L is at linear order

$$\langle\zeta_S^2\rangle_{\zeta_L}(\Delta x) = \langle\zeta_S^2\rangle_0(\Delta x) + \zeta_L \cdot \frac{d}{d\zeta_L}\langle\zeta_S^2\rangle\Big|_0 + \cdots , \tag{5.3.103}$$

where the subscript 0 denotes a quantity in the absence of ζ_L (or in the limit $\zeta_L \to 0$). Using $\frac{d}{d\zeta_L} \to \frac{d}{d\ln\Delta x}$, we find

$$\langle\zeta_S^2\rangle_{\zeta_L} = \langle\zeta_S^2\rangle_0 + \zeta_L \cdot (1 - n_s) \cdot \langle\zeta_S^2\rangle_0 . \tag{5.3.104}$$

To get the three-point function in (5.3.102), we multiply (5.4.113) by ζ_L and average over it

$$\langle\langle\zeta_S^2\rangle_{\zeta_L}\zeta_L\rangle = (1 - n_s)\langle\zeta_L^2\rangle\langle\zeta_S^2\rangle_0 . \tag{5.3.105}$$

Going to Fourier space gives eq. (5.3.101),

$$B_\zeta(k_S, k_S, k_L \to 0) = (1 - n_s)P_\zeta(k_L)P_\zeta(k_S) . \tag{5.3.106}$$

QED.

From the proof it is clear that we didn't have to assume any details about inflation, except that the existence of ζ_L modifies $\langle\zeta_S^2\rangle$. This assumption is satisfied for a single field, but for multiple fields $\langle\zeta_S^2\rangle$ can be modified by other things as well. That is why this is a theorem for *all* single-field models, and can be used to ruled them out!

5.4 Classical Non-Gaussianities

In the previous section we computed the non-Gaussianity generated at horizon crossing. In this section we discuss a second source of non-Gaussianity arising from non-linearities after horizon crossing when all modes have become classical. A convenient way to describe these non-Gaussianities is the δN formalism.

5.4.1 The δN Formalism

Scalar perturbations to the spatial metric on a fixed time slice t can be written as a local perturbation to the scalar factor

$$a(\mathbf{x}, t) \equiv a(t)e^{\psi(\mathbf{x},t)} . \tag{5.4.107}$$

The local number of e-folds of expansion between two time slices t_1 and t_2 is

$$\delta N_{12}(\mathbf{x}) = \int_{t_1}^{t_2} \frac{d\ln a}{dt}\, dt = \psi(\mathbf{x}, t_2) - \psi(\mathbf{x}, t_1) . \tag{5.4.108}$$

Define $\delta N(\mathbf{x}, t)$ as the number of e-folds from a fixed flat slice[23] ($\psi = 0$) to a uniform density slice ($\psi = \zeta$) at time t. Then,

$$\zeta(\mathbf{x}, t) = \delta N(\mathbf{x}, t) \ . \tag{5.4.109}$$

This leads to a simple algorithm to compute the superhorizon evolution of the primordial curvature perturbation ζ: to illustrate the procedure consider a set of scalars ϕ_i. A linear combination of these fields will be the inflaton. The remaining fields are 'isocurvatons'. We assume that all fields have become superhorizon at some initial time. At this time we choose a spatially flat time-slice, on which there are no scalar metric fluctuations, but only fluctuations in the matter fields, $\bar{\phi}_i + \delta\phi_i(\mathbf{x})$. Choose the final time-slice to have uniform density, i.e. the inflaton field is unperturbed and all fluctuations are in the metric and the isocurvatons. Evolve the unperturbed fields $\bar{\phi}_i$ in the initial slice 'classically' to the unperturbed final slice, and denote the corresponding number of e-folds $\bar{N}(\bar{\phi}_i)$. Next, evolve the perturbed initial field configuration $\bar{\phi}_i + \delta\phi_i(\mathbf{x})$ 'classically' to the perturbed final slice. (Notice that in single-field inflation the 'perturbed' final slice is equal to the unperturbed slice, while in general they are different.) The corresponding number of e-folds is $N(\bar{\phi}_i + \delta\phi_i)$. The difference between those two answers,

$$\delta N = N(\bar{\phi}_i + \delta\phi_i) - \bar{N}(\bar{\phi}_i) \ , \tag{5.4.110}$$

is the curvature perturbation ζ, cf. eq. (5.4.109). Taylor expanding, we obtain an expression for ζ in terms of the scalar field fluctuations $\delta\phi_i$ and derivatives of N defined on the initial slice,

$$\zeta = N_i \delta\phi_i + \frac{1}{2} N_{ij} \delta\phi_i \delta\phi_j + \cdots \tag{5.4.111}$$

where $N_i \equiv \partial_i N$, $N_{ij} = \partial_i \partial_j N$, etc. are derivatives evaluated on the initial slice.

We see that there are two sources of non-Gaussianity for ζ:

1. Intrinsic non-Gaussianity of the fields $\delta\phi_i$.

2. Non-Gaussianity resulting from the non-linear relationship between ζ and $\delta\phi_i$.

In this section we are interested in the second effect, so we assume that the inflaton and the isocurvatons are Gaussian on the initial slice. Any non-Gaussianity arises from the subsequence non-linear evolution. Under this assumption, the correlation functions for ζ can be written as

$$\langle \zeta(\mathbf{x}_1)\zeta(\mathbf{x}_2)\rangle = N_i N_j \langle \delta\phi_i(\mathbf{x}_1)\delta\phi_j(\mathbf{x}_2)\rangle \tag{5.4.112}$$

$$\langle \zeta(\mathbf{x}_1)\zeta(\mathbf{x}_2)\zeta(\mathbf{x}_3)\rangle = N_{ij} N_k N_l \langle \delta\phi_i(\mathbf{x}_1)\delta\phi_j(\mathbf{x}_1)\delta\phi_k(\mathbf{x}_2)\delta\phi_l(\mathbf{x}_3)\rangle + 2 \text{ perms.} \tag{5.4.113}$$

The four-point function in (5.4.113) can be written as a product of two-point functions for Gaussian fluctuations. Using the power spectrum for nearly massless fields in de Sitter,

$$\langle \delta\phi_i(\mathbf{k}_1)\delta\phi_j(\mathbf{k}_2)\rangle = \frac{H^2}{2k_1^3}(2\pi)^3 \delta(\mathbf{k}_1 + \mathbf{k}_2)\delta_{ij} \ , \tag{5.4.114}$$

we find

$$B_\zeta(k_1, k_2, k_3) = \frac{N_{ij} N_i N_j}{(N_l^2)^2} \cdot (P(k_1)P(k_2) + 2 \text{ perms.}) \ . \tag{5.4.115}$$

[23] Previously, we called ζ the curvature perturbation in 'comoving' gauge. On superhorizon scales this is equal to the curvature perturbation in 'uniform density' gauge, so we don't introduce a separate variable.

where

$$P_\zeta(k) = N_i^2 \cdot \frac{H^2}{2k^3} \ . \tag{5.4.116}$$

Unsurprisingly, we find a bispectrum of the local shape. (The non-Gaussianity is coming from local non-linearities in real space.) The amplitude is

$$f_{\rm NL}^{\rm loc.} = \frac{5}{6}\frac{N_{ij}N_iN_j}{(N_l^2)^2} \ . \tag{5.4.117}$$

5.4.2 Inhomogeneous Reheating

In this section, we present two applications of the δN formalism. In the curvaton model inhomogeneous reheating occurs because the amplitude of an oscillating field after inflation is modulated by long-wavelength fluctuations set up during inflation. In scenarios of modulated reheating the same long-wavelength fluctuations lead to a spatially dependent inflaton decay rate. Both mechanism lead to a new source for curvature perturbations with potentially large non-Gaussianity.

Modulated Curvaton Oscillations

Let there be a second light field σ during inflation with potential $V(\sigma) = \frac{1}{2}m_\sigma^2\sigma^2$ and mass sufficiently small, $m_\sigma \ll H$, to allow long-wavelength quantum fluctuations $\delta\sigma$ to be unsuppressed. The energy density associated with σ is negligible initially. However, its quantum fluctuations $\delta\sigma$ will become the primary source for the primordial curvature perturbation ζ. For this reason, σ is given the name 'curvaton'. Let $\sigma_\star = \bar\sigma + \delta\sigma(\mathbf{x})$ be the amplitude of the curvaton at horizon-exit. Assume that the fluctuations are Gaussian at horizon-exit. The curvaton amplitude remains frozon until the Hubble parameter drops below m_σ. At that time the curvaton starts oscillation around the minimum of its potential and behaves as pressureless matter. Initially, the universe is radiation-dominated after inflation, but the fraction of the matter energy density associated with the curvaton oscillations quickly grows.[24] We assume that the curvaton is unstable and decays when its decay rate equals the Hubble expansion rate, $H = \Gamma$ (sudden-decay approximation).

We will use the δN formalism to compute the statistics of the resulting curvature perturbations. The initial spatially flat slice t_0 has radiation density $\rho_{\gamma,0}$ and curvaton density $\rho_{\sigma,0}$. The scale factor at that time is a_0. Both components initially redshift as radiation, until $H = m_\sigma$ at time t_1, and the curvaton starts oscillating. The Friedmann equation at t_1 is

$$3M_{\rm pl}^2 m_\sigma^2 = \left(\frac{a_0}{a_1}\right)^4 (\rho_{\gamma,0} + \rho_{\sigma,0}) \approx \left(\frac{a_0}{a_1}\right)^4 \rho_{\gamma,0} \ . \tag{5.4.118}$$

The curvaton now dilutes as matter, until it decays at t_2 when $H = \Gamma$. The Friedmann equation at t_2 is

$$3M_{\rm pl}^2\Gamma^2 = \left(\frac{a_0}{a_2}\right)^4 \rho_{\gamma,0} + \left(\frac{a_0}{a_1}\right)^4 \left(\frac{a_1}{a_2}\right)^3 \rho_{\sigma,0} \ . \tag{5.4.119}$$

Since the time-slice t_2 is defined by a constant Hubble parameter (and hence constant energy density), we automatically are on the final uniform density slice required by the δN formula.[25]

[24] Recall that matter dilutes as a^{-3}, while radiation redshifts as a^{-4}.

[25] After t_2, both components become radiation and the evolution is the same everywhere, so there is no further contribution to δN.

To apply the δN formula for ζ, we need to determine the number of e-folds of expansion from t_0 to t_2 as a function of the initial field value of the curvaton σ. First, we note that (5.4.119) implies

$$e^{-4N} + e^{-3N}\alpha = const. \tag{5.4.120}$$

where $\alpha \equiv (a_0/a_1)(\rho_{\sigma,0}/\rho_{\gamma,0})$. At leading order, a_0/a_1 is independent of σ, cf. eq. (5.4.118). Since $\rho_{\sigma,0} \propto V(\sigma) \propto \sigma^2$, we infer that $\alpha \propto \sigma^2$. Differentiating (5.4.120) with respect to σ and using (5.4.117), we find

$$f_{\rm NL}^{\rm loc.} = \frac{5}{6}\frac{N_{\sigma\sigma}}{N_\sigma^2} = \frac{5}{3x_{\sigma,2}} - \frac{5(4+9x_{\sigma,2})}{12(4+3x_{\sigma,2})} \,, \tag{5.4.121}$$

where we defined $x_{\sigma,2} = e^N\alpha = \rho_{\sigma,2}/\rho_{\gamma,2}$. Sometimes the final answer is written as

$$f_{\rm NL}^{\rm loc.} = \frac{5}{4x} - \frac{5}{3} - \frac{5x}{6} \,, \tag{5.4.122}$$

where

$$x \equiv \frac{3\rho_{\sigma,2}}{4\rho_{\gamma,2} + 3\rho_{\sigma,2}} \,. \tag{5.4.123}$$

This can lead to large non-Gaussianities if $x \ll 1$.

Inhomogeneous Decay Rate

In string theory and supergravity it is natural that coupling constants are functions of additional moduli fields. It is therefore quite natural to imagine that the decay rate of the inflation is modulated by a second light field $\Gamma(\sigma)$. Quantum fluctuations of σ during inflation will lead to fluctuations in the density after inhomogeneous reheating. We will use the δN formalism to compute the resulting non-Gaussianity.

As in the curvaton scenario, the spacetime is practically unperturbed until σ decays at $H = \Gamma$. The δN formula requires the number of e-folds from the intial time, when the spacetime is unperturbed, to a final time after decay, when the energy density has some value. Denoting these times by the subscripts i and f, and labelling the time of decay by 'dec', we can write

$$e^N = \frac{a_f}{a_i} = \frac{a_f}{a_{\rm dec}}\frac{a_{\rm dec}}{a_i} \,. \tag{5.4.124}$$

For purposes of illustration we assume a matter-dominated universe by the time of reheating. We can then take the initial slice to be during matter domination. During matter domination we have $a \propto H^{-2/3}$, while after the decay, during radiation domination, we have $a \propto H^{-1/2}$. Since the decay occurs at $H = \Gamma$, we therefore have

$$\frac{a_{\rm dec}}{a_i} \propto \Gamma^{-2/3} \quad \text{and} \quad \frac{a_f}{a_{\rm dec}} \propto \Gamma^{1/2} \,. \tag{5.4.125}$$

This gives $e^N \propto \Gamma^{-1/6}$, and

$$\frac{\partial N}{\partial \Gamma} = -\frac{1}{6}\frac{1}{\Gamma} \quad \text{and} \quad \frac{\partial^2 N}{\partial^2 \Gamma} = \frac{1}{6}\frac{1}{\Gamma^2} \,. \tag{5.4.126}$$

and hence

$$\delta N = -\frac{1}{6}\left[\frac{\delta\Gamma}{\Gamma} - \frac{1}{2}\left(\frac{\delta\Gamma}{\Gamma}\right)^2\right] \,. \tag{5.4.127}$$

Using $\delta\Gamma = \Gamma'\delta\sigma + \frac{1}{2}\Gamma''(\delta\sigma)^2 + \cdots$ we find

$$f_{\rm NL}^{\rm loc.} = 5\left[\frac{\Gamma''\Gamma}{(\Gamma')^2} - 1\right] \ . \tag{5.4.128}$$

This can lead to large non-Gaussianity if the dependence of the decay rate on the modulus σ is non-linear, $\Gamma''\Gamma \gg (\Gamma')^2$.

5.5 Large-Scale Structure and Non-Gaussianity

`derive scale-dependent bias in peak-background split.`

5.6 Future Prospects

The current WMAP constraints on local, equilateral, and orthogonal non-Gaussianity are

$$f_{\rm NL}^{\rm loc.} = 32 \pm 21 \ , \qquad f_{\rm NL}^{\rm equil.} = 26 \pm 140 \ , \qquad f_{\rm NL}^{\rm ortho.} = -202 \pm 104 \ . \tag{5.6.129}$$

The Planck satellite will improve error bars by at least a factor of 5. The results are expected in less than two years. Stay tuned.

`LSS constraints.`

Part II

The Physics of Inflation

6 Effective Field Theory

In the absence of a fundamental theory of high energy physics including gravitation, we can still make progress. A systematic way of parameterizing our ignorance is to construct effective field theories valid at the energy scale of the experiment. Effective field theory considerations will give us a valuable perspective on the physics of inflation. Before we discuss this in more detail in the next chapter, we will take this chapter as an opportunity to introduce the basic principles of effective field theory.

6.1 Introduction

Few concepts in theoretical physics are more widely applicable than effective field theory (EFT). Progress in diverse fields, such as condensed matter physics, particle physics, and cosmology, has relied heavily on the power and universality of EFT techniques. EFT isolates the relevant low-energy degrees of freedom, while systematically including the effects of high-energy degrees of freedom as non-renormalizable corrections. The low-energy physics is then described by an effective action for the light fields that includes all operators that are consistent with the symmetries of the problem. These notes give a first introduction to this important topic. The text is adapted from the excellent TASI lectures of Witold Skiba[1] and Markus Luty[2].

Phenomena involving distinct energy, or length, scales can often be analyzed by considering one relevant scale at a time. In most branches of physics, this is such an obvious statement that it does not require any justification. The multipole expansion in electrodynamics is useful because the short-distance details of charge distribution are not important when observed from far away. One does not worry about the sizes of planets, or their geography, when studying orbital motions in the Solar System. Similarly, the hydrogen spectrum can be calculated quite precisely without knowing that there are quarks and gluons inside the proton.

Taking advantage of scale separation in quantum field theories leads to effective field theories (EFTs). Fundamentally, there is no difference in how scale separation manifests itself in classical mechanics, electrodynamics, quantum mechanics, or quantum field theory. The effects of large energy scales, or short distance scales, are suppressed by powers of the ratio of scales in the problem. This observation follows from the equations of mechanics, electrodynamics, or quantum mechanics. Calculations in field theory require extra care to ensure that large energy scales decouple.

Decoupling of large energy scales in field theory seems to be complicated by the fact that inte-

[1] W. Skiba, *TASI Lectures on Effective Field Theory*, (arXiv:1006.2142).
[2] M. Luty, *TASI Lectures on Supersymmetry Breaking*, (arXiv:hep-th/0509029).

gration over loop momenta involves all scales. However, this is only a superficial obstacle which is straightforward to deal with in a convenient regularization scheme, for example dimensional regularization. The decoupling of large energy scales takes place in renormalizable quantum field theories whether or not EFT techniques are used. There are many precision calculations that agree with experiments despite neglecting the effects of heavy particles. For instance, the original calculation of the anomalous magnetic moment of the electron, by Schwinger, neglected the one-loop effects arising from weak interactions. Since the weak interactions were not understood at the time, Schwinger's calculation included only the photon contribution, yet it agreed with the experiment within a few percent. Without decoupling, the weak gauge boson contribution would be of the same order as the photon contribution. This would result in a significant discrepancy between theory and experiment and QED would likely have never been established as the correct low-energy theory.

The decoupling of heavy states is, of course, the reason for building high-energy accelerators. If quantum field theories were sensitive to all energy scales, it would be much more useful to increase the precision of low-energy experiments instead of building large colliders. By now, the anomalous magnetic moment of the electron is known to more than ten significant digits. Calculations agree with measurements despite that the theory used for these calculations does not incorporate any TeV-scale dynamics, grand unification, or any notions of quantum gravity.

If decoupling of heavy scales is a generic feature of field theory, why would one consider EFTs? That depends on whether the dynamics at high energy is known and calculable or else the dynamics is either non-perturbative or unknown. If the full theory is known and perturbative, EFTs often simplify calculations. Complex computations can be broken into several easier tasks. If the full theory is not known, EFTs allow one to parameterize the unknown interactions, to estimate the magnitudes of these interactions, and to classify their relative importance. EFTs are applicable to both cases with the known and with the unknown high-energy dynamics because in an effective description only the relevant degrees of freedom are used. The high-energy physics is encoded indirectly though interactions among the light states.

6.2 EFT Fundamentals

6.2.1 Basic Principles

The first step in constructing effective field theories is identifying the relevant degrees of freedom for the measurements of interest. For instance, in particle physics, *light* particles $\phi_{\rm L}$ are included in the effective theory while *heavy* particles $\phi_{\rm H}$ are integrated out. Here, we distinguish light and heavy degrees of freedom on the basis of whether the corresponding particles can be produced on shell at the energies available to the experiment of interest.

Effective Actions

Formally, the heavy fields are *integrated out* by performing a path integral over the heavy degrees of freedom only. This results in an *effective action* for the light degrees of freedom,

$$e^{iS_{\rm eff}(\phi_{\rm L})} \equiv \int \mathcal{D}\phi_{\rm H}\, e^{iS(\phi_{\rm L},\phi_{\rm H})} \ . \tag{6.2.1}$$

In practice, only lattice gauge theorists actually do path integrals, the rest of us converts path integrals into Feynman rules and Feynman diagrams. We will see examples of this perturbative

procedure below. The effective Lagrangian will contain a finite number of renormalizable terms of dimension four or less, and a infinite tower of non-renormalizable[3] terms of dimension larger than four,

$$\mathcal{L}_{\text{eff}}(\phi_{\text{L}}) = \mathcal{L}_{\Delta<4} + \sum_i c_i \, \frac{\mathcal{O}_i(\phi_{\text{L}})}{\Lambda^{\Delta_i-4}} \ , \tag{6.2.2}$$

where Δ_i are the dimensions of the operators $\mathcal{O}_i$. The operators $\mathcal{O}_i$ are made out of the light degrees of freedom ϕ_{L} and are *local* in spacetime (for $E < M_{\text{H}} \sim \Lambda$). In weakly interacting theories the dimensions of the composite operators $\mathcal{O}_i$ is determined simply by adding the dimensions of all fields (and possibly derivatives) making up $\mathcal{O}_i$. The field dimensions are determined from the kinetic terms. In strongly interaction theories the dimensions of operators typically differ significantly from the sum of the constituent field dimensions determined in the free theory.

The sum over higher dimensional operators in eq. (6.2.2) is in principle an infinite sum. In practice, just a few terms are pertinent. Only a finite number of terms needs to be kept because the theory needs to reproduce experiments to finite accuracy and also because the theory can be tailored to specific processes of interest. The higher the dimension of an operator, the smaller its contribution to low-energy observables. Hence, obtaining results to a given accuracy requires a finite number of terms.[4] This is the reason why non-renormalizable theories are just as good as renormalizable theories. An infinite tower of operators is truncated and a finite number of parameters is needed for making predictions, which is exactly the same situation as in renormalizable theories.

It is a simplification to assume that different higher dimensional operators in eq. (6.2.2) are suppressed by the same scale Λ. Different operators can arise from exchanges of distinct heavy states that are not part of the effective theory. The scale Λ is often referred to as the cutoff of the EFT. This is a somewhat misleading term that is not to be confused with a regulator used in loop calculations, for example a momentum cutoff. Λ is related to the scale where the effective theory breaks down. However, dimensionless coefficients do matter. One could redefine Λ by absorbing dimensionless numbers into the definitions of operators. The breakdown scale of an EFT is a physical scale that does not depend on the convention chosen for Λ. This scale could be estimated experimentally by measuring the energy dependence of amplitudes at small momentum. In EFTs, amplitudes grow at high energies and exceed the limits from unitarity at the breakdown scale. It is clear that the breakdown scale is physical since it corresponds to on-shell contributions from heavy states.

Finally, let us remark that terms in the $\mathcal{L}_{\Delta\leq4}$ part of eq. (6.2.2) also receive contributions from the heavy fields. Such contributions may not lead to observable consequences as the coefficients of interactions in $\mathcal{L}_{\Delta\leq4}$ are determined from low-energy observables. In some cases, the heavy fields violate symmetries that would have been present in the full Lagrangian $\mathcal{L}(\phi_{\text{L}}, \phi_{\text{H}} = 0)$ if the heavy fields are neglected. Symmetry-violating effects of heavy fields are certainly observable in $\mathcal{L}_{\Delta\leq4}$.

[3] As we will see, effective theories are just as renormalizable as so-called renormalizable theories.

[4] Not all terms of a given dimension need to be kept. For example, one may be studying $2 \to 2$ scattering. Some operators may contribute only to other scattering processes, for example $2 \to 4$, and may not contribute indirectly through loops to the processes of interest at a given loop order.

Relevant, Irrelevant and Marginal

At energies below Λ, the behavior of the different operators in eq. (6.2.2) is determined by their dimension. We can distinguish three types of operators: *relevant* ($\Delta < 4$), *marginal* ($\Delta = 4$), and *irrelevant* ($\Delta > 4$).

Irrelevant operators deserve their name because their effects are suppressed by powers of E/Λ and are thus small at low energies. Of course, this does not mean that they are not important. In fact, they usually contain the interesting information about the underlying dynamics at higher scales. The point is that irrelevant operators are weak at low energies. In contrast, relevant operators become more important at lower energies. In four-dimensional relativistic field theories, the number of possible relevant operators is rather low: $\Delta = 0$ (the unit operator), $\Delta = 2$ (bosonic mass term ϕ^2), $\Delta = 3$ (fermionic mass term $\bar{\psi}\psi$ and cubic scalar interactions ϕ^3). Finite mass effects are negligible at very high energies ($E \gg m$), however they become relevant when the energy scale is comparable to the mass.

Operators of dimension four are equally important at all energy scales and are called marginal operators. They lie between relevancy and irrelevancy because quantum effects could modify their scaling behaviour on either side. Well-known examples of marginal operators are ϕ^4, the QED and QCD interactions and the Yukawa $\bar{\psi}\psi\phi$ interactions.

In any situation where there is a large mass gap between the energy scale being analyzed and the scale of any heavier states (i.e. $m, E \ll \Lambda$), the effects induced by irrelevant operators are always suppressed by powers of E/Λ, and can usually be neglected. The resulting EFT, which only contains *relevant* and *marginal* operators, is called *renormalizable*. Its predictions are valid up to E/Λ corrections.

These consideration offer a new perspective on the old concept of renormalizability. Take QED as an example. The theory was constructed to be the most general renormalizable ($\Delta \leq 4$) Lagrangian consistent with the electromagnetic $U(1)$ gauge symmetry. However, there exist other interactions (exchanges of Z bosons) that contribute to $e^+e^- \to e^+e^-$, which at low energies ($E \ll M_Z$) generate additional *non-renormalizable* local couplings of higher dimensions. The reason why QED is so successful to describe the low-energy scattering of electrons with positrons is not renormalizability, but rather the fortunate fact that M_Z is very heavy and the leading non-renormalizable contributions are suppressed by E^2/M_Z^2.

Uses of EFTs

Effective Lagrangians are typically used in one of two ways:

1. *The "full theory" $S[\phi_\mathrm{L}, \phi_\mathrm{H}]$ is known*:

   In this case, integrating out the heavy modes gives a simple way of systematically analyzing the effects of the heavy physics on low-energy observables. Because only the low energy scale appears explicitly in the Feynman diagrams of the EFT, amplitudes are easier to calculate and to power count than in the full theory.

   Examples: Integrating out the W, Z bosons from the $SU(2)_L \times U(1)_Y$ Electroweak Lagrangian at energies $E \ll m_{W,Z}$ results in the Fermi theory of weak decays plus corrections suppressed by powers of $E^2/m_{W,Z}^2$. It is easier to analyze electromagnetic or QCD corrections to weak decays in the four-Fermi theory than in the full Electroweak Lagrangian. Another example is the use of EFTs to calculate heavy particle threshold corrections to

low energy gauge couplings. This has applications, for instance, in Grand Unified Theories and in QCD.

2. *The full theory is unknown (or known but strongly coupled)*:

   Whatever the physics at the scale Λ is, by decoupling it must manifest itself at low energies as an effective Lagrangian of the form eq. (6.2.2). If the symmetries (e.g. Poincare, gauge, global) that survive at low energies are known, then the operators $\mathcal{O}_i(x)$ appearing in $\mathcal{L}_{\text{eff}}[\phi_{\text{L}}]$ must respect those symmetries. Thus by writing down an effective Lagrangian containing the most general set of operators consistent with the symmetries, we are necessarily accounting for the UV physics in a completely model independent way.

   Examples: The QCD chiral Lagrangian below the scale $\Lambda_{\chi SB}$ of $SU(3)_L \times SU(3)_R \to SU(3)_V$ chiral symmetry breaking. Here the full theory, QCD, is known, but because $\Lambda_{\chi SB}$ is of order the scale $\Lambda_{\text{QCD}} \sim 1$ GeV, where the QCD coupling is strong, it is impossible to perform the functional integral in eq. (6.2.1) analytically. Another example is general relativity below the scale $M_{\text{pl}} \sim 10^19$ GeV. This theory can be used to calculate, e.g., graviton-graviton scattering at energies $E \ll M_{\text{pl}}$. Above those energies, however, scattering amplitudes calculated in general relativity start violating unitarity bounds, and the effective field theory necessarily breaks down. Thus general relativity is an effective Lagrangian for quantum gravity below the strong coupling scale M_{pl}. Finally, it is believed that the Standard Model itself is an effective field theory below scales of order 1 TeV or so. This scale manifests itself indirectly, in the form of $SU(2)_L \times U(1)_Y$ gauge-invariant operators of dimension $\Delta > 4$ constructed from Standard Model fields, certain linear combinations of which have been constrained experimentally using collider data from the LEP experiments at CERN (c.f. precision electroweak constraints using effective Lagrangians). If there is indeed new physics at the TeV scale, it will be seen directly, at the LHC.

Power Counting

EFTs are based on several systematic expansions. In addition to the usual loop expansion in quantum field theory, one expands in the ratios of energy scales. There can be several scales in the problem: the masses of heavy particles, the energy at which the experiment is done, the momentum transfer, and so on. In an EFT, one can independently keep track of powers of the ratio of scales and of the logarithms of scale ratios. This could be useful, especially when logarithms are large. Ratios of different scales can be kept to different orders depending on the numerical values, which is something that is nearly impossible to do without using EFTs.

When constructing an EFT one needs to be able to formally predict the magnitudes of different operators $\mathcal{O}_i$ in the effective Lagrangian. This is referred to as power counting the terms in the Lagrangian and it allows one to predict how different terms scale with energy. In the simple EFTs discussed here, power counting is the same as dimensional analysis using the natural $\hbar = c = 1$ units, in which [mass] = [length]$^{-1}$. From now on, dimensions will be expressed in the units of [mass], so that energy has dimension 1, while length has dimension -1. The Lagrangian density $\mathcal{L}$ has dimension 4, since the action $S = \int \mathcal{L}\, d^4x$ must be dimensionless. The dimensions of fields are determined from their kinetic energies because in weakly-interacting theories these terms always dominate. The kinetic energy term for a scalar field, $\partial_\mu \phi \partial^\mu \phi$, implies that ϕ has dimension 1, while that of a fermion, $i\,\overline{\psi} \partial\!\!\!/ \psi$, implies that ψ has dimension $\frac{3}{2}$ in four space-time dimensions.

Matching and Running

In the next section, we will consider an example where the full theory is known and weakly coupled. In that case, it is, in principle, possible to perform the functional integral in eq. (6.2.1) and derive the effective theory in terms of the couplings of the full theory. In partice, it is easier to fix the low-energy parameters through a procedure called matching.

The idea is very simple: calculate some observable, for instance the scattering amplitude of the light particles, in two ways. First, calculate the amplitude in the full theory, expanding the result in powers of E/Λ. Then, calculate the same quantity in the effective field theory, adjusting the EFT parameter in order to reproduce the result of the full theory. In general, the coefficients in the effective Lagrangian are calculable as a series expansion in the parameters of the full theory, such as the couplings $\lambda \ll 1$ between the light and the heavy fields. Matching will be performed to a certain order in λ (e.g. tree level, one loop, etc.). Loop graphs in matching calculation typically contain logarithms $\log \Lambda/\mu$. In order to avoid possible large logarithms that could render perturbation theory invalid, one must choose a matching scale μ that is of the same order as the masses of the fields that are being integrated out, i.e. $\mu \sim \Lambda$. To calculate observables at low energy scale $E \ll \Lambda$, it is better to evaluate loop graphs in the EFT at a renormalization point $\mu \sim E$, indicating that in the EFT one should use the couplings $c_i(\mu \sim E)$. These can be obtained in terms of the coefficients $c_i(\mu \sim \Lambda)$ obtained through matching by renormalization group (RG) evolution between the scales Λ and E within the EFT.

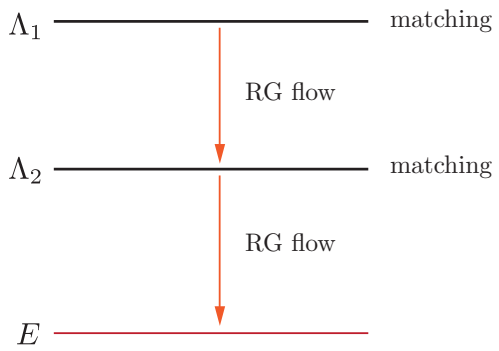


Figure 6.1: *Construction of a low-energy EFT for a theory with scale $\Lambda_1 \gg \Lambda_2 \gg E$.*

For a theory with multiple scales, the procedure is similar. A typical example is shown in fig. 6.1, which depicts the construction of an EFT at a low scale $E \ll \Lambda_2 \ll \Lambda_1$ starting from a theory of light fields φ coupled to heavy fields Φ_1, Φ_2 (masses of order Λ_1, Λ_2). One first constructs an EFT_1 for φ, Φ_2 (regarded as approximately massless) by integrating out the fields Φ_1. This generates a theory defined by its coupling constants at the renormalization scale $\mu \sim \Lambda_1$. This EFT_1 is then used to RG evolve the couplings down to the threshold $\mu \sim \Lambda_2$, at which point Φ_2 is treated as heavy and removed from the theory. This finally generates an EFT_2 for the light fields φ which can be used to calculate at the scale E. In EFT_2, logarithms of E/Λ_2 can be resummed by RG running.

6.2.2 A Toy Model

We will study a simple toy example that nevertheless illustrates all the important lessons you need to know about effective field theory.

Let the full theory be a massless fermion ψ and two massive bosons Φ and φ,

$$\mathcal{L} \;=\; i\bar{\psi}\partial\!\!\!/\psi + \tfrac{1}{2}(\partial_\mu\Phi)^2 - \tfrac{1}{2}M^2\Phi^2 + \tfrac{1}{2}(\partial_\mu\varphi)^2 - \tfrac{1}{2}m^2\varphi^2 - \lambda\Phi\bar{\psi}\psi - \eta\varphi\bar{\psi}\psi \;, \tag{6.2.3}$$

where the parameters λ and η are dimensionless Yukawa couplings. Let us assume $M \gg m$ (for now we won't worry if this hierarchy is natural). We want to determine the effective theory for the light fields: the fermion ψ and the scalar φ. The interactions generated by the exchanges of the heavy field Ψ will be mocked up by new interactions involving the light fields.

6.2.3 Tree-Level Matching

We start with tree-level effects. Consider $\psi\psi \to \psi\psi$ scattering to order (λ^2, η^0) in the coupling constants and keep terms to second order in the external momenta.

As we described above, integrating out the heavy fields is accomplished by comparing, or *matching*, amplitudes in the full UV-theory and the effective IR-theory. Specifically, the process $\psi\psi \to \psi\psi$ has the following amplitude in the UV-theory,

$$\mathcal{A}_{\rm UV} = \bar{u}_3 u_1 \bar{u}_4 u_2 (-i\lambda)^2 \frac{i}{(p_3 - p_1)^2 - M^2} - \{3 \leftrightarrow 4\} \;, \tag{6.2.4}$$

where we defined $u_i \equiv u(p_i)$. The exchange term $\{3 \leftrightarrow 4\}$ comes with a minus sign as required by Fermi statistics. We can expand the propagator in inverse power of the scale M,

$$(-i\lambda)^2 \frac{i}{(p_3 - p_1)^2 - M^2} \approx i\frac{\lambda^2}{M^2}\left(1 + \frac{(p_3 - p_1)^2}{M^2} + \cdots\right) \;. \tag{6.2.5}$$

As advertised, we will neglect terms higher than second order in external momenta, i.e. we construct the effective theory to finite order in the expansion parameter p^2/M^2. The effective theory is therefore only valid at energies below the mass of the heavy scalar, $p^2 \ll M^2$.

Let us find the corresponding effective theory. To zeroth order in external momenta, we can reproduce the $\psi\psi \to \psi\psi$ scattering amplitude by the following four-fermion Lagrangian

$$\mathcal{L}_{\rm eff}^{\{p^0,\lambda^2\}} = i\bar{\psi}\partial\!\!\!/\psi + \frac{c}{2}\bar{\psi}\psi\bar{\psi}\psi \;. \tag{6.2.6}$$

The amplitude calculated with this effective Lagrangian is

$$\mathcal{A}_{\rm IR} = \bar{u}_3 u_1 \bar{u}_4 u_2 (ic) - \{3 \leftrightarrow 4\} \;. \tag{6.2.7}$$

Comparing $\mathcal{A}_{\rm IR}$ to $\mathcal{A}_{\rm UV}$, we find $c = \lambda^2/M^2$. At next order in external momenta, we can write the following Lagrangian[5]

$$\mathcal{L}_{\rm eff}^{\{p^2,\lambda^2\}} = i\bar{\psi}\partial\!\!\!/\psi + \frac{1}{2}\frac{\lambda^2}{M^2}\bar{\psi}\psi\bar{\psi}\psi + d\,\partial_\mu\bar{\psi}\partial^\mu\psi\bar{\psi}\psi \;. \tag{6.2.8}$$

We want to compare the terms obtained from this effective Lagrangian with the amplitude derived from the UV-theory, eq. (6.2.5). The effective Lagrangian needs to be valid both on-shell and off-shell. For the matching, we can use any choice of external momenta that is convenient. In particular, we can choose the momenta to be both on-shell and off-shell. The external particles, here ψ, are identical in the full and effective theories. The choice of external momenta has

[5]There is a second independent two-derivative operator, $\partial_\mu\bar{\psi}\psi\bar{\psi}\partial^\mu\psi$. It turns out that it isn't required to match the UV-theory, so we won't have to consider it.

nothing to do with UV dynamics. In other words, for any momenta below the cutoff the full and effective theories must be identical, thus one is allowed to make opportunistic choices of momenta to simplify calculations.

In the present example, we choose $p_1^2 = p_2^2 = p_3^2 = p_4^2 = 0$. The amplitudes then can only depend on $p_i \cdot p_j$, with $i \neq j$. The effective theory now needs to reproduce the $-2i\frac{\lambda^2}{M^2}\frac{p_1\cdot p_3}{M^2} - \{3 \leftrightarrow 4\}$ part of the amplitude in eq. (6.2.5). The term proportional to d in $\mathcal{L}_{\text{eff}}^{\{p^2,\lambda^2\}}$ gives[6]

$$\mathcal{A}_{\text{IR}} = id(p_1 \cdot p_3 + p_2 \cdot p_4)\, \bar{u}_3 u_1 \bar{u}_4 u_2 - \{3 \leftrightarrow 4\} \ . \tag{6.2.9}$$

Conservation of momentum, $p_1 + p_2 = p_3 + p_4$, implies $p_1 \cdot p_3 = p_2 \cdot p_4$. Hence, we match UV and IR amplitudes for $d = -\lambda^2/M^4$.

At tree-level, the effective Lagrangian is[7]

$$\mathcal{L}_{\text{eff}}^{\{p^2,\lambda^2\}} = i\bar{\psi}\partial\psi + \frac{c}{2}\bar{\psi}\psi\bar{\psi}\psi + d\partial_\mu\bar{\psi}\partial^\mu\psi\bar{\psi}\psi + \tfrac{1}{2}(\partial_\mu\varphi)^2 - \tfrac{1}{2}m^2\varphi^2 - \eta\varphi\bar{\psi}\psi \ , \tag{6.2.10}$$

where

$$c = \frac{\lambda^2}{M^2} \qquad \text{and} \qquad d = -\frac{\lambda^2}{M^4} \ . \tag{6.2.11}$$

6.2.4 RG Running

We should think of the matching as being performed at the scale associated with the mass of the heavy particle, i.e. eq. (6.2.11) should be interpreted as

$$c(\mu = M) = \frac{\lambda^2}{M^2} \qquad \text{and} \qquad d(\mu = M) = -\frac{\lambda^2}{M^4} \ . \tag{6.2.12}$$

At tree level, the couplings don't run and therefore apply unchanged at lower energies. However, at loop level, the couplings will run as we go to low energy scales, e.g. $\mu = m$. Since this RG flow only involves energies below M, it can be computed in the effective theory. Consider, for example, loop contributions to the $\psi\psi \to \psi\psi$ amplitude to lowest order in the momenta and to order $\lambda^2\eta^2$ in the UV coupling constants. Naively, this will lead to a correction to the tree-level amplitude of order $\eta^2/(4\pi)^2$. However, this will not be a good estimate if there are several scales in the problem. For instance, since we have extra light scalars, $m \ll M$, the scattering amplitude could contain *large logs*, such as $\log(M/m)$.

In fact, in an EFT one separates logarithm-enhanced contributions and contributions independent of large logs. The log-independent contributions arise from matching and the log-dependent ones are accounted for by the RG evolution of parameters. By definition, while matching one compares theories with different field contents. This needs to be done using the same renormalization scale in both theories. This so-called matching scale is usually the mass of the heavy particle that is being integrated out. No large logarithms can arise in the process since only one scale is involved. The logs of the matching scale divided by a low-energy scale must be identical in the two theories since the two theories are designed to be identical at low energies. We will illustrate loop matching in the next section. It is very useful that one can compute the matching and running contributions independently. This can be done at different orders in perturbation theory as dictated by the magnitudes of couplings and ratios of scales.

[6] Think Wick contractions.

[7] Note that this effecitve Lagrangian is not complete to order λ^2 and p^2, it only contains all tree-level terms of this order. For example, the Yukawa coupling $\varphi\bar{\psi}\psi$ receives corrections proportional to $\eta\lambda^2$ at one loop.

In our effective theory described in eq. (6.2.10) we need to find the RG equation for the Lagrangian parameters. For concreteness, let us assume we want to know the amplitude at the scale $\mu = m$. Since we will be interested in the momentum-independent part of the amplitude, we can neglect the term proportional to d. By dimensional analysis, the two-derivative term will always be suppressed by $\frac{m^2}{M^2}$ compared to the leading term arising from the non-derivative term.[8]

Let us first compute RG running of the coupling c:

First, a fermion self-energy diagram leads to a one-loop correction to the kinetic term

$$\begin{aligned} &= (-i\eta)^2 \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{i(\not{k}+\not{p})}{(k+p)^2} \frac{i}{k^2-m^2} \\ &= \eta^2 \int \frac{\mathrm{d}^d q}{(2\pi)^d} \int_0^1 \mathrm{d}x \, \frac{\not{q}+(1-x)\not{p}}{[q^2-\Delta^2]^2} \\ &= \frac{i\eta^2}{(4\pi)^2} \frac{1}{\epsilon} \left(\int_0^1 \mathrm{d}x (1-x)\not{p} \right) + \text{ finite} \\ &= \frac{i\eta^2 \not{p}}{2(4\pi)^2} \frac{1}{\epsilon} + \text{ finite} \,, \end{aligned} \tag{6.2.13}$$

where we used Feynman parameters to combine the denominators and shifted the loop momentum $q = k + xp$. We then used the standard result for loop integrals (see eq. (A.44) in Peskin and Schroeder) and expanded $d = 4 - 2\epsilon$. Only the $\frac{1}{\epsilon}$ pole is kept as the finite term does not enter the RG calculation.

Next, we compute the loop corrections due to the four-fermion vertex.

There are six diagrams with a scalar exchange because there are six different pairings of the external lines. The diagrams are depicted in fig. 6.2 and there are two diagrams in each of the three topologies. All of these diagrams are logarithmically divergent in the UV, so we can neglect the external momenta and masses if we are interested in the divergent parts. The divergent terms must be local and therefore be analytic in the external momenta. Extracting positive powers of momenta from a diagram reduces its degree of divergence which is apparent from dimensional analysis. Diagrams (a) in fig. 6.2 are the most straightforward to deal with and the divergent part is easy to extract

$$2(-i\eta)^2 ic \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{i\not{k}}{k^2} \frac{i\not{k}}{k^2} \frac{i}{k^2} = -2c\eta^2 \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{1}{k^4} = -\frac{2ic\eta^2}{(4\pi)^2} \frac{1}{\epsilon} + \text{ finite.} \tag{6.2.14}$$

We did not mention the cross diagrams here, denoted $\{3 \leftrightarrow 4\}$ in the previous section, since they go along for the ride, but they participate in every step. Diagrams (b) in fig. 6.2 require more care as the loop integral involves two different fermion lines. To keep track of this we indicate the external spinors and abbreviate $u(p_i) = u_i$. The result is

$$2(-i\eta)^2 ic \int \frac{\mathrm{d}^d k}{(2\pi)^d} \overline{u}_3 \frac{i\not{k}}{k^2} u_1 \, \overline{u}_4 \frac{-i\not{k}}{k^2} u_2 \frac{i}{k^2} = \frac{ic\eta^2}{2(4\pi)^2} \frac{1}{\epsilon} \overline{u}_3 \gamma^\mu u_1 \, \overline{u}_4 \gamma_\mu u_2 + \text{ finite.} \tag{6.2.15}$$

[8] This reasoning only holds if one uses a mass-independent regulator, like dimensional regularization with minimal subtraction. In dimensional regularization, the renormalization scale μ only appears in logs. In less suitable regularization schemes, the two-derivative term could contribute as much as the non-derivative term as the extra power of $\frac{1}{M^2}$ could become $\frac{\Lambda^2}{M^2}$, where Λ is the regularization scale. With the natural choice $\Lambda \approx M$, the two-derivative term is not suppressed at all. Since the same argument holds for terms with more and more derivatives, all terms would contribute exactly the same and the momentum expansion would be pointless. This is, for example, how hard momentum cut off and Pauli-Villars regulators behave. Such regulators do their job, but they needlessly complicate power counting. From now on, we will only be using dimensional regularization.

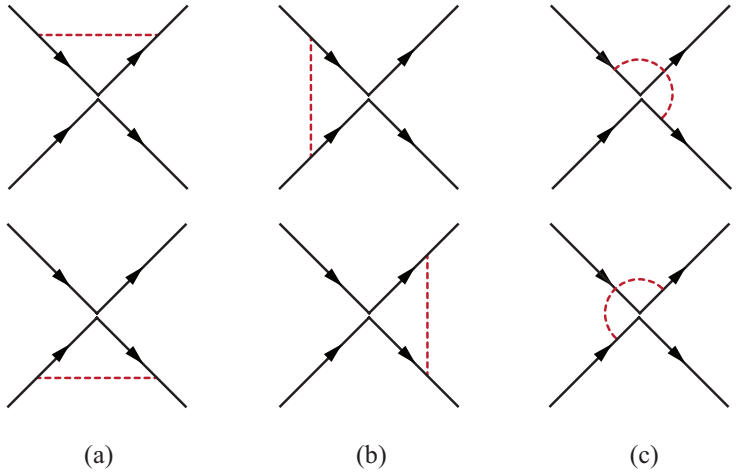


Figure 6.2: *Diagrams contributing to the renormalization of the four-fermion interaction. The (red) dashed lines represent the light scalar φ. The four-fermion vertices are represented by the kinks on the fermion lines. The fermion lines do not touch even though the interaction is point-like. This is not due to limited graphic skills of the author, but rather to illustrate the fermion number flow through the vertices.*

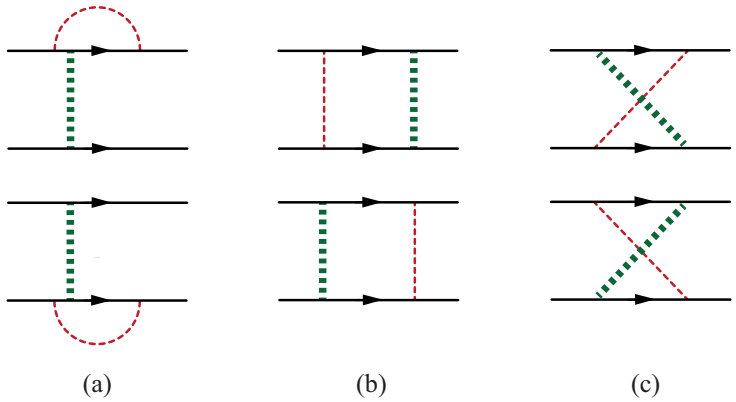


Figure 6.3: *The full theory analogs of the Feynman diagrams in fig. 6.2. The (green) thick dashed lines represents the heavy scalar Φ.*

This divergent contribution is canceled by diagrams (c) in fig. 6.2 because one of the momentum lines carries the opposite sign

$$2(-i\eta)^2 ic \int \frac{\mathrm{d}^d k}{(2\pi)^d}\, \overline{u}_3 \frac{i\not{k}}{k^2} u_1\, \overline{u}_4 \frac{i\not{k}}{k^2} u_2 \frac{i}{k^2} \ . \tag{6.2.16}$$

If the divergent parts of the diagrams (b) and (c) did not cancel this would lead to operator mixing which often takes place among operators with the same dimensions. We will illustrate this shortly.

To calculate the RG equations (RGEs) we can consider just the fermion part of the Lagrangian in eq. (6.2.10) and neglect the derivative term proportional to d. We can think of the original Lagrangian as being expressed in terms of the bare fields and bare coupling constants and rescale $\psi_0 = \sqrt{Z_\psi}\psi$ and $c_0 = c\mu^{2\epsilon} Z_c$. As usual in dimensional regularization, the mass dimensions of the fields depend on the dimension of space-time. In $d = 4 - 2\epsilon$, the fermion dimension is $[\psi] = \frac{3}{2} - \epsilon$ and $[\mathcal{L}] = 4 - 2\epsilon$. We explicitly compensate for this change from the usual 4 space-time dimensions by including the factor $\mu^{2\epsilon}$ in the interaction term. This way, the coupling c

does not alter its dimension when $d = 4 - 2\epsilon$. The Lagrangian is then

$$\begin{aligned}\mathcal{L}_{p^0,\lambda^2\eta^2\log} &= i\overline{\psi}_0 \not{\partial} \psi_0 + \frac{c_0}{2}\,\overline{\psi}_0\psi_0\,\overline{\psi}_0\psi_0 = iZ_\psi\overline{\psi}\not{\partial}\psi + \frac{c}{2}\,Z_c Z_\psi^2 \mu^{2\epsilon}\overline{\psi}\psi\,\overline{\psi}\psi \\ &= i\overline{\psi}\not{\partial}\psi + \mu^{2\epsilon}\frac{c}{2}\,\overline{\psi}\psi\,\overline{\psi}\psi + i(Z_\psi - 1)\overline{\psi}\not{\partial}\psi + \mu^{2\epsilon}\frac{c}{2}(Z_c Z_\psi^2 - 1)\,\overline{\psi}\psi\,\overline{\psi}\psi\ , \end{aligned} \tag{6.2.17}$$

where in the last line we separated the counterterms. We can read off the counterterms from eqs. (6.2.13) and (6.2.14) by insisting that the counterterms cancel the divergences we calculated previously.

$$Z_\psi - 1 = -\frac{\eta^2}{2(4\pi)^2}\frac{1}{\epsilon} \quad \text{and} \quad c(Z_c Z_\psi^2 - 1) = \frac{2c\eta^2}{(4\pi)^2}\frac{1}{\epsilon}\ , \tag{6.2.18}$$

where we used the minimal subtraction (MS) prescription and hence retained only the $\frac{1}{\epsilon}$ poles. Comparing the two equations in (6.2.18), we obtain $Z_c = 1 + \frac{3\eta^2}{(4\pi)^2}\frac{1}{\epsilon}$.

The standard way of computing RGEs is to use the fact that the bare quantities do not depend on the renormalization scale

$$0 = \mu\frac{d}{d\mu}c_0 = \mu\frac{d}{d\mu}(c\mu^{2\epsilon}Z_c) = \beta_c\mu^{2\epsilon}Z_c + 2\epsilon c\mu^{2\epsilon}Z_c + c\mu^{2\epsilon}\mu\frac{d}{d\mu}Z_c, \tag{6.2.19}$$

where $\beta_c \equiv \mu\frac{dc}{d\mu}$. We have $\mu\frac{d}{d\mu}Z_c = \frac{3}{(4\pi)^2}2\eta\beta_\eta\frac{1}{\epsilon}$. Just like we had to compensate for the dimension of c, the renormalized coupling η needs an extra factor of μ^ϵ to remain dimensionless in the space-time where $d = 4 - 2\epsilon$. Repeating the same manipulations we used in Eq. (6.2.19), we obtain $\beta_\eta = -\epsilon\eta - \eta\frac{d\log Z_\eta}{d\log\mu}$. Keeping the derivative of Z_η would give us a term that is of higher order in η as for any Z factor the scale dependence comes from the couplings. Thus, we keep only the first term, $\beta_\eta = -\epsilon\eta$, and get $\mu\frac{d}{d\mu}Z_c = -\frac{6\eta^2}{(4\pi)^2}$. Finally,

$$\beta_c = \frac{6\eta^2}{(4\pi)^2}c\ . \tag{6.2.20}$$

We can now complete our task and compute the low-energy coupling, and thus the scattering amplitude, to the leading log order

$$c(m) = c(M) - \frac{6\eta^2}{(4\pi)^2}c\log\left(\frac{M}{m}\right) = \frac{\lambda^2}{M^2}\left[1 - \frac{6\eta^2}{(4\pi)^2}\log\left(\frac{M}{m}\right)\right]. \tag{6.2.21}$$

Of course, at this point it requires little extra work to re-sum the logarithms by solving the RGEs. First, one needs to solve for the running of η. We will not compute it in detail here, but $\beta_\eta = \frac{5\eta^3}{(4\pi)^2}$. Solving this equation gives

$$\frac{1}{\eta^2(\mu_2)} - \frac{1}{\eta^2(\mu_1)} = \frac{10}{(4\pi)^2}\log\frac{\mu_1}{\mu_2}\ . \tag{6.2.22}$$

Putting the μ dependence of η from Eq. (6.2.22) into Eq. (6.2.20) and performing the integral yields

$$c(m) = C(M)\left(\frac{\eta^2(m)}{\eta^2(M)}\right)^{\frac{3}{5}}, \tag{6.2.23}$$

which agrees with eq. (6.2.21) to the linear order in $\log\left(\frac{M}{m}\right)$.

6.2.5 One-Loop Matching

Construction of effective theories is a systematic process. We saw how RG equations can account for each ratio of scales, and we now increase the accuracy of matching calculations. To improve our $\psi\psi \to \psi\psi$ scattering calculation we compute matching coefficients to one-loop order. As an example, we examine terms proportional to λ^4. This calculation illustrates several important points about matching calculations.

Our starting point is again the full theory with two scalars and a fermion. Since we are only interested in the heavy scalar field, we can neglect the light scalar for the time being and consider

$$\mathcal{L} = i\overline{\psi}\,\partial\!\!\!/\,\psi - \sigma\overline{\psi}\psi + \frac{1}{2}(\partial_\mu\Phi)^2 - \frac{M^2}{2}\Phi^2 - \lambda\,\overline{\psi}\psi\Phi + \mathcal{O}(\varphi)\ . \tag{6.2.24}$$

We added a small mass, σ, for the fermion to avoid possible IR divergences and also to be able to obtain a nonzero answer for terms proportional to $\frac{1}{M^4}$.

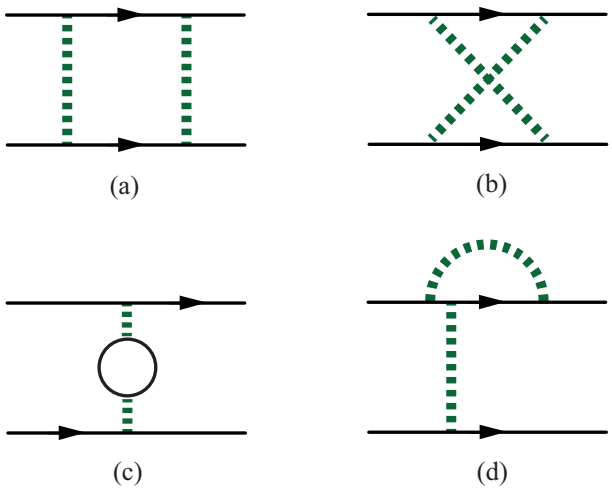


Figure 6.4: Diagrams in the full theory to order λ^4. Diagram (d) stands in for two diagrams that differ only by the placement of the loop.

The diagrams that contribute to the scattering at one loop are illustrated in Fig. 6.4. As we did before, we will focus on the momentum-independent part of the amplitude and we will not explicitly write the terms related by exchange of external fermions. The first diagram gives

$$\begin{aligned}(a) &= (-i\lambda)^4 \int \frac{\mathrm{d}^d k}{(2\pi)^d}\ \overline{u}_3 \frac{i(\not{k}+\sigma)}{k^2-\sigma^2} u_1\, \overline{u}_4 i\frac{i(-\not{k}+\sigma)}{k^2-\sigma^2} u_2 \frac{i^2}{(k^2-M^2)^2} \\ &= \lambda^4 \left[-\overline{u}_3\gamma^\alpha u_1\, \overline{u}_4\gamma^\beta u_2 \int \frac{\mathrm{d}^d k}{(2\pi)^d}\ \frac{k_\alpha k_\beta}{(k^2-\sigma^2)^2(k^2-M^2)^2} \right. \\ &\qquad \left. + \overline{u}_3 u_1\, \overline{u}_4 u_2 \int \frac{\mathrm{d}^d k}{(2\pi)^d}\ \frac{\sigma^2}{(k^2-\sigma^2)^2(k^2-M^2)^2} \right].\end{aligned} \tag{6.2.25}$$

The loop integrals are straightforward to evaluate using Feynman parameterization

$$\frac{1}{(k^2-\sigma^2)^2(k^2-M^2)^2} = 6\int_0^1 \mathrm{d}x\ \frac{x(1-x)}{(k^2-xM^2-(1-x)\sigma^2)^4}\ . \tag{6.2.26}$$

The final result for diagram (a) is

$$\begin{aligned}(a)_{\rm UV} &= \frac{i\lambda^4}{(4\pi)^2}\left[U_V\frac{1}{2}\int_0^1 {\rm d}x\,\frac{x(1-x)}{xM^2+(1-x)\sigma^2}+\sigma^2 U_S\int_0^1 {\rm d}x\,\frac{x(1-x)}{(xM^2+(1-x)\sigma^2)^2}\right]\\ &= \frac{i\lambda^4}{(4\pi)^2}\left[U_V\left(\frac{1}{4M^2}+\frac{\sigma^2}{4M^4}\left(3-2\log\left(\frac{M^2}{\sigma^2}\right)\right)\right)+U_S\frac{\sigma^2}{M^4}\left(\log\left(\frac{M^2}{\sigma^2}\right)-2\right)\right]+\cdots\,,\end{aligned} \tag{6.2.27}$$

where we abbreviated $U_S = \overline{u}_3 u_1\,\overline{u}_4 u_2$, $U_V = \overline{u}_3\gamma^\alpha u_1\,\overline{u}_4\gamma_\alpha u_2$, and in the last line omitted terms of order $\frac{1}{M^6}$ and higher. The subscript $(...)_{\rm UV}$ stands for the full theory, We will denote the corresponding amplitudes in the effective theory with the subscript $(...)_{\rm IR}$. The cross box amplitude (b) is nearly identical, except for the sign of the momentum in one of the fermion propagators

$$(b)_{\rm UV} = \frac{i\lambda^4}{(4\pi)^2}\left[-U_V\left(\frac{1}{4M^2}+\frac{\sigma^2}{4M^4}\left(3-2\log\left(\frac{M^2}{\sigma^2}\right)\right)\right)+U_S\frac{\sigma^2}{M^4}\left(\log\left(\frac{M^2}{\sigma^2}\right)-2\right)\right]+\cdots\,. \tag{6.2.28}$$

Diagrams (c) and (d) are even simpler to evaluate, but they are divergent:

$$(c)_{\rm UV} = -4\frac{i\lambda^4}{(4\pi)^2}\frac{\sigma^2}{M^4}U_S\left[3\frac{1}{\bar\epsilon}+3\log\left(\frac{\mu^2}{\sigma^2}\right)+1\right]+\cdots\,, \tag{6.2.29}$$

where $\frac{1}{\bar\epsilon} = \frac{1}{\epsilon}-\gamma+\log(4\pi)$. Here, μ is the regularization scale and it enters since coupling λ carries a factor of μ^ϵ in dimensional regularization. The four Yukawa couplings give $\lambda^4\mu^{4\epsilon}$. However, $\mu^{2\epsilon}$ should be factored out of the calculation to give the proper dimension of the four-fermion coupling, while the remaining $\mu^{2\epsilon}$ is expanded for small ϵ and yields $\log(\mu^2)$. In the following expression a factor of two is included to account for two diagrams

$$(d)_{\rm UV} = -2\frac{i\lambda^4}{(4\pi)^2M^2}U_S\left[\frac{1}{\bar\epsilon}+1+\log\left(\frac{\mu^2}{M^2}\right)+\frac{\sigma^2}{M^2}\left(2-3\log\left(\frac{M^2}{\sigma^2}\right)\right)\right]+\cdots\,. \tag{6.2.30}$$

The sum of all of these contributions is

$$(a{+}\ldots{+}d)_{\rm UV} = \frac{2i\lambda^4U_S}{(4\pi)^2M^2}\left[-\frac{1}{\bar\epsilon}-1-\log\left(\frac{\mu^2}{M^2}\right)+\frac{\sigma^2}{M^2}\left(-\frac{6}{\bar\epsilon}-6\log\left(\frac{\mu^2}{\sigma^2}\right)-6+4\log\left(\frac{M^2}{\sigma^2}\right)\right)\right]. \tag{6.2.31}$$

We also need the fermion two-point function in order to calculate the wave function renormalization in the effective theory. The calculation is identical to that in eq. (6.2.13). We need the finite part as well. The amplitude linear in momentum is

$$i\not{p}\,\frac{\lambda^2}{2(4\pi)^2}\left(\frac{1}{\bar\epsilon}+\log\left(\frac{\mu^2}{M^2}\right)+\frac{1}{2}+\cdots\right)\,. \tag{6.2.32}$$

It is time to calculate in the effective theory. The effective theory has a four-fermion interaction that was induced at tree level. Again, we neglect the light scalar φ as it does not play any role in our calculation. The effective Lagrangian is

$$\mathcal{L} = iz\overline{\psi}\not{\partial}\psi-\sigma\overline{\psi}\psi+\frac{c}{2}\,\overline{\psi}\psi\,\overline{\psi}\psi\,. \tag{6.2.33}$$

We established that at tree level, $c = \frac{\lambda^2}{M^2}$, but do not yet want to substitute the actual value of c as not to confuse the calculations in the full and effective theories. To match the amplitudes

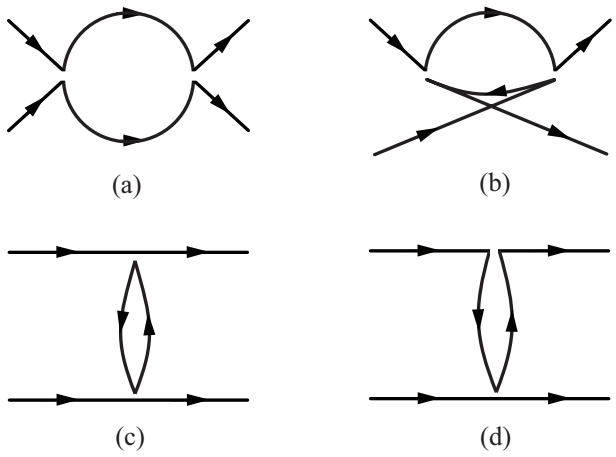


Figure 6.5: Diagrams in the effective theory to order c^2. Diagram (d) stands in for two diagrams that are related by an upside-down reflection. As we drew in Fig. 6.2, the four-fermion vertices are not exactly point-like, so one can follow each fermion line.

we also need to compute one-loop scattering amplitude in the effective theory. The two-point amplitude for the fermion kinetic energy vanishes in the effective theory. The four-point diagrams in the effective theory are depicted in fig. 6.5. Diagrams in an effective theory have typically higher degrees of UV divergence as they contain fewer propagators. For example, diagram $(a)_{\rm IR}$ is quadratically divergent, while $(a)_{\rm UV}$ is finite. This is not an obstacle. We simply regulate each diagram using dimensional regularization.

Exactly like in the full theory, the fermion propagators in diagrams $(a)_{\rm IR}$ and $(b)_{\rm IR}$ have opposite signs of momentum, thus the terms proportional to U_V cancel. The parts proportional to U_S are the same and the sum of these diagrams is

$$(a+b)_{\rm IR} = 2\frac{ic^2\sigma^2}{(4\pi)^2}U_S\left[\frac{1}{\bar{\epsilon}} + \log\left(\frac{\mu^2}{\sigma^2}\right)\right] + \dots . \tag{6.2.34}$$

If one was careless with drawing these diagrams, one might think that there is a closed fermion loop and assign an extra minus sign. However, the way of drawing the effective interactions in fig. 6.5 makes it clear that the fermion line goes around the loop without actually closing. Diagram $(c)_{\rm IR}$ is identical to its counterpart in the full theory. Since we are after the momentum-independent part of the amplitude, the heavy scalar propagators in $(c)_{\rm UV}$ were simply equal to $\frac{-i}{M^2}$. Therefore,

$$(c)_{\rm IR} = -4\frac{ic^2\sigma^2}{(4\pi)^2}U_S\left[3\frac{1}{\bar{\epsilon}} + 3\log\left(\frac{\mu^2}{\sigma^2}\right) + 1\right] + \dots . \tag{6.2.35}$$

As in the full theory, $(d)_{\rm IR}$ includes a factor of two for two diagrams

$$(d)_{\rm IR} = 2\frac{ic^2\sigma^2}{(4\pi)^2}U_S\left[3\frac{1}{\bar{\epsilon}} + 3\log\left(\frac{\mu^2}{\sigma^2}\right) + 1\right] + \dots . \tag{6.2.36}$$

The sum of these diagrams is

$$(a+\ldots+d)_{\rm IR} = -\frac{2ic^2\sigma^2}{(4\pi)^2}U_S\left[\frac{2}{\bar{\epsilon}} + 2\log\left(\frac{\mu^2}{\sigma^2}\right) + 1\right]. \tag{6.2.37}$$

Of course, we should set $c = \frac{\lambda^2}{M^2}$ at this point.

Before we compare the results let us make two important observations. There are several logs in the amplitudes. In the full theory, $\log(\frac{\mu^2}{M^2})$, $\log(\frac{\mu^2}{\sigma^2})$ and $\log(\frac{M^2}{\sigma^2})$ appear, while in the effective theory only $\log(\frac{\mu^2}{\sigma^2})$ shows up. Interestingly, comparing the full and effective theories diagram by diagram, the corresponding coefficients in front of $\log(\sigma^2)$ are identical. This means that $\log(\sigma^2)$ drops out of the difference between the full and effective theories so $\log(\sigma^2)$ never appears in the matching coefficients. It had to be this way. We already argued that the two theories are identical in the IR, so non-analytic terms depending on the light fields must be the same. This would hold for all other quantities in the low-energy theory, for instance for terms that depend on the external momenta. This correspondence between logs of low-energy quantities does not have to happen, in general, diagram by diagram, but it has to hold for the entire calculation. This provides a useful check on matching calculations. When the full and effective theory are compared, the only log that turns up is the $\log(\frac{\mu^2}{M^2})$. This is good news as it means that there is only one scale in the matching calculation and we can minimize the logs by setting $\mu = M$.

The $\frac{1}{\bar{\epsilon}}$ poles are different in the full and effective theories as the effective theory diagrams are more divergent. We simply add appropriate counterterms in the full and the effective theories to cancel the divergences. The counterterms in the two theories are not related. We compare the renormalized, or physical, scattering amplitudes and make sure they are equal. We are going to use the $\overline{MS}$ prescription and the counterterms will cancel just the $\frac{1}{\bar{\epsilon}}$ poles. It is clear that since the counterterms differ on the two sides, the coefficients in the effective theory depend on the choice of regulator. Of course, physical quantities will not depend on the regulator.

Setting $\mu = M$, the difference between eqs. (6.2.31) and (6.2.37) gives

$$c(\mu = M) = \frac{\lambda^2}{M^2} - \frac{2\lambda^4}{(4\pi)^2 M^2} - \frac{10\lambda^4\sigma^2}{(4\pi)^2 M^4}. \tag{6.2.38}$$

To reproduce the two-point function in the full theory we set $z = 1 + \frac{\lambda^2}{4(4\pi)^2}$ in the $\overline{MS}$ prescription since there are no contributions in the effective theory. To obtain physical scattering amplitude, the fermion field needs to be canonically normalized by rescaling $\sqrt{z}\psi \to \psi_{\text{canonical}}$. This rescaling gives an additional contribution to the $\frac{\lambda^4}{(4\pi)^2 M^2}$ term in the scattering amplitude from the product of the tree-level contribution and the wave function renormalization factor. Without further analysis, it is not obvious that it is consistent to keep the last term in the expression for $c(\mu = M)$. One would have to examine if there are any other terms proportional to $\frac{1}{M^4}$ that were neglected. For example, the momentum-dependent operator proportional to d in Eq. (**??**) could give a contribution of the same order when the RG running in the effective theory is included. Such contribution would be proportional to $\frac{\lambda^2\eta^2\sigma^2}{(4\pi)^2 M^4}\log(\frac{M^2}{m^2})$. There can also be contributions to the fermion two-point function arising in the full theory from the heavy scalar exchange. We were originally interested in a theory with massless fermions which means that $\sigma = 0$. It was a useful detour to do the matching calculation including the $\frac{1}{M^4}$ terms as various logs and UV divergences do not fully show up in this example at the $\frac{1}{M^2}$ order.

We calculated the scattering amplitudes arising from the exchanges of the heavy scalar. In the calculation of the $\psi\psi \to \psi\psi$ scattering cross section, both amplitudes coming from the exchanges of the heavy and light scalars have to be added. These amplitudes depend on different coupling constants, but they can be difficult to disentangle experimentally since the measurements are done at low energies. The amplitude associated with the heavy scalar is measurable only if the mass and the coupling of the light scalar can be inferred. This can be accomplished, for example, if the light scalar can be produced on-shell in the s channel. Near the resonance

corresponding to the light scalar, the scattering amplitude is dominated by the light scalar and its mass and coupling can be determined. Once the couplings of the light scalar are established, one could deduce the amplitude associated with the heavy scalar by subtracting the amplitude with the light scalar exchange. If the heavy and light states did not have identical spins one could distinguish their contributions more easily as they would give different angular dependence of the scattering cross section.

6.2.6 Naturalness

Integrating out a fermion in the Yukawa theory emphasizes several important points. We are going to study the same "full" Lagrangian again, but this time assume that the fermion is heavy and the scalar φ remains light

$$\mathcal{L} = i\overline{\psi}\partial\!\!\!/\psi - M\overline{\psi}\psi + \frac{1}{2}(\partial_\mu\varphi)^2 - \frac{m^2}{2}\varphi^2 - \eta\,\overline{\psi}\psi\varphi\ , \tag{6.2.39}$$

where $M \gg m$. We will integrate out ψ and keep φ in the effective theory. As we did earlier, we have neglected the potential for φ assuming that it is zero. There are no tree-level diagrams involving fermions ψ in the internal lines only. We are going to examine diagrams with two scalars and four scalars for illustration purposes. The diagrams resemble those of the Coleman-Weinberg effective potential calculation, but we do not necessarily neglect external momenta. The momentum dependence could be of interest. The two point function gives

$$\begin{aligned} &= (-1)(-i\eta\mu^\epsilon)^2 \int \frac{\mathrm{d}^d k}{(2\pi)^d}\ i^2 \frac{\mathrm{Tr}[(k\!\!\!/ + p\!\!\!/ + M)(k\!\!\!/ + M)]}{[(k+p)^2 - M^2](k^2 - M^2)} \\ &= -\frac{4i\eta^2}{(4\pi)^2}\left[\left(\frac{3}{\overline{\epsilon}} + 1 + 3\log\left(\frac{\mu^2}{M^2}\right)\right)\left(M^2 - \frac{p^2}{6}\right) + \frac{p^2}{2} - \frac{p^4}{20M^2} + \cdots\right], \end{aligned} \tag{6.2.40}$$

where we truncated the momentum expansion at order p^4. The four-point amplitude, to the lowest order in momentum is

$$= -\frac{8i\eta^4}{(4\pi)^2}\left[3\left(\frac{1}{\overline{\epsilon}} + \log\left(\frac{\mu^2}{M^2}\right)\right) - 8 + \cdots\right]. \tag{6.2.41}$$

There are no logarithms involving m^2 or p^2 in eqs. (6.2.40) and (6.2.41). Our effective theory at the tree level contains a free scalar field only, so in that effective theory there are no interactions and no loop diagrams. Thus, logarithms involving m^2 or p^2 do not appear because they could not be reproduced in the effective theory. Setting $\mu = M$ and choosing the counterterms to cancel the $\frac{1}{\overline{\epsilon}}$ poles we can read off the matching coefficients in the scalar theory

$$\mathcal{L} = \left(1 - \frac{4\eta^2}{3(4\pi)^2}\right)\frac{(\partial_\mu\varphi)^2}{2} - \left(m^2 + \frac{4\eta^2 M^2}{(4\pi)^2}\right)\frac{\varphi^2}{2} + \frac{\eta^2}{5(4\pi)^2 M^2}\frac{(\partial^2\varphi)^2}{2} + \frac{64\eta^2}{(4\pi)^2}\frac{\varphi^4}{4!} + \cdots \tag{6.2.42}$$

To obtain physical scattering amplitudes one needs to absorb the $1 - \frac{4\eta^2}{3(4\pi)^2}$ factor in the scalar kinetic energy, so the field is canonically normalized. The scalar effective Lagrangian in Eq. (6.2.42) is by no means a consistent approximation. For example, we did not calculate the tadpole diagram and did not calculate the diagram with three scalar fields. Such diagrams do not vanish since the Yukawa interaction is not symmetric under $\varphi \to -\varphi$. There are no new features in those calculations so we skipped them.

The scalar mass term, $m^2 + \frac{4\eta^2 M^2}{(4\pi)^2}$, contains a contribution from the heavy fermion. If the sum $m^2 + \frac{4\eta^2 M^2}{(4\pi)^2}$ is small compared to $\frac{4\eta^2 M^2}{(4\pi)^2}$ one calls the scalar "light" compared to the heavy mass scale M. This requires a cancellation between m^2 and $\frac{4\eta^2 M^2}{(4\pi)^2}$. Cancellation happens when the two terms are of opposite signs and close in magnitude, yet their origins are unrelated. No symmetry of the theory can relate the tree-level and the loop-level terms. If there was a symmetry that ensured the tree-level and loop contributions are equal in magnitude and opposite in sign, then small breaking of such symmetry could make make the sum $m^2 + \frac{4\eta^2 M^2}{(4\pi)^2}$ small. But no symmetry is present in our Lagrangian. This is why light scalars require a tuning of different terms unless there is a mechanism protecting the mass term, for example the shift symmetry or supersymmetry.

The sensitivity of the scalar mass term to the heavy scales is often referred to as the quadratic divergence of the scalar mass term. When one uses mass-dependent regulators, the mass terms for scalar fields receive corrections proportional to $\frac{\Lambda^2}{(4\pi)^2}$. Having light scalars makes fine tuning necessary to cancel the large regulator contribution. There are no quadratic divergences in dimensional regularization, but the fine tuning of scalar masses is just the same. In dimensional regularization, the scalar mass is quadratically sensitive to heavy particle masses. This is a much more intuitive result compared to the statement about an unphysical regulator. Fine tuning of scalar masses would not be necessary in dimensional regularization if there were no heavy particles. For example, if the Standard Model (SM) was a complete theory there would be no fine tuning associated with the Higgs mass. Perhaps the SM is a complete theory valid even beyond the grand unification scale, but there is gravity and we expect Planck-scale particles in any theory of quantum gravity. Another term used for the fine tuning of the Higgs mass in the SM is the hierarchy problem. Having a large hierarchy between the Higgs mass and other large scales requires fine tuning, unless the Higgs mass is protected by symmetry.

It is apparent from our calculation that radiative corrections generate all terms allowed by symmetries. Even if zero at tree level, there is no reason to assume that the potential for the scalar field vanishes. The potential is generated radiatively. We obtained nonzero potential in the effective theory when we integrated out a heavy fermion. However, generation of terms by radiative corrections is not at all particular to effective theory. The RG evolution in the full theory would do the same. We saw another example of this in §6.2.4, where an operator absent at one scale was generated radiatively. Therefore, having terms smaller than the sizes of radiative corrections requires fine tuning. A theory with all coefficients whose magnitudes are not substantially altered by radiative corrections is called technically natural. Technical naturalness does not require that all parameters are of the same order, it only implies that none of the parameters receives radiative corrections that significantly exceed its magnitude. As our calculation demonstrated, a light scalar that is not protected by symmetry is not technically natural.

Naturalness is a stronger criterion. Dirac's naturalness condition is that all dimensionless coefficients are of order one and the dimensionful parameters are of the same magnitude. A weaker naturalness criterion, due to 't Hooft, is that small parameters are natural if setting a small parameter to zero enhances the symmetry of the theory. Technical naturalness is yet a weaker requirement. The relative sizes of terms are dictated by the relative sizes of radiative corrections and not necessarily by symmetries, although symmetries obviously affect the magnitudes of radiative corrections. Technical naturalness has to do with how perturbative field theory works.

6.2.7 Summary

We have constructed several effective theories so far. It is a good moment to pause and review the observations we made. To construct an EFT one needs to identify the light fields and their symmetries, and needs to establish a power counting scheme. If the full theory is known then an EFT is derived perturbatively as a chain of matching calculations interlaced by RG evolutions. Each heavy particle is integrated out and new effective theory matched to the previous one, resulting in a tower of effective field theories. Consecutive ratios of scales are accounted for by the RG evolution.

This is a systematic procedure which can be carried out to the desired order in the loop expansion. Matching is done order by order in the loop expansion. When two theories are compared at a given loop order, the lower order results are included in the matching. For example, in §6.2.5 we calculated loop diagrams in the effective theory including the effective interaction we obtained at the tree level. At each order in the loop expansion, the effective theory valid below a mass threshold is amended to match the results valid just above that threshold. Matching calculations do not depend on any light scales and if logs appear in the matching calculations, these have to be logs of the matching scale divided by the renormalization scale. Such logs can be easily minimized to avoid spoiling perturbative expansion. The two theories that are matched across a heavy threshold have in general different UV divergences and therefore different counterterms.

EFTs naturally contain higher-dimensional operators and are therefore non-renormalizable. In practice, this is of no consequence since the number of operators, and therefore the number of parameters determined from experiment, is finite. To preserve power counting and maintain consistent expansion in the inverse of large mass scales one needs to employ a mass-independent regulator, for instance dimensional regularization. Consequently, the renormalization scale only appears in dimensionless ratios inside logarithms and so it does not alter power counting. Contributions from the heavy fields do not automatically decouple when using dimensional regularization, thus decoupling should be carried out explicitly by constructing effective theories.

Large logarithms arise from the RG running only as one relates parameters of the theory at different renormalization scales. The field content of the theory does not change while its parameters are RG evolved. However, distinct operators of the same dimension can mix with one another. The RG running and matching are completely independent and can be done at unrelated orders in perturbation theory. The magnitudes of coupling constants and the ratios of scales dictate the relative sizes of different contributions and dictate to what orders in perturbation theory one needs to calculate. A commonly repeated phrase is that two-loop running requires one-loop matching. This is true when the logarithms are very large, for example in grand unified theories. The $\log(M_{\rm GUT}/M_{\rm weak})$ is almost as large as $(4\pi)^2$, so the logarithm compensates the loop suppression factor. This is not the case for smaller ratios of scales.

The contributions of the heavy particles to an effective Lagrangian appear in both renormalizable terms and in higher dimensional terms. For the renormalizable terms, the contributions from heavy fields are often unobservable as the coefficients of the renormalizable terms are determined from low-energy experiments. The contributions of the heavy fields simply redefine the coefficients that were determined from experiments instead of being predicted by the theory. The coefficients of the higher-dimensional operators are suppressed by inverse powers of the heavy masses. As one increases the masses of the heavy particles, their effects diminish. This typical situation is referred to as the decoupling of heavy fields.

When the high-energy theory is not known, or it is not perturbative, one still benefits from constructing an EFT. One can power count the operators and then enumerate the pertinent operators to the desired order. One cannot calculate the coefficients, but one can estimate them. In a perturbative theory, explicit examples tell us what magnitudes of coefficients to expect at any order of the loop expansion.

6.3 The Standard Model as an Effective Theory*

We are now ready to start building effective field theory models. If we believe in the naturalness principle articulated in the previous section, then the models should be defined by specifying the particle content and the symmetries of the theory. Then we should write down all possible couplings consistent with the symmetries.

6.3.1 The Standard Model

Let us apply these ideas to the standard model. The standard model is defined to be a theory with gauge group

$$SU(3)_C \times SU(2)_W \times U(1)_Y. \tag{6.3.43}$$

The fermions of the standard model can be written in terms of 2-component Weyl spinor fields as

$$Q^i \sim (\mathbf{3}, \mathbf{2})_{+\frac{1}{6}}, \tag{6.3.44}$$

$$(u^c)^i \sim (\bar{\mathbf{3}}, \mathbf{1})_{-\frac{2}{3}}, \tag{6.3.45}$$

$$(d^c)^i \sim (\bar{\mathbf{3}}, \mathbf{1})_{+\frac{1}{3}}, \tag{6.3.46}$$

$$L^i \sim (\mathbf{1}, \mathbf{2})_{-\frac{1}{2}}, \tag{6.3.47}$$

$$(e^c)^i \sim (\mathbf{1}, \mathbf{1})_{+1}, \tag{6.3.48}$$

where $i = 1, 2, 3$ is a generation index. In addition, the model contains a single scalar multiplet

$$H \sim (1, \mathbf{2})_{+\frac{1}{2}}. \tag{6.3.49}$$

According to the ideas above, we must now write the most general interactions allowed by the symmetries. The most important interactions are the marginal and relevant ones. The marginal interactions include kinetic terms for the Higgs field, the fermion fields, and the gauge fields:

$$\mathcal{L}_{\text{kinetic}} = (D^\mu H)^\dagger D_\mu H + Q_i^\dagger i\tilde{\sigma}^\mu D_\mu Q_i + \cdots - \tfrac{1}{4} B^{\mu\nu} B_{\mu\nu} + \cdots \tag{6.3.50}$$

Note that these include the gauge self interactions. Also marginal is the quartic interaction for the Higgs

$$\Delta\mathcal{L}_{\text{quartic}} = -\frac{\lambda}{4}(H^\dagger H)^2 \tag{6.3.51}$$

and Yukawa interactions:

$$\Delta\mathcal{L}_{\text{Yukawa}} = (y_u)_{ij} Q^i H (u^c)^j + (y_d)_{ij} Q^i H^\dagger (d^c)^j + (y_e)_{ij} L^i H^\dagger (e^c)^j. \tag{6.3.52}$$

Note that the Yukawa interactions are the only interactions that break a $SU(3)^5$ global symmetry that would otherwise act on the generation indices of the fermion fields. This means that the

Yukawa interactions can be naturally small without any fine tuning. This is reassuring, since it means that the small electron Yukawa coupling $y_e \sim 10^{-5}$ is perfectly natural.

Finally, the marginal interactions include 'vacuum angle' (v.a.) terms for each of the gauge groups:

$$\mathcal{L}_{\text{v.a.}} = \frac{g_1^2\Theta_1}{16\pi^2}\tilde{B}^{\mu\nu}B_{\mu\nu} + \frac{g_2^2\Theta_2}{8\pi^2}\text{Tr}(\tilde{W}^{\mu\nu}W_{\mu\nu}) + \frac{g_3^2\Theta_3}{8\pi^2}\text{Tr}(\tilde{G}^{\mu\nu}G_{\mu\nu}), \tag{6.3.53}$$

where $\tilde{B}^{\mu\nu} = \frac{1}{4}\epsilon^{\mu\nu\rho\sigma}B_{\rho\sigma}$, *etc.* These terms break CP, and are therefore very interesting. These terms are total derivatives, e.g.

$$\tilde{B}^{\mu\nu}B_{\mu\nu} = \partial^\mu K_\mu, \qquad K^\mu = \tfrac{1}{2}\epsilon^{\mu\nu\rho\sigma}A_\nu F_{\rho\sigma}. \tag{6.3.54}$$

This is enough to ensure that they do not give physical effects to all orders in perturbation theory. They can give non-perturbative effects with parametric dependence $\sim e^{1/g^2}$, but these are completely negligible for the $SU(2)_W \times U(1)_Y$ terms, since these gauge couplings are never strong. The strong vacuum angle gives rise to CP-violating non-perturbative effects in QCD, most importantly the electric dipole moment of the neutron. Experimental bounds on the neutron electric dipole moment require $\Theta_3 \lesssim 10^{10}$. Explaining this small number is the 'strong CP problem.' There are a number of proposals to solve the strong CP problem. For example, there may be a spontaneously broken Peccei-Quinn symmetry leading to an axion, or there may be special flavor structure at high scales that ensures that the determinant of the quark masses is real.

There is one relevant interaction that is allowed, namely a mass term for the Higgs field:

$$\mathcal{L}_{\text{relevant}} = -m_H^2 H^\dagger H. \tag{6.3.55}$$

Note that mass terms for the fermions such as Le^c are not gauge singlets, and therefore forbidden by gauge symmetry. The Higgs mass parameter cannot be forbidden by any obvious symmetry, and therefore must be fine tuned in order to be light compared to heavy thresholds such as the GUT scale. For example, in GUT models there are massive gauge bosons with masses of order M_{GUT} that couple to the Higgs with strength g, where g is the unified gauge coupling. These will contribute to the effective Higgs mass below the GUT scale

$$\Delta m_H^2 \sim \frac{g^2 M_{\text{GUT}}^2}{16\pi^2} \sim 10^{30}\text{GeV}^2 \tag{6.3.56}$$

for $M_{\text{GUT}} \sim 10^{16}$GeV. In order to get a Higgs mass of order 100 GeV we must fine tune to one part in 10^{26}!

We can turn this around and ask what is the largest mass threshold that is naturally compatible with the existence of a light Higgs boson. The top quark couples to the Higgs with coupling strength $y_t \sim 1$, and top quark loops give a quadratically divergent contribution to the Higgs mass. Assuming that this is cut off by a new threshold at the scale M, we find a contribution to the Higgs mass of order

$$\Delta m_H^2 \sim \frac{y_t^2 M^2}{16\pi^2}, \tag{6.3.57}$$

which is naturally small for $M \lesssim 1$TeV. We get a similar estimate for M from loops involving $SU(2) \times U(1)$ gauge bosons. So the standard model is natural as an effective field theory only if there is new physics at or below a TeV. This is the principal motivation for the Large Hadron Collider (LHC) at CERN, which will start operation in 2007-2008 with a center of mass energy of 14TeV. It is expected that the LHC will discover the mechanism of electroweak symmetry breaking and the new physics that makes it natural.

6.3.2 The GIM Mechanism

One very important feature of the standard model is that it violates flavor in just the right way. The quark mass matrices are proportional to the up-type and down-type Yukawa couplings. Diagonalizing the quark mass matrices requires that we perform independent unitary transformations on the two components of the quark doublet Q_i. This gives rise to the CKM mixing matrix, which appears in the interactions of the mass eigenstate quarks with the $W^\pm$ ('charged currents'). Crucially, the interactions with the photon and the Z ('neutral currents') are automatically diagonal in the mass basis. This naturally explains the phenomenology of flavor-changing decays observed in nature, including the 'GIM suppression' of flavor changing neutral current processes such as K^0–$\bar{K}^0$ mixing.

For our purposes, what is important is that this comes about because the quark Yukawa couplings are the only source of flavor violation in the standard model. If there were other couplings that violated quark flavor, these would not naturally be diagonal in the same basis that diagonalized the quark masses, and would in general lead to additional flavor violation. A simple example of this is a general model with 2 Higgs doublets, in which there are twice as many Yukawa coupling matrices.

6.3.3 Accidental Symmetries

It is noteworthy that the standard model was completely defined by its particle content gauge symmetries. In particular, we did not have to impose any additional symmetries to suppress unwanted interactions. If we look back at the terms we wrote down, we see that all of the relevant and marginal interactions are actually invariant under some additional global symmetries. One of these is baryon number, a $U(1)$ symmetry with charges

$$B(Q) = \tfrac{1}{3}, \quad B(u^c) = B(d^c) = -\tfrac{1}{3}, \quad B(L) = B(e^c) = B(H) = 0. \tag{6.3.58}$$

Another symmetry is lepton number, another $U(1)$ symmetry with charges

$$L(Q) = L(u^c) = L(d^c) = 0, \quad L(L) = +1, \quad L(e^c) = -1, \quad L(H) = 0. \tag{6.3.59}$$

These symmetries can be broken by higher-dimension operators. For example, the lowest-dimension operators that violate baryon number are dimension-six:

$$\Delta\mathcal{L} \sim \frac{1}{M^2} QQQL + \frac{1}{M^2} u^c u^c d^c e^c \ , \tag{6.3.60}$$

where the color indices are contracted using the $SU(3)_C$ invariant antisymmetric tensor. Consistency with the experimental limit on the proton lifetime of 10^{33} yr gives a bound $M \gtrsim 10^{22}$GeV. Although this is larger than the Planck mass, these couplings also violate flavor symmetries, and it seems reasonable that whatever explains the small values of the light Yukawa couplings can suppresses these operators.

A very appealing consequence of this is that if the standard model is valid up to a high scale M, then the proton is automatically long-lived, without having to assume that baryon number is an exact or approximate symmetry of the fundamental theory. Baryon number emerges as an 'accidental symmetry' in the sense that the other symmetries of the model (in this case gauge symmetries) do not allow any relevant or marginal interactions that violate the symmetry.

6.3.4 Neutrino Masses

Lepton number can be violated by the dimension-five operator

$$\Delta\mathcal{L} \sim \frac{1}{M}(LH)(LH). \tag{6.3.61}$$

When the Higgs gets a VEV, these gives rise to Majorana masses for the neutrinos of order

$$m_\nu \sim \frac{v^2}{M}. \tag{6.3.62}$$

In order to get neutrino masses in the interesting range $m_\nu \sim 10^{-2}$eV for solar and atmospheric neutrino mixing, we require $M \sim 10^{15}$GeV, remarkably close to the GUT scale. The interaction (6.3.61) also has a nontrivial flavor structure, so the actual scale of new physics depends on the nature of flavor violation in the fundamental theory, like the baryon number violating interactions considered above.

The experimental discovery of neutrino masses has been heralded as the discovery of physics beyond the standard model, but it can also be viewed as a triumph of the standard model. The standard model *predicts* that neutrino masses (if present) are naturally small, since they can only arise from an irrelevant operator. We can view the discovery of neutrino masses as evidence for the existence of a new scale in physics. This is analogous to the discovery of weak β decay, which can be described by an effective 4-fermion interaction with coupling strength $G_{\rm F} \sim 1/(100\text{GeV})^2$. (Therefore, Fermi was doing effective quantum field theory in the 1930's!)

6.3.5 Beyond the Standard Model Physics

The steps in constructing an extension of the standard model are the same ones we followed in constructing the standard model above. The model should be defined by its particle content and symmetries. We then write down all couplings allowed by these principles. The goal is to find an extension of the standard model that cures the naturalness problem, but preserves the successes of the standard model described above.

6.4 Conclusions

Effective field theory is the language in which all of modern theoretical physics is phrased (or should be phrased). It it well worth becoming proficient in speaking it.

7 Effective Field Theory and Inflation

Let us describe inflation as a low-energy effective theory. First, we identify the relevant light degrees of freedom at the energy scale of the "experiment" (for inflation, this is the Hubble scale, $E \sim H$). The EFT will contain at least one light scalar, the inflaton ϕ. Next, we write down the effective action for the inflaton, cf. eq. (6.2.2). We are obliged to down all operators consistent with the assumed symmetries of the inflaton,

$$\frac{\mathcal{O}_\delta}{\Lambda^{\delta-4}}\,, \tag{7.0.1}$$

where δ denotes the mass dimension of the operator. The purpose of this chapter is to explain that for inflation even Planck-suppressed operators don't decouple. Instead they make critical contributions to the dynamics. This indirectly makes inflation a window into quantum gravity.

7.1 UV Sensitivity

What value should we choose for the cutoff Λ? At what scale do we expect new degrees of freedom to become important? The larger the cutoff, the more suppressed the effects of the operators $\mathcal{O}_\delta$. However, the largest we can make the cutoff is the Planck-scale. The presence of some form of new physics at the Planck scale is required in order to render graviton-graviton scattering sensible, just as unitarity of W-W scattering requires new physics at the TeV scale. Although we know that new degrees of freedom must emerge, we cannot say whether the physics of the Planck scale is a finite theory of quantum gravity, such as string theory, or is instead simply an effective theory for some unimagined physics at yet higher scales.

Sensitivity to higher-dimension operators is commonplace in particle physics: as we saw in the previous section, bounds on flavor-changing processes place limits on physics above the TeV scale, and lower bounds on the proton lifetime even allow us to constrain GUT-scale operators that would mediate proton decay. However, particle physics considerations alone do not often reach beyond operators of dimension $\delta = 6$, nor go beyond $M \sim M_{\rm GUT}$. (Scenarios of gravity-mediated supersymmetry breaking are one exception.) Equivalently, Planck-scale processes, and operators of very high dimension, are irrelevant for most of particle physics: they decouple from low-energy phenomena.

In inflation, however, the flatness of the potential in Planck units introduces sensitivity to $\delta \leq 6$ *Planck-suppressed* operators, such as

$$\frac{\mathcal{O}_6}{M_{\rm pl}^2}\,. \tag{7.1.2}$$

As we explain in §7.2, an understanding of such operators is required to address the smallness of the eta parameter, i.e. to ensure that the theory supports at least 60 e-folds of inflationary expansion. This sensitivity to dimension-six Planck-suppressed operators is therefore common to all models of inflation.

For large-field models of inflation the UV sensitivity of the inflaton action is dramatically enhanced. As we discuss in §7.3, in this important class of inflationary models the potential becomes sensitive to an *infinite* series of operators of arbitrary dimension.

7.2 The Eta Problem

The most common field theory mechanisms for inflation involve a scalar field ϕ with mass m_ϕ parametrically smaller than the Hubble scale H:

$$\eta = \frac{m_\phi^2}{3H^2} \ll 1 \ . \tag{7.2.3}$$

It is difficult to protect this hierarchy against high-energy corrections. We know that some new degrees of freedom must appear at $\Lambda \lesssim M_{\rm pl}$ to give a UV-completion of GR. In string theory this scale is often found to be significantly below the Planck scale, $\Lambda \lesssim M_s \lesssim M_{\rm pl}$. If ϕ has $\mathcal{O}(1)$ couplings to the 'Planck slop' fields ψ, then integrating out the fields ψ yields the following corrections to the low-energy effective action

$$\Delta\mathcal{L}_\phi = \frac{\mathcal{O}_\Delta(\phi)}{\Lambda^{\Delta-4}} \ , \tag{7.2.4}$$

with $\mathcal{O}$ the set of all allowed operators in the effective theory of the inflaton field ϕ. Consider an inflationary lagrangian of slow-roll form

$$\mathcal{L}_\phi = -\frac{1}{2}(\partial_\mu\phi)^2 - V_\phi(\phi) \ . \tag{7.2.5}$$

The above argument makes us worry that integrating out the massive fields ψ yields corrections to the potential of the form

$$\Delta V = c\, V_0(\phi)\frac{\phi^2}{\Lambda^2} \ . \tag{7.2.6}$$

If this term arises, the eta parameter receives the following correction

$$\Delta\eta = \frac{M_{\rm pl}^2}{V_0}(\Delta V)'' \approx 2c\left(\frac{M_{\rm pl}}{\Lambda}\right)^2 \ . \tag{7.2.7}$$

Since $\Lambda \lesssim M_{\rm pl}$ and Wilson suggests $c \sim \mathcal{O}(1)$, we find

$$\Delta\eta \gtrsim 1 \ . \tag{7.2.8}$$

In supersymmetric theories a minor miracle occurs. Above the scale H, the theory is supersymmetric and the contributions from bosons and fermions precisely cancel. However, supersymmetry is spontaneously broken during inflation, leading to an inflation mass of order Hubble, $m_\phi \sim H$, and an eta parameter of order one, $\Delta\eta \sim 1$.

7.3 Large-Field Inflation

The Planck-scale sensitivity of inflation is dramatically enhanced in models with observable gravitational waves, $r > 0.01$. In this case, the inflaton field moves over a super-Planckian range during the last 60 e-folds of inflation, $\Delta\phi > M_{\rm pl}$. This observation makes an effective field theorists nervous and a string theorist curious. Let us explain why.

It is postulated that the theory below a cutoff Λ is

$$\mathcal{L}_{\rm s.r.} = -\frac{1}{2}(\partial_\mu\phi)^2 - V(\phi) \ . \tag{7.3.9}$$

The problem with field excursions larger than Λ is that the effective potential becomes sensitive to the couplings to massive degrees of freedom. Let us parameterize these degrees of freedom by a set of scalar fields ψ_i with nearly equal masses $m_{\psi_i} \sim \Lambda$. As good Wilsonian's we write down generic couplings between the inflaton ϕ and the massive fields ψ_i

$$V(\phi,\psi_i) = V(\phi) + \Lambda^2\psi_i^2 + g_i^2\phi^2\psi_i^2 + g_i^4\phi^4\frac{\psi_i^2}{\Lambda^2} + \cdots \tag{7.3.10}$$

For simplicity let us also assume that $g_i \sim g$. The potential then takes the form

$$V(\phi,\psi_i) = V(\phi) + \Lambda^2\psi_i^2 f_i\Big(\frac{g^2\phi_i^2}{\Lambda^2}\Big) \ . \tag{7.3.11}$$

For $\Delta\phi \gtrsim \frac{\Lambda}{g}$ the masses of the heavy fields change by order one, $\Delta m_\psi \sim \Lambda$. This implies that some of the fields ψ_i will leave the EFT ($m_\psi \gg \Lambda$), while others will join it ($m_\psi \ll \Lambda$). The EFTs at $\langle\phi\rangle = 0$ and $\langle\phi\rangle = \frac{\Lambda}{g}$ will be different; the effective potential for the inflaton will be different. Often the problem is presented as follows: assuming $\Lambda \to M_{\rm pl}$ and $g \sim 1$, the same EFT is valid throughout observable inflation only if

$$\Delta\phi \ll \frac{\Lambda}{g} \sim M_{\rm pl} \ . \tag{7.3.12}$$

The problem can also be expressed as a constraint on the couplings g. Since field excursions smaller than $\frac{\Lambda}{g}$ are still ok, observable gravity waves $\Delta\phi \gg M_{\rm pl}$, require

$$\frac{\Lambda}{g} \gg M_{\rm pl} \qquad \text{or} \qquad g \ll \frac{\Lambda}{M_{\rm pl}} \ . \tag{7.3.13}$$

This can be a strong constraint on the coupling in the generic situation $\Lambda \ll M_{\rm pl}$. This leads us to a key question: Why does the inflaton couple so weakly (not even gravitationally!) to Planck-scale degrees of freedom? Obviously, this is a question for a Planck-scale theory like string theory.

Let us compare the Planck-scale sensitivity of small-field and large-field inflation:

(i) In small-field inflation we worry that dimension-six Planck-suppressed operators $\frac{\mathcal{O}_6}{M_{\rm pl}^2}$ lead to unacceptably large corrections to the inflaton mass, $\Delta\eta \sim 1$. However, operators of higher dimensions are harmless, e.g. $\frac{\mathcal{O}_7}{M_{\rm pl}^3}$ gives $\Delta\eta \sim \frac{\Delta\phi}{M_{\rm pl}} \ll 1$. The important corrections in small-field inflation are dimension-six and smaller, $\Delta \leq 6$.

(ii) In contrast, for large-field inflation the terms $\frac{\mathcal{O}_\Delta}{M_{\rm pl}^{\Delta-4}}$ become larger for larger Δ.

In case (i), we have a finite number of operators $\mathcal{O}_{\Delta\leq 6}$ that contribute important corrections to the inflationary dynamics. We can therefore hope to enumerate all $\mathcal{O}_\Delta$ with $\Delta \leq 6$ and balance them against each other (fine-tuning). In case (ii), enumeration is not an option. One needs a sufficiently powerful symmetry to protect the inflaton from all corrections.

7.4 Non-Gaussianity

In is well-known that single-field slow-roll inflation produces Gaussian fluctuations. What kind of high-energy effects could deform slow-roll inflation in such a way as to produce large non-Gaussianity without disrupting the inflationary background solution? In effective field theory the effects of high-energy physics are encoded in high-dimension operators for the inflaton lagrangian.[1] Non-derivative operators such as ϕ^n/Λ^{n-4} form part of the inflaton potential and are therefore strongly constrained by the background. In other words, the existence of a slow-roll phase requires the non-Gaussianity associated with these operators to be small.[2] This naturally leads us to consider higher-derivative operators of the form $(\partial_\mu\phi)^{2n}/\Lambda^{4n-4}$. These operators don't affect the background, but in principle they could lead to strong interactions. Let us consider the leading correction to the slow-roll lagrangian

$$\mathcal{L} = \mathcal{L}_{\text{s.r.}} + \frac{(\partial_\mu\phi)^4}{8\Lambda^4} \ . \tag{7.4.14}$$

We split the inflaton field into background $\bar{\phi}(t)$ and fluctuations $\varphi(\mathbf{x},t)$. For $\dot{\bar{\phi}} \ll \Lambda^2$, we can ignore the correction to the quadratic lagrangian for φ,

$$\mathcal{L}_2 \approx -\frac{1}{2}(\partial_\mu\varphi)^2 \ . \tag{7.4.15}$$

We get the cubic lagrangian for φ by evaluating one of the legs of the interaction $(\partial\phi)^4$ on the background $\dot{\bar{\phi}}$,

$$\mathcal{L}_3 = -\frac{\dot{\bar{\phi}}}{2\Lambda^4}\,\dot{\varphi}(\partial_\mu\varphi)^2 \ . \tag{7.4.16}$$

We estimate the size of the non-Gaussianity as follows,

$$f_{\text{NL}} \sim \frac{1}{\zeta}\frac{\mathcal{L}_3}{\mathcal{L}_2} \sim \frac{1}{\zeta}\frac{\dot{\bar{\phi}}\,\dot{\varphi}}{\Lambda^4} \ . \tag{7.4.17}$$

Using $\dot{\varphi} \sim H\varphi$ and $\zeta = \frac{H}{\dot{\bar{\phi}}}\varphi$, we get

$$f_{\text{NL}} \sim \frac{\dot{\bar{\phi}}^2}{\Lambda^4} \ . \tag{7.4.18}$$

We therefore find that we only get significant non-Gaussianity when $\dot{\bar{\phi}} > \Lambda^2$, in which case we can't trust our derivative expansion. In other words, for $\dot{\bar{\phi}} > \Lambda^2$ there is no reason to truncate the expansion at finite order as we did in (7.4.14). Instead, operators of arbitrary dimensions become important in this limit,

$$P(X,\phi) = \sum c_n(\phi)\frac{X^n}{\Lambda^{4n-4}} \ , \qquad \text{where} \quad X \equiv -\frac{1}{2}(\partial_\mu\phi)^2 \ . \tag{7.4.19}$$

As an effective-field theory, eq. (7.4.19) makes little sense when $X > \Lambda^4$. All of the coefficients c_n are radiatively unstable. Hence, if we want to use a theory like (7.4.19) to generate large non-Gaussianity, we require a UV-completion. Interestingly, an example for such a UV-completion

[1] In addition, there could be high-energy modifications to the vacuum state. These effects require a separate discussion.

[2] A possibility to get large non-Gaussianities is to have additional light fields (i.e. fields with mass smaller than H) during inflation. These new degrees of freedom are not constrained by the slow-roll requirements and if their fluctuations are somehow converted into curvature perturbations, these can be much less Gaussian than in single-field slow-roll inflation.

exists in string theory. In Dirac-Born-Infeld (DBI) inflation,

$$P(X,\phi) = \frac{\Lambda^4}{f(\phi)}\sqrt{1 - f(\phi)\frac{X}{\Lambda^4}} - V(\phi) \ , \tag{7.4.20}$$

the form of the action is protected by a higher-dimensional boost symmetry. This symmetry protects eq. (A.3.36) from radiative corrections and allows a predictive inflationary model with large non-Gaussianity. It would be interesting to explore if there are other examples of $P(X)$ theories that are radiatively stable.

8 Supersymmetry and Inflation

8.1 Introduction

Why inflation and supersymmetry?

8.2 Facts about SUSY

Readers who don't know about SUSY won't learn it here.[1] Instead, this section is just a quick reminder of some essential facts about SUSY that we will need in our discussion of SUSY inflation.

8.2.1 SUSY and Naturalness

Scalar fields play a prominent role in many cosmological theory. Since scalar field theories suffer from UV divergencies, we have to worry about naturalness. Let us remind ourselves of the hierarchy problem in particle physics and the role of SUSY in its resolution.[2] This will serve as a useful analogy for the corresponding problems in cosmology.

The standard model (SM) of particle physics is well-known to be unreasonably effective, since it is in accord with all the experimental data. However, the consistency of the model relies on the Higgs field having a vacuum expectation value (vev) of 246 GeV even though this is highly unstable under quantum loop corrections. This instability can be seen by computing the loop corrections to the Higgs mass term. The fact that these corrections diverge quadratically with the high-energy cutoff is the signal that this instability is a severe problem. Much of the recent interest in supersymmetry (SUSY) has been driven by the possibility that SUSY can cure this instability.

The largest contribution to the Higgs mass correction in the SM of particle physics comes from the top quark loop. The top quark acquires a mass from the vev, $\langle H \rangle \equiv v$, of the, real, neutral component of the Higgs field H. Given the coupling of the Higgs to the top quark:

$$\mathcal{L}_{\text{Yukawa}} = -\frac{y_i}{\sqrt{2}} H \bar{t}_L t_R + \text{h.c.} \,, \tag{8.2.1}$$

where t_L and t_R are the left-handed and right-handed components of the top quark, y_t is the

[1] A good starting point for learning about SUSY is: Stephan Martin, *A Supersymmetry Primer*, (arXiv:hep-ph/9709356).

[2] See Terning, *Modern Supersymmetry*.

top Yukawa coupling. Expanding H around its vev, $H = v + h$, we find

$$m_t = \frac{y_t v}{\sqrt{2}} \ . \tag{8.2.2}$$

Given the coupling in eq. (8.2.1), we can easily evaluate the top loop contribution to the Higgs mass

$$\begin{aligned} -i\delta m_h^2\big|_{\rm top} &= (-1)N_c \int \frac{d^4k}{(2\pi)^4} \,{\rm Tr}\left[\frac{-iy_t}{\sqrt{2}}\frac{i}{\not{k} - m_t}\frac{-iy_t^*}{\sqrt{2}}\frac{i}{\not{k} - m_t}\right] \\ &= -2N_c|y_t|^2 \int \frac{d^4k}{(2\pi)^4}\frac{k^2 + m_t^2}{(k^2 - m_t^2)^2} \ . \end{aligned} \tag{8.2.3}$$

After a Wick rotation, $k_0 \to ik_4$ and $k^2 \to -k_{\rm E}^2$, we can perform the angular integration and impost a hard momentum cutoff, $k_{\rm E}^2 < \Lambda^2$. This yields

$$-i\delta m_h^2\big|_{\rm top} = \frac{iN_c|y_t|^2}{8\pi^2}\int_0^{\Lambda^2} dk_{\rm E}^2 \,\frac{k_{\rm E}^2(k_{\rm E}^2 - m_t^2)}{(k_{\rm E}^2 + m_t^2)^2} \ . \tag{8.2.4}$$

Changing variable to $x = k_{\rm E}^2 + m_t^2$, results in

$$\begin{aligned} -i\delta m_h^2\big|_{\rm top} &= -\frac{N_c|y_t|^2}{8\pi^2}\int_{m_t^2}^{\Lambda^2} dx \left(1 - \frac{3m_t^2}{x} + \frac{2m_t^4}{x^2}\right) \\ &= -\frac{N_c|y_t|^2}{8\pi^2}\left[\Lambda^2 - 3m_t^2 \ln\left(\frac{\Lambda^2 + m_t^2}{m_t^2}\right) + \cdots\right] \ , \end{aligned} \tag{8.2.5}$$

where $\cdots$ indicates finite terms in the limit $\Lambda \to \infty$. So we find that there are quadratically and logarithmically divergent corrections which (in the absence of a severe fine-tuning) push the natural value of the Higgs mass term (and hence the Higgs VEV) up toward the cutoff. Another way of saying this is that the SM can only be an effective field theory with a cutoff near 1 TeV, and some new physics must come into play near the TeV scale which can stabilize the Higgs vev. SUSY is (was?) the leading candidate for such new physics.

There is a simple way to stabilize the Higgs VEV by cancelling the divergent corrections to the Higgs mass term. Suppose there are N new scalar particles ϕ_L and ϕ_R that are lighter than a TeV with the following interactions:

$$\mathcal{L}_{\rm scalar} = -\frac{\lambda}{2}h^2(|\phi_L|^2 + |\phi_R|^2) - h(\mu_L|\phi_L|^2 + \mu_R|\phi_R|^2) - m_L^2|\phi_L|^2 - m_R^2|\phi_R|^2 \ . \tag{8.2.6}$$

The interactions in eq. (8.2.6) produce two new corrections to the Higgs mass term:

$$\delta m_h^2\big|_1 = \frac{\lambda N}{16\pi^2}\left[2\Lambda^2 - m_L^2 \ln\left(\frac{\Lambda^2 + m_L^2}{m_L^2}\right) - m_R^2 \ln\left(\frac{\Lambda^2 + m_R^2}{m_R^2}\right)\right] \ , \tag{8.2.7}$$

and

$$\delta m_h^2\big|_2 = \frac{N}{16\pi^2}\left[-\mu_L^2 \ln\left(\frac{\Lambda^2 + m_L^2}{m_L^2}\right) - \mu_R^2 \ln\left(\frac{\Lambda^2 + m_R^2}{m_R^2}\right)\right] \ . \tag{8.2.8}$$

Notice that if $N = N_c$ and $\lambda = |y_t|^2$ the quadratic divergences in eqs. (8.2.5) and (8.2.7) are canceled. If we also have $m_t = m_L = m_R$ and $\mu_L^2 = \mu_R^2 = 2\lambda m_t^2$, the logarithmic divergences in eqs. (8.2.5), (8.2.7) and (8.2.8) are canceled as well. SUSY is a symmetry between fermions and bosons that guarantees just these conditions. The cancellation of the logarithmic divergence is more than is needed to resolve the hierarchy problem; it is the consequence of powerful non-renormalization theorems.

8.2.2 Superspace and Superfields

Points in *superspace* are labeled by the following coordinates:

$$x^\mu\,,\ \theta^\alpha\,,\ \bar\theta_{\dot\alpha}\ . \tag{8.2.9}$$

Here, θ^α and $\bar\theta_{\dot\alpha}$ are constant complex anti-commuting two-component spinors with dimension $[\text{mass}]^{-1/2}$. In the superspace formalism, the components of a supermultiplet are united into a single *superfield*, which is a function of the superspace coordinates. Infinitesimal translations in superspace correspond to global supersymmetry transformations. Any superfield can be expanded in a power series in the anti-commuting variables[3]

$$S(x,\theta,\bar\theta) = a + \theta\psi + \bar\theta\bar\chi + \theta\theta M + \bar\theta\bar\theta N + \theta\sigma^\mu\bar\theta V_\mu + \theta\theta\bar\theta\bar\lambda + \bar\theta\bar\theta\theta\rho + \theta\theta\bar\theta\bar\theta D\ . \tag{8.2.10}$$

In these notes, we will only consider *chiral superfields*. These have the folllowing component expansion

$$\Phi(y,\theta) = \phi(y) + \sqrt{2}\theta\psi(y) + \theta^2 F(y)\ , \tag{8.2.11}$$

where

$$y^\mu \equiv x^\mu - i\theta\sigma^\mu\bar\theta\ . \tag{8.2.12}$$

Taylor expanding eq. (8.2.11) in the Grassmann variables θ and $\bar\theta$, we find

$$\Phi(x,\theta,\bar\theta) = \phi(x) - i\theta\sigma^\mu\bar\theta\partial_\mu\phi(x) - \tfrac{1}{4}\theta^2\bar\theta^2\partial^2\phi(x) + \sqrt{2}\theta\psi(x) + \tfrac{i}{\sqrt{2}}\theta^2\partial_\mu\psi(x)\sigma^\mu\bar\theta + \theta^2 F(x)\ . \tag{8.2.13}$$

Under a supersymmetry transformation the chiral superfield transforms as

$$\delta\Phi = i(\epsilon\mathcal{Q} + \bar\epsilon\bar{\mathcal{Q}})\Phi\ , \tag{8.2.14}$$

where

$$\mathcal{Q}_\alpha \ \equiv\ -i\frac{\partial}{\partial\theta^\alpha} - (\sigma^\mu)_{\alpha\dot\beta}\bar\theta^{\dot\beta}\frac{\partial}{\partial x^\mu}\ , \tag{8.2.15}$$

$$\bar{\mathcal{Q}}_{\dot\alpha} \ \equiv\ +i\frac{\partial}{\partial\bar\theta^{\dot\alpha}} + \theta^\beta(\sigma^\mu)_{\beta\dot\alpha}\frac{\partial}{\partial x^\mu}\ . \tag{8.2.16}$$

Exercise. Show that

$$\{\mathcal{Q}_\alpha, \bar{\mathcal{Q}}_{\dot\alpha}\} = 2(\sigma^\mu)_{\alpha\dot\alpha}\mathcal{P}_\mu\ , \qquad \text{where} \quad \mathcal{P}_\mu \equiv= -i\partial_\mu\ . \tag{8.2.17}$$

Show that eq. (8.2.14) implies

$$\delta_\epsilon\phi = \epsilon\psi\ , \tag{8.2.18}$$

$$\delta_\epsilon\psi_\alpha = -i(\sigma^\mu\bar\epsilon)_\alpha\partial_\mu\phi + \epsilon_\alpha F\ , \tag{8.2.19}$$

$$\delta_\epsilon F = -i\bar\epsilon\bar\sigma^\mu\partial_\mu\psi\ . \tag{8.2.20}$$

We also have anti-chiral superfields

$$\bar\Phi(y,\bar\theta) = \bar\phi(y) + \sqrt{2}\bar\theta\bar\psi(y) + \bar\theta^2\bar F(y)\ . \tag{8.2.21}$$

[3]Since there are two components of θ^α and likewise for $\bar\theta_{\dot\alpha}$, the expansion always terminates, with each term containing at most two θ's and two $\bar\theta$'s

8.2.3 Supersymmetric Lagrangians

A key observation is that the integral of a general superfield over all of superspace is automatically invariant under supersymmetry transformations

$$\delta_\epsilon A = 0 \ , \qquad \text{for} \quad A = \int \mathrm{d}^4x \int d^2\theta d^2\bar\theta \ S(x,\theta,\bar\theta) \ . \tag{8.2.22}$$

This follows immediately from the fact that $\mathcal{Q}$ and $\bar{\mathcal{Q}}$ in eqs. (8.2.15) and (8.2.16) are sums of total derivatives with respect to the superspace coordinates x^μ, θ, $\bar\theta$, so that $(\epsilon\mathcal{Q}+\bar\epsilon\bar{\mathcal{Q}})S$ vanishes upon integration.

In the special case of a chiral superfield, the integral over half the superspace is supersymmetric

$$\delta_\epsilon B = 0 \ , \qquad \text{for} \quad B = \int \mathrm{d}^4x \int d^2\theta \ \Phi(x,\theta,\bar\theta) \ . \tag{8.2.23}$$

To see this, we note that the F-term of a chiral superfield transforms into a total derivative, see eq. (8.2.20).

Let us construct the supersymmetric Lagrangian for a chiral superfield Φ. Consider the composite superfield

$$\begin{aligned}\bar\Phi\Phi \ = \ & \bar\phi\phi + \sqrt{2}\theta\psi\bar\phi + \sqrt{2}\bar\theta\bar\psi\phi + \theta\theta\bar\phi F + \bar\theta\bar\theta\phi\bar F + \bar\theta\bar\sigma^\mu\theta\left[i\bar\phi\partial_\mu\phi - i\phi\partial_\mu\bar\phi - \bar\psi\bar\sigma_\mu\psi\right] \\ & + \tfrac{i}{\sqrt{2}}\theta\theta\bar\theta\bar\sigma^\mu\left(\psi\partial_\mu\bar\phi - \partial_\mu\psi\bar\phi\right) + \sqrt{2}\theta\theta\bar\theta\bar\psi F + \tfrac{i}{\sqrt{2}}\bar\theta\bar\theta\theta\sigma^\mu\left(\bar\psi\partial_\mu\phi - \partial_\mu\bar\psi\phi\right) + \sqrt{2}\bar\theta\bar\theta\theta\psi\bar F \\ & + \theta\theta\bar\theta\bar\theta\left[\bar F F - \tfrac{1}{2}\partial^\mu\bar\phi\partial_\mu\phi + \tfrac{1}{4}\bar\phi\partial^2\phi + \tfrac{1}{4}\phi\partial^2\bar\phi + \tfrac{i}{2}\bar\psi\bar\sigma^\mu\partial_\mu\psi + \tfrac{i}{2}\psi\sigma^\mu\partial_\mu\bar\psi\right] \ ,\end{aligned} \tag{8.2.24}$$

where all fields are evaluated at x^μ. For the special case of a single chiral superfield, consider the integral over superspace

$$\int d^2\theta d^2\bar\theta \, \bar\Phi\Phi = -\partial^\mu\bar\phi\partial_\mu\phi + i\bar\psi\bar\sigma^\mu\partial_\mu\psi + \bar F F + \partial_\mu(\cdots) \ . \tag{8.2.25}$$

This is the Lagrangian density for the massless free Wess-Zumino model.

To obtain interactions and mass terms, we consider products of chiral superfields, such as Φ^2 and Φ^3. It is easy to see that any holomorphic function $W(\Phi)$ of a chiral superfield is also a chiral superfield. From this we can construct a supersymmetric Lagrangian by integrating over half the superspace

$$\int d^2\theta \, W(\Phi) + h.c. = \partial_\phi W F - \tfrac{1}{2}\partial_\phi^2 W\psi\psi - \partial_\mu(\cdots) + h.c. \ , \tag{8.2.26}$$

where $\partial_\phi W \equiv \partial_\Phi W|_{\Phi=\phi}$ and $\partial_\phi^2 W \equiv \partial_\Phi^2 W|_{\Phi=\phi}$. We added the complex conjugate, $\int d^2\bar\theta \, \bar W(\bar\Phi)$, the make the term real. For this to correspond to a Lagrangian density, $W(\phi)$ must have dimension $[\text{mass}]^3$. The most general renormalizable superpotential is

$$W(\Phi) = \tfrac{1}{2}m\Phi^2 + \tfrac{1}{3}g\Phi^3 \ . \tag{8.2.27}$$

The total Lagrangian is

$$\mathcal{L} = \int d^4\theta \ \bar\Phi\Phi + \left(\int d^2\theta \ W(\Phi) + h.c.\right) \tag{8.2.28}$$

$$= -\partial^\mu\bar\phi\partial_\mu\phi + i\bar\psi\bar\sigma^\mu\partial_\mu\psi + \bar F F + (\partial_\phi W F + h.c.) - \tfrac{1}{2}(\partial_\phi^2 W\psi\psi + h.c.) \ . \tag{8.2.29}$$

The part of the Lagrangian depending on the auxiliary field F take the simple form

$$\mathcal{L}_{\rm aux} = \bar{F}F + \partial_\phi W F + \partial_{\bar{\phi}} \bar{W} \bar{F} \ . \tag{8.2.30}$$

Notice that this is quadratic and without derivatives. This means that the field F does not propagate. Hence, we can solve the field equations for F,

$$F = -\partial_{\bar{\phi}} \bar{W} \ , \tag{8.2.31}$$

and substitute the result back into the Lagrangian

$$\mathcal{L}_{\rm aux} = -|\partial_\phi W|^2 \equiv -V_F(\phi) \ . \tag{8.2.32}$$

We see that the superpotential $W(\Phi)$ leads to a (F-term) potential $V_F(\phi)$ for the scalar field ϕ. Notice that the scalar potential is positive semi-definite, $V_F \geq 0$.

Exercise. Using eq. (8.2.27) show that:

1) the mass of the scalar ϕ equals the mass of the spinor ψ (after all the theory is supersymmetric).
2) the coefficient g of the Yukawa coupling $g(\phi\psi\psi)$ also determined the scalar self-coupling $g^2|\phi|^3$. Explain why this is the source of the "miraculous" cancellations in SUSY perturbation theory.

So far, we have only discussed renormalizable supersymmetric Lagrangians. As we have seen in previous chapters, inflation is an effective theory and we are therefore interested in non-renormalizable theories.

A non-renormalizable theory involving a set of chiral superfield Φ_i can be constructed as

$$\mathcal{L} = \int d^4\theta \ K(\Phi_i, \bar{\Phi}_j) + \left(\int d^2\theta \ W(\Phi_i) + h.c. \right) \tag{8.2.33}$$

- The *superpotential* W, is an arbitrary holomorphic of the chiral superfields treated as complex variables. It has dimension [mass]^3.
- The *Kähler potential* K is a function of both chiral and anti-chiral superfields. It is real, and has dimension [mass]^2.

The part of the Lagrangian coming from the superpotential is

$$\int d^2\theta \ W(\Phi_i) = W_i F^i - \tfrac{1}{2} W_{ij} \psi^i \psi^j \ , \tag{8.2.34}$$

where $W_i \equiv \partial_{\phi_i} W$ and $W_{ij} \equiv \partial_{\phi_i} \partial_{\phi_j} W$. After integrating out the auxiliary fields F^i, the part of the scalar coming from the superpotential is

$$V_F = K^{i\bar{\jmath}} W_i \bar{W}_{\bar{\jmath}} \ , \tag{8.2.35}$$

where $K^{i\bar{\jmath}}$ is the inverse of the Kähler metric

$$K_{i\bar{\jmath}} = \partial_{\phi_i} \partial_{\bar{\phi}_j} K \ . \tag{8.2.36}$$

Exercise. Derive eq. (8.2.35).

8.2.4 Miraculous Cancellations

Let us show explicit how SUSY enforces cancellations between fermion and boson loops. As a concrete example, consider the following renormalizable theory, $K = \bar{\Phi}\Phi$ and $W = \frac{1}{2}m\Phi^2 + \frac{1}{3}g\Phi^3$, with Lagrangian

$$\mathcal{L} = \partial^\mu\bar{\phi}\partial_\mu\phi + i\bar{\psi}\bar{\sigma}^\mu\partial_\mu\psi - |m\phi + g\phi^2|^2 - (\tfrac{1}{2}m + g\phi)\psi\psi - (\tfrac{1}{2}m + g\bar{\phi})\bar{\psi}\bar{\psi} \ . \tag{8.2.37}$$

Defining $\phi \equiv \frac{1}{\sqrt{2}}(A + iB)$ and $\Psi \equiv (\psi, \bar{\psi})$, we can write this as

$$\begin{aligned}\mathcal{L} = {} & \tfrac{1}{2}(\partial_\mu A)^2 - \tfrac{1}{2}m^2A^2 + \tfrac{1}{2}(\partial_\mu B)^2 - \tfrac{1}{2}m^2B^2 + \tfrac{1}{2}\bar{\Psi}(i\partial\!\!\!/ - m)\Psi \\ & - \tfrac{1}{\sqrt{2}}mgA(A^2 + B^2) - \tfrac{1}{4}g^2(A^4 + B^4 + 2A^2B^2) - \tfrac{1}{\sqrt{2}}g\bar{\Psi}(A - iB\gamma^5)\Psi \ .\end{aligned} \tag{8.2.38}$$

We can draw 5 one-loop corrections to the mass of A: 4 boson loops and 1 fermion loop. Using the usual Feynman rules for non-supersymmetric field theory, we find

$$(\mathrm{B}_1) + (\mathrm{B}_2) = 4g^2\int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{1}{k^2 - m^2} \ , \tag{8.2.39}$$

$$(\mathrm{B}_3) + (\mathrm{B}_4) = 4g^2m^2\int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{1}{(k^2 - m^2)((k-p)^2 - m^2)} \ , \tag{8.2.40}$$

and

$$\begin{aligned}(\mathrm{F}_1) &= -\left(-\frac{ig}{\sqrt{2}}\right)^2 2\int\frac{\mathrm{d}^4k}{(2\pi)^4}\,\mathrm{Tr}\left\{\frac{i(k\!\!\!/ + m)}{k^2 - m^2}\frac{i(k\!\!\!/ - p\!\!\!/ + m)}{(k-p)^2 - m^2}\right\} \\ &= -2g^2\left(\int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{1}{k^2 - m^2} + \int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{1}{(k-p)^2 - m^2}\right. \\ &\qquad \left. + \int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{4m^2 - p^2}{(k^2 - m^2)((k-p)^2 - m^2)}\right) \ .\end{aligned} \tag{8.2.41}$$

The total one-loop correction to the mass of A therefore is

$$2g^2\left\{\int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{1}{k^2 - m^2} - \int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{1}{(k-p)^2 - m^2} + \int\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{p^2 - 2m^2}{(k^2 - m^2)((k-p)^2 - m^2)}\right\} \ .$$

The signs here are crucial, arising from the relative sign between the bosonic and fermionic loops. The UV-divergences of the first to terms cancel, and the last term is only log-divergent

$$\int_\Lambda\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{1}{(k^2 - m^2)((k-p)^2 - m^2)} \approx \int_0^\Lambda\frac{2\pi^2k^3\mathrm{d}k}{(2\pi)^4}\frac{1}{k^4} \sim \ln\Lambda \ . \tag{8.2.42}$$

In contrast, non-supersymmetric theories usually produce quadratic divergences

$$\int_\Lambda\frac{\mathrm{d}^4k}{(2\pi)^4}\frac{1}{k^2 - m^2} \approx \int_0^\Lambda\frac{2\pi^2k^3\mathrm{d}k}{(2\pi)^4}\frac{1}{k^2} \sim \Lambda^2 \ . \tag{8.2.43}$$

In SUSY theories quadratic divergences cancel because of boson-fermion degeneracy of the spectrum.

8.2.5 Non-Renormalization Theorem

We have just seen an example for the power of SUSY to protect a theory against radiative corrections. Can this be generalized? How do theories with general K and W behave under quantum corrections?

The answer is remarkable:

- K gets corrections order-by-order in perturbation theory.
- W is *not* renormalized in perturbation theory.

In this section, we will follow Seiberg and prove the non-renormalization theorem for the superpotential. This will be a very slick argument, using only symmetries and holomorphy of the superpotential.

Consider a single chiral superfield Φ, with superpotential $W_{\text{tree}}(\Phi)$. As our canonical example, we will again use the Wess-Zumino model, $W_{\text{tree}}(\Phi) = \frac{1}{2}m\Phi^2 + \frac{1}{3}g\Phi^3$. For $m = g = 0$ the theory has a global $U(1) \times U(1)_R$. The R-symmetry acts as a phase factor on the superspace coordinate θ, i.e. $\theta \to e^{-i\alpha}\theta$. By convention $R[\theta] = 1$. Since Grassmann integrals define $\int d^2\theta\theta^2$ as a pure number ($= 1$), we also have $R[d^2\theta] = -2$. Thus $U(1)_R$ is a symmetry if $R[W] = 2$. Assign $R[\Phi] = 1$, so the free theory ($g = 0$) preserves an R-symmetry. Treat the coupling g as a background field (spurion field) that has charge $R[g] = -1$, so that the R-symmetry is preserved by the interaction. An ordinary $U(1)$ symmetry does not act on θ, so the superpotential W is neutral under such a symmetry. The following table summarizes the symmetries of the WZ-model:

	$U(1)$	$U(1)_R$
Φ	1	1
m	-2	0
g	-3	-3

To get an effective theory valid below some scale μ, we integrate out modes from Λ down to μ. In practice, you would start with W_{tree} at the scale Λ, expand in Fourier modes, and carry out the path integral over modes above μ (but below Λ). The result can be assembled into an effective superpotential W_{eff} that depends only on modes below μ. The claim is that $W_{\text{eff}}(\Phi) = W_{\text{tree}}(\Phi)$.

We will exploit symmetries to proves this result without any work. The superpotential must have R-charge 2 and vanishing $U(1)$ charge. For example, the term $m\Phi^2$ has those charges. Moreover, the combination $g\Phi/m$ is neutral under both $U(1)$'s. The most general form of the effective superpotential therefore is

$$W_{\text{eff}} = m\Phi^2 h\left(\frac{g\phi}{m}\right) \ , \tag{8.2.44}$$

where h is an unknown holomorphic function. Consider a power series expansion of the effective superpotential

$$W_{\text{eff}} = m\Phi^2 h\left(\frac{g\phi}{m}\right) = \sum_n a_n g^n m^{1-n}\Phi^{n+2} \ . \tag{8.2.45}$$

In the weak coupling limit, $g \to 0$, this must just give the mass term since in that case there are no interactions. Hence, we find $n \geq 0$. Next, we have to ensure that the massless limit, $m \to 0$,

is sensible. This requires $n \leq 1$. Hence, we have found that only $n = 0$ and $n = 1$ are allowed. Renaming the coefficients, we find

$$W_{\text{eff}} = \tfrac{1}{2}m\Phi^2 + \tfrac{1}{3}g\Phi^3 = W_{\text{tree}} \ . \tag{8.2.46}$$

The superpotential is not renormalized!

Exercise. Generalize the proof to arbitrary W_{tree}.

We have therefore reached the important conclusion that holomorphic couplings in the superpotential are not renormalized. Quantum effects in chiral field theory are completely captured by the renormalization of the Kähler potential (wave function renormalization).

8.2.6 Supersymmetry Breaking

SUSY is broken when the vacuum state $|0\rangle$ is not invariant under SUSY transformations

$$Q_\alpha|0\rangle \neq 0 \qquad \text{and} \qquad \bar{Q}_{\dot{\alpha}}|0\rangle \ . \tag{8.2.47}$$

In global SUSY, the Hamiltonian operator H is related to the SUSY generators through the SUSY algebra

$$H = P^0 = \frac{1}{4}(Q_1\bar{Q}_1 + \bar{Q}_1Q_1 + Q_2\bar{Q}_2 + \bar{Q}_2Q_2) \ . \tag{8.2.48}$$

If SUSY is unbroken in the vacuum state, it follows that $H|0\rangle = 0$ and the vacuum has zero energy. Conversely, if SUSY is spontaneously broken in the vacuum state, then the vacuum state must have positive energy, since

$$\langle 0|H|0\rangle = \frac{1}{4}\left(||\bar{Q}_1|0\rangle||^2 + ||Q_1|0\rangle||^2 + ||\bar{Q}_2|0\rangle||^2 + ||Q_2|0\rangle||^2\right) > 0 \ , \tag{8.2.49}$$

if the Hilbert space is to have positive norm.

F-term Breaking

If spacetime-dependent effects and fermion condensates can be neglected, then $\langle 0|H|0\rangle = \langle 0|V|0\rangle$, where

$$V = \bar{F}_i F^i \ . \tag{8.2.50}$$

In general the scalar potential will also have a D-term contribution, $V = \frac{1}{2}D^aD^a$. We have ignored this in our enormously simplified discussion of SUSY. From eq. (8.2.50) If the equations $F_i = \partial_{\phi_i}W = 0$ (and $D^a = 0$) can't be solved simultaneously, then SUSY is spontaneously broken.

O'Raifeartaigh Model

Consider a theory with three chiral superfields $\Phi_{1,2,3}$ and superpotential

$$W = -k\Phi_1 + m\Phi_2\Phi_3 + \tfrac{1}{2}y\Phi_1\Phi_3^2 \ , \tag{8.2.51}$$

where without loss of generality we can choose k, m and y to be real and positive (by phase rotations of the fields). Note that W contains a linear term. Such a term is required to achieve F-term breaking at tree-levle in renormalizable superpotential, since otherwise setting all $\phi_i = 0$

will always giv a supersymmetric global minimum with all $F_i = 0$. The linear term is allowed if the corresponding chiral supermultiplet is a gauge singlet. The scalar potential (8.2.50) becomes

$$V = |F_1|^2 + |F_2|^2 + |F_3|^3 \ , \tag{8.2.52}$$

where

$$F_1 = k - \tfrac{1}{2} y \bar{\phi}_3^2 \ , \tag{8.2.53}$$
$$F_2 = -m\bar{\phi}_3 \ , \tag{8.2.54}$$
$$F_3 = -m\bar{\phi}_2 - y\bar{\phi}_1\bar{\phi}_3 \ . \tag{8.2.55}$$

Clearly, $F_1 = 0$ and $F_2 = 0$ are not compatible, so SUSY must be broken. If $m^2 > yk$, then the absolute minimum of the classical potential is at $\phi + 2 = \phi_3 = 0$ with ϕ_1 undetermined, so $F_1 = k$ and $V = k^2$ at the minimum. The fact that ϕ_1 is undetermined at tree level is an example of a "flat direction" in the scalar potential; this is a common feature of supersymmetric models. However, as usual, the flat direction parameterized by ϕ_1 is an accidental feature of the classical scalar potential and doesn't survive quantum corrections, i.e. loop corrections will lift the potential.

8.2.7 Supergravity

Gravity exists, so if supersymmetry is realized in Nature, it has to be a local symmetry. Gauging global SUSY leads to supergravity (SUGRA). This is a huge and complicated topic. Learning the full machinery of SUGRA would keep us busy for a while. We will save our energy for applications of SUSY to inflation. The only result from SUGRA that we will need for that discussion is the F-term potential,

$$V_F = e^{K/M_{\rm pl}^2}\left[K^{i\bar{j}} D_i W \overline{D_j W} - 3|W|^2\right] \ , \tag{8.2.56}$$

where $D_i W = \partial_i W + \frac{1}{M_{\rm pl}^2}(\partial_i K)W$ is called the Kähler covariant derivative of W.

Exercise. Show that eq. (8.2.56) reduces to eq. (8.2.35) in the limit $M_{\rm pl} \to \infty$.

8.2.8 Further Reading

8.3 SUSY Inflation: Generalities

8.3.1 The Supergravity Eta-Problem

An important instance of the eta problem arises in locally-supersymmetric theories, i.e. in supergravity.[4] In $\mathcal{N} = 1$ supergravity, a key term in the scalar potential is the F-term potential,

$$V_F = e^{K/M_{\rm pl}^2}\left[K^{\varphi\bar{\varphi}} D_\varphi W \overline{D_\varphi W} - \tfrac{3}{M_{\rm pl}^2}|W|^2\right], \tag{8.3.57}$$

where $K(\varphi, \bar{\varphi})$ and $W(\varphi)$ are the Kähler potential and the superpotential, respectively; φ is a complex scalar field which is taken to be the inflaton; and we have defined $D_\varphi W \equiv \partial_\varphi W +$

[4] This case is relevant for many string theory models of inflation because four-dimensional supergravity is the low-energy effective theory of supersymmetric string compactifications.

$M_{\rm pl}^{-2}(\partial_\varphi K)W$. For simplicity of presentation, we have assumed that there are no other light degrees of freedom, but generalizing our expressions to include other fields is straightforward.

The Kähler potential determines the inflaton kinetic term, $-K_{,\varphi\bar\varphi}\,\partial_\mu\varphi\partial^\mu\bar\varphi$, while the superpotential determines the interactions. To derive the inflaton mass, we expand K around some chosen origin, which we denote by $\varphi \equiv 0$ without loss of generality, i.e. $K(\varphi,\bar\varphi) = K_0 + K_{,\varphi\bar\varphi}|_0\ \varphi\bar\varphi + \cdots$. The inflationary Lagrangian then becomes

$$\mathcal{L} \approx -K_{,\varphi\bar\varphi}\,\partial\varphi\partial\bar\varphi - V_0\Big(1 + K_{,\varphi\bar\varphi}|_0\,\frac{\varphi\bar\varphi}{M_{\rm pl}^2} + \dots\Big) \tag{8.3.58}$$

$$\equiv -\partial_\mu\phi\partial^\mu\bar\phi - V_0\Big(1 + \frac{\phi\bar\phi}{M_{\rm pl}^2}\Big) + \dots\,, \tag{8.3.59}$$

where we have defined the canonical inflaton field $\phi\bar\phi \approx K_{\varphi\bar\varphi}|_0\,\varphi\bar\varphi$ and $V_0 \equiv V_F|_{\varphi=0}$. We have retained the leading correction to the potential originating in the expansion of $e^{K/M_{\rm pl}^2}$ in eq. (8.3.57), which could plausibly be called a universal correction in F-term scenarios. The omitted terms, some of which can be of the same order as the terms we keep, arise from expanding $\Big[K^{\varphi\bar\varphi}D_\varphi W\overline{D_\varphi W} - \frac{3}{M_{\rm pl}^2}|W|^2\Big]$ in eq. (8.3.57) and clearly depend on the model-dependent structure of the Kähler potential and the superpotential. This results in a large, model-independent contribution to the eta parameter

$$\Delta\eta = 1\,, \tag{8.3.60}$$

as well as a model-dependent contribution which is typically of the same order. It is therefore clear that in an inflationary scenario driven by an F-term potential, eta will generically be of order unity.

Under what circumstances can inflation still occur, in a model based on a supersymmetric Lagrangian? One obvious possibility is that the model-dependent contributions to eta approximately cancel the model-independent contribution, so that the smallness of the inflaton mass is a result of fine-tuning. Clearly, it would be far more satisfying to exhibit a mechanism that *removes* the eta problem by ensuring that $\Delta\eta \ll 1$. This requires either that the F-term potential is negligible, or that the inflaton does not appear in the F-term potential. The first case does not often arise, because F-term potentials play an important role in presently-understood models for stabilization of the compact dimensions of string theory. However, in the next section, we will present a scenario in which the inflaton is a Goldstone boson and does not appear in the Kähler potential, or in the F-term potential, to any order in perturbation theory. This evades the particular incarnation of the eta problem that we have described above.

8.3.2 Goldstone Bosons in Supergravity

A promising approach to realize a technically natural small value for η is to make the inflaton a Goldstone boson with small mass protected by a shift symmetry. Consider, for example, a superpotential which spontaneously breaks a global $U(1)$ symmetry

$$W = S(\Phi\tilde\Phi - f^2)\,, \tag{8.3.61}$$

where Φ and $\tilde\Phi$ are two independent chiral superfields whose bottom components are the scalar fields ϕ and $\tilde\phi$. Let the expectation values of the fields be

$$\Phi = f e^{\theta/f} \quad \text{and} \quad \tilde\Phi = f e^{-\theta/f}\,, \tag{8.3.62}$$

where $\theta = \rho + i\varphi$ is a complex scalar field.[5] The canonical Kähler potential then becomes

$$K = \Phi^\dagger\Phi + \tilde{\Phi}^\dagger\tilde{\Phi} = 2f^2 \cosh\left(\frac{\theta+\theta^\dagger}{f}\right) , \tag{8.3.63}$$

and the supergravity potential for θ is

$$V = \exp\left[2\frac{f^2}{M_{\rm pl}^2}\cosh\frac{\theta+\theta^\dagger}{f}\right]\left[\sigma^4 + \dots\right] . \tag{8.3.64}$$

We notice that only the real part of θ acquires a mass; the shift symmetry of the Goldstone boson is protecting the imaginary component.[6]

8.4 SUSY Inflation: A Case Study

8.4.1 Hybrid Inflation and Naturalness

To illustrate the role of SUSY in inflationary model-building, we consider hybrid inflation as an example. Recall the basic elements of hybrid models: i) an inflaton ϕ with potential $V(\phi)$ that vanishes at the origin, $V(\phi = 0) = 0$, ii) a waterfall field ψ with symmetry breaking potential

$$\kappa(v^2 - \psi^2)^2 = V_0 - m_\psi^2\psi + \cdots \tag{8.4.65}$$

where $V_0 \equiv \kappa v^4$ and $m_\psi^2 \equiv 2\kappa v^2$. iii) a coupling between the inflaton field ϕ and the waterfall field ψ that controls the end of inflaton and removes the vacuum energy.

For concreteness, consider the following inflaton-waterfall coupling

$$\lambda\phi^2\psi^2 . \tag{8.4.66}$$

This generates a loop correction to the inflaton mass

$$\delta m_\phi^2 \sim \frac{\lambda}{16\pi^2}\Lambda_{\rm uv}^2 , \tag{8.4.67}$$

where $\Lambda_{\rm uv}$ is the cutoff of the loop integral. In order for ϕ to act as a switch on the waterfall field, we require

$$\lambda\phi_i^2 > m_\psi^2 , \tag{8.4.68}$$

where ϕ_i is the initial value of the infaton. This implies a minimal value for the natural size of the inflaton mass

$$m_\phi^2 > \frac{1}{16\pi^2}\frac{\Lambda^2 m_\psi^2}{\phi_i^2} . \tag{8.4.69}$$

The hybrid mechanism requires the hierarchy $m_\phi^2 \ll m_\psi^2$. This puts a strong upper limit on the cutoff $\Lambda_{\rm uv}$. Supersymmetry is one of the few ways we know to achieve a low cutoff in a

[5] In the following we use θ both for the chiral superfield and its bottom component. Which is meant should be clear from the context.

[6] This looks like a nice solution to the eta problem; however, it *assumes* that shift symmetry breaking contributions in the UV are small—*i.e.* we have to assume that there are *no* non-trivial corrections to (8.3.63). However, generic UV-completions are expected to break continuous global symmetries, so symmetries of the Kähler potential are not believed to persist beyond leading order. For the moment, we will assume that these effects are small, but we will return to it later.

controlled way. Even with SUSY, we may (at best) cut off the loop integral at $\Lambda_{\rm uv}^2 \sim m_\psi^2$. In that case, we get

$$m_\phi^2 > \frac{1}{16\pi^2}\frac{m_\psi^4}{\phi_i^2} \ . \tag{8.4.70}$$

We have to ensure that this lower limit on inflaton mass is consistent with the upper limit coming from smallness of the η-parameter

$$m_\phi^2 \ll \eta H^2 \sim \eta \frac{V_0}{M_{\rm pl}^2} \ . \tag{8.4.71}$$

Using $\phi_i \ll M_{\rm pl}$, we find consistency only if $m_\psi^4 \ll V_0$, i.e. the waterfall field has to be light compared to the scale of the total vacuum energy it controls. A naturally small mass for ψ again requires some symmetry explanation. In contrast with the slow-roll field, SUSY alone can protect the lightness of the waterfall.

We see that naturalness puts very specific requirements on the structures for hybrid models. In the next section, we present an example in the context of SUSY.

8.4.2 SUSY Pseudo-Natural Inflation

For purposes of illustration, we will consider the specific supergravity model of Arkani-Hamed et al.[7] (see also Kaplan and Weiner[8]).

The superpotential is

$$W = \lambda_0 S(\Phi\tilde{\Phi} - f^2) + \frac{\lambda_1}{2}(\Phi + \tilde{\Phi})\Psi^2 + \lambda_2 X(\Psi^2 - v^2) \ , \tag{8.4.72}$$

where $\lambda_1^2 f^2 > 2\lambda_2^2 v^2$ and

$$\Phi \equiv (f + \rho)e^{i\varphi/f} \ , \tag{8.4.73}$$

$$\tilde{\Phi} \equiv (f - \rho)e^{-i\varphi/f} \ . \tag{8.4.74}$$

This term preserves a $U(1)$ symmetry which is spontaneously broken. As before, the Goldstone boson φ associated with the broken symmetry will be the inflaton. Without loss of generality, we assume that the flat modulus ρ is stabilized at $\rho \equiv 0$ after supersymmetry breaking. The second term in W breaks the $U(1)$ explicitly and gives the Goldstone mode a potential. The field Ψ is the standard waterfall field of hybrid models of inflation. During inflation it is stabilized at $\Psi = 0$. Finally, the last term in W includes the field X whose F-term dominates the inflationary potential energy, $V_0 \approx |F_X|^2 = \lambda_2^2 v^4$. The Kähler potential takes the same form as in (8.3.63). In particular, it respects the $U(1)$ symmetry. Given this input (and for now assuming no other contributions to W and K), the inflationary potential receives two main contributions:

i) a loop-suppressed supergravity coupling

$$\delta K = \frac{\bar{\lambda}_1^2}{16\pi^2}(\Phi^\dagger\tilde{\Phi} + h.c.) \qquad \Rightarrow \qquad V_1 = V_0\Big(1 - \frac{\bar{\lambda}_1^2}{4\pi^2}\frac{f^2}{M_{\rm pl}^2}\sin^2\frac{\varphi}{f}\Big) \ , \tag{8.4.75}$$

where $\bar{\lambda}_1^2 \equiv \lambda_1^2 \log(\frac{\Lambda}{f})$ and we dropped a small constant term, $V_0(1 + \frac{\bar{\lambda}_1^2}{8\pi^2}\frac{f^2}{M_{\rm pl}^2}) \approx V_0$.

[7] Arkani-Hamed et al., *Pseudo-Natural Inflation*, (arXiv:hep-th/0302034).
[8] Kaplan and Weiner, (arXiv:hep-ph/0302014).

ii) a one-loop Coleman-Weinberg contribution

$$V_2 = V_0 \frac{\lambda_2^2}{4\pi^2} \log\Big(\frac{\lambda_1 \cos(\varphi/f)}{\mu/f}\Big)\,, \tag{8.4.76}$$

where μ is the renormalization scale.

The complete inflaton potential hence is

$$V = V_0 \left(1 - \frac{\bar{\lambda}_1^2}{4\pi^2}\frac{f^2}{M_{\rm pl}^2}\sin^2(\varphi/f) + \frac{\lambda_2^2}{4\pi^2}\log\big(\cos(\varphi/f)\big)\right)\,, \tag{8.4.77}$$

where have absorbed small constants into V_0. Small ε and η can be achieved with $\bar{\lambda}_1 \lesssim 1$, $\lambda_2 \ll 1$ and $f \ll M_{\rm pl}$. This is easily seen from (8.4.77) for the regime $\varphi \ll f$: in this case we find

$$\eta \simeq -\frac{\bar{\lambda}_1^2}{2\pi^2} - \frac{\lambda_2^2}{4\pi^2}\frac{M_{\rm pl}^2}{f^2} \qquad \text{and} \qquad \varepsilon \simeq \eta^2 \frac{\varphi^2}{M_{\rm pl}^2}\,, \tag{8.4.78}$$

and inflation with $\eta \lesssim 10^{-2}$ therefore requires

$$\bar{\lambda}_1 \lesssim 1 \qquad \text{and} \qquad \lambda_2 \lesssim \frac{f}{M_{\rm pl}} \ll 1\,. \tag{8.4.79}$$

Supersymmetry makes the small value of λ_2 technically natural.

8.4.3 UV Sensitivity

8.5 Signatures of SUSY Inflation

8.5.1 Hubble-Mass Scalars

8.5.2 The Squeezed Limit

8.5.3 Scale-Dependent Halo Bias

8.6 Conclusions

9 String Theory and Inflation

9.1 Introduction

When I find the energy I will add content to this chapter.

9.2 Elements of String Theory

9.2.1 Fields and Effective Actions

9.2.2 String Compactifications

9.3 Warped D-brane Inflation

9.4 Axion Monodromy Inflation

9.5 Conclusions

Outlook

Part III

Supplementary Material

A The Effective Theory of Single-Field Inflation

A.1 Introduction

The standard approach to study inflation is to assume the existence of a fundamental scalar field – the inflaton – and postulate a specific form for its action. Given this action one finds a background that describes an accelerating spacetime, $|\dot{H}| \ll H^2$. Perturbing around this background gives the action for fluctuations. In this note I want to describe an interesting alternative approach in which the most general *effective action* for the *fluctuations* is written down *directly*[1], without model-dependent assumptions about the microscopic physics sourcing the background. Instead, the inflationary quasi-de Sitter background $H(t)$ is assumed as given and the focus is on small fluctuations around this background. This formalism is particularly powerful in the study of non-Gaussianity.

By definition, $|\dot{H}| \ll H^2$, inflation implies approximate time-translation invariance of the background. However, inflation has to end, so the symmetry has to be spontaneously broken. As in gauge theory, there is a Goldstone boson associated with the symmetry breaking. In the case of inflation, this field characterizes fluctuations in the 'clock' measuring time during inflation

$$\pi \sim \delta t \sim \frac{\delta\phi}{\dot{\bar{\phi}}} \ . \tag{A.1.1}$$

The final equality is for the case of a fundamental scalar field ϕ, but we emphasize that the description in terms of π is more general than that. The Goldstone boson is related by a simple rescaling to the comoving curvature perturbation ζ (and hence to cosmological observables such as the CMB temperature fluctuations),

$$\zeta \sim \frac{\delta a}{a} \sim H\pi \ . \tag{A.1.2}$$

Being a Goldstone boson, the action for π is constrained by the symmetries of the background. The construction of the low-energy effective action for this Goldstone degree of freedom is the fundamental objective of these notes.

Before we embark on our journey to the effective theory of inflation, let me summarize what awaits us at the promised land:

- First and foremost, effective field theory is the central organizing principle of theoretical physics. Applying it to inflation will allow us to systematically classify all models of

[1] Cheung et al., *The Effective Theory of Inflation.*

inflation (as long as they are characterized by a single clock). In a systematic way the effective field theory explores the full range of possibilities.

- The effective theory doesn't commit to a specific microscopic realization of the physics of inflation. In particular, the effective theory of inflation allows cosmologists to stop apologizing for using scalar fields to describe inflation. It shows that the basic predictions of inflation don't rely on that assumption. It never mattered what was creating the background. The fluctuations are scalars because of Goldstone's theorem.

- Describing inflationary fluctuations in terms of the Goldstone degrees of freedom will help greatly in identifying the most relevant low-energy degrees of freedom. For instance, the Goldstone interpretation will allow a systematic decoupling of metric perturbations. Studying the Goldstone fluctuations in the *un*perturbed background allows the most direct way to obtain the leading order results. In alternative formulations of inflation this physical decoupling property is often much less manifest. All of this is summarized by Weinberg's First Law of Theoretical Physics: "You can use any variables you like to analyse a problem, but if you use the wrong variables you'll be sorry". For many applications Goldstone bosons are simply the right description of the physics.

- The Goldstone picture will make it clear what is dictated by the underlying symmetries of the de Sitter background (and hence model-insensitive) and what is not (and hence allows us to distinguish between different models).

- The symmetries of the background are non-linearly realized in the effective theory. This will lead to important correlations between different orders in the perturbation expansion. For instance, it shows that a small speed of sound (a term in the quadratic lagrangian) is related by symmetry to large interactions (a term in the cubic lagrangian).

- Physical energy scales are readily identified in the Goldstone language. This greatly facilitates understanding the dynamics of the theory in different energy regimes.

- Finally, the effective theory of inflation simply seems to be the most physical way to describe non-Gaussianities in single-clock inflation.

A.2 Spontaneously Broken Symmetries

Physical systems with spontaneously broken symmetries–i.e. systems for which a symmetry of the action is not a symmetry of the ground state–are ubiquitous in nature. Some of the key physical characteristics of such systems are captured by the low-energy effective theory of the Goldstone bosons associated with the symmetry breaking. Our goal in these notes is to formulate inflation as an example of spontaneous symmetry breaking. In this case, the symmetry is approximate time translation invariance of the de Sitter background. Before we develop the effective theory of inflation we review the standard treatment of symmetry breaking in gauge theory.

A.2.1 Global Symmetries

The physics of spontaneous symmetry breaking is based on two powerful theorems due to Noether and Goldstone. We assume that the reader is familiar with *Noether's theorem* which states that

for every global continuous symmetry of the action there exists a conserved current j^μ, with

$$\partial_\mu j^\mu = 0 \ . \tag{A.2.3}$$

Spontaneous breaking of a global symmetry naturally leads to Goldstone bosons[2] whose low-energy properties are largely governed by the nature of the symmetries which are spontaneously broken. The Goldstone state $|\pi\rangle$ is obtained by performing a symmetry transformation on the ground state $|0\rangle$, with spacetime-dependent transformation parameter. One can show that this implies that the following matrix element cannot vanish

$$\langle \pi | j^0(\mathbf{x}, t) | 0 \rangle \neq 0 \ , \tag{A.2.4}$$

where j^0 is the charge density guaranteed to exist by Noethers theorem. Eq. (A.2.4) implies the following two critical properties of Goldstone bosons:

1. *gaplessness*

   The energy of the Goldstone boson must vanish in the limit of vanishing 3-momentum:

   $$\lim_{p \to 0} E(p) = 0 \ . \tag{A.2.5}$$

   To prove this, we note that

   $$j^\mu(\mathbf{x}, t) = e^{-iHt} j^0(\mathbf{x}, 0) e^{iHt} \ , \tag{A.2.6}$$
   $$j^i(\mathbf{x}, t) = e^{-i\mathbf{P}\cdot\mathbf{x}} j^i(0, t) e^{i\mathbf{P}\cdot\mathbf{x}} \ , \tag{A.2.7}$$

   where $H|0\rangle = \mathbf{P}|0\rangle = 0$, $\mathbf{P}|\pi\rangle = \mathbf{p}|\pi\rangle$ and $H|g\rangle = E_p|\pi\rangle$. Differentiating (A.2.4) with respect to time, we find

   $$\begin{aligned} -iE_p e^{-iE_p t} \langle \pi | j^0(\mathbf{x}, 0) | 0 \rangle &= \langle \pi | \partial_0 j^0(\mathbf{x}, t) | 0 \rangle \\ &= -\langle \pi | \partial_i j^i(\mathbf{x}, t) | 0 \rangle \\ &= -ip^i \langle \pi | j^i(\mathbf{x}, t) | 0 \rangle \ . \end{aligned} \tag{A.2.8}$$

   The r.h.s. vanishes in the limit $p \to 0$. However, because of (A.2.4), the l.h.s. only vanishes if (A.2.5) holds. In relativistic theories–$E(p) = \sqrt{p^2 + m^2}$–this implies that the Goldstone bosons are *massless*. Not that we have arrived at this conclusion without every writing down an action.[3] We just followed Noether and Goldstone.

2. *low-energy decoupling*

   The above argument can be extended to more complicated matrix elements. In this way one finds that the Goldstone bosons must decouple from all interactions in the limit $p \to 0$.[4] This observation significantly constrains the low-energy action parameterizing the Goldstone interactions.[5]

[2] For spontaneously broken supersymmetry the Goldstone mode is a fermion.

[3] In fact, *gaplessness* and *low-energy decoupling* inform us what the effective action has to look like.

[4] Basically, in the zero-momentum limit, the spacetime-dependent symmetry transformation that generated the Goldstone boson from the ground state becomes spacetime-*independent* and the Goldstone mode becomes indistinguishable from the ground state, i.e. the Goldstone state becomes a symmetry transformation of the ground state.

[5] The description of the physics in terms of the low-energy lagrangian of the Goldstone boson is useful even when the underlying symmetry is not an exact symmetry. In this case, the small breaking of the symmetry can be treated perturbatively, leading to pseudo-Goldstone bosons with a small mass and small non-derivative interactions in the effective action, i.e. both gaplessness and low-energy decoupling become approximate.

A.2.2 Effective Lagrangian

Consider a gauge group G that is spontaneously broken to a subgroup H. Let T^a be the generators of the broken symmetry which live in the coset G/H. The Goldstone modes $|\pi\rangle$ are then obtained from the ground state $|0\rangle$ by performing a symmetry transformation with spacetime-dependent transformation parameter

$$U = e^{i\pi(x)/f_\pi} \ , \quad \text{where} \quad \pi \equiv \pi^a T^a \ . \tag{A.2.9}$$

The effective action for the Goldstone boson π has to be consistent with gaplessness and low-energy decoupling. At lowest order in a low-energy expansion – $\mathcal{O}(E^2)$ – the unique invariant lagrangian for the Goldstone boson is

$$\mathcal{L}_{\text{eff}} = -\frac{f_\pi^2}{2}\partial_\mu U\cdot\partial^\mu U^\dagger \ , \tag{A.2.10}$$

where $\partial_\mu U\cdot\partial^\mu U^\dagger \equiv \text{Tr}[\partial_\mu U\partial^\mu U^\dagger]$. In terms of π this becomes

$$\mathcal{L}_{\text{eff}} \ \rightarrow \ -\frac{1}{2}(\partial_\mu\pi)^2 + \frac{1}{6f_\pi^2}\left[(\pi\cdot\partial_\mu\pi)^2 - \pi^2(\partial_\mu\pi)^2\right] + \cdots \tag{A.2.11}$$

Notice the appearance of an infinity series of non-renormalizable interactions. Moreover, we see that there are special relations between the interactions dictated by the broken symmetry and that the couplings are determined by the single parameter f_π. These interactions are called *universal*. At higher-order in the energy expansion we obtain additional, *non-universal* interactions. For example, at $\mathcal{O}(E^4)$, we find the following single-derivative terms

$$\mathcal{L}_{\text{eff}} \ = \ -\frac{f_\pi^2}{2}\partial_\mu U\cdot\partial^\mu U^\dagger + c_1\,[\partial_\mu U\cdot\partial^\mu U^\dagger]^2 + c_2\,\partial_\mu U\cdot\partial_\nu U^\dagger\,\partial^\mu U\cdot\partial^\nu U^\dagger + \cdots \tag{A.2.12}$$

where c_1 and c_2 are model-dependent, dimensionless constants. If f_π sets the natural scale relative to which the low-energy limit is to be taken, then we expect $c_i \sim \mathcal{O}(1)$. In terms of π, we again find a series on non-renormalizable interactions,

$$\mathcal{L}_{\text{eff}} \ \rightarrow \ \cdots + \frac{c_1}{4f_\pi^4}\left((\partial_\mu\pi)^4 - \frac{2}{3f_\pi^2}(\partial_\mu\pi)^2(\pi\cdot\partial_\mu\pi)^2 + \cdots\right) + \cdots \ . \tag{A.2.13}$$

Individual interactions are again related by the non-linearly realized symmetry. Going beyond single-derivative terms we may include terms involving higher derivatives such as terms involving $\partial^2 U$. At $\mathcal{O}(E^4)$ this allows a number of additional terms.

A.2.3 A Toy UV-Completion

We digress briefly to present a simple field theory—the sigma model—that at low energies reduces to the effective lagrangian of the previous section:

$$\mathcal{L} = -\frac{1}{2}\partial_\mu\Sigma\cdot\partial^\mu\Sigma^\dagger + \frac{\mu^2}{2}\Sigma\cdot\Sigma^\dagger - \frac{\lambda}{4}[\Sigma\cdot\Sigma^\dagger]^2 \ , \tag{A.2.14}$$

with $\Sigma = \sigma + i\tilde{\pi}$, where $\tilde{\pi} \equiv T^a\tilde{\pi}^a$. At high energies, this provides a toy UV-completion of the effective Goldstone theory. At low energies, the classical minimum of the potential in (A.2.14) leads to a symmetry breaking vev, $\langle\sigma\rangle^2 = v^2 = \frac{\mu^2}{\lambda}$. To identify the Goldstone mode we introduce the following parameterization

$$\Sigma = (v + \rho(x))U(x) \ , \quad \text{where} \quad U = e^{i\pi(x)/v} \ , \tag{A.2.15}$$

such that $\pi \equiv T^a \pi^a = \tilde{\pi} + \cdots$. The lagrangian (A.2.14) becomes

$$\mathcal{L} = -\frac{1}{2}\left((\partial_\mu \rho)^2 - 2\mu^2 \rho^2\right) + \frac{(v+\rho)^2}{2}\partial_\mu U \cdot \partial^\mu U^\dagger - \lambda v \rho^3 - \frac{\lambda}{4}\rho^4 \ . \tag{A.2.16}$$

Integrating out the massive field ρ gives precisely the effective lagrangian (A.2.12),

$$\mathcal{L}_{\text{eff}} = -\frac{f_\pi^2}{2}\partial_\mu U \cdot \partial^\mu U^\dagger + c_1 [\partial_\mu U \cdot \partial^\mu U^\dagger]^2 + \cdots \ , \tag{A.2.17}$$

where $f_\pi \equiv v$ and $c_1 \equiv \frac{v^2}{8\mu^2} = \frac{1}{8\lambda}$. We see that the non-universal terms in (A.2.12) arise from integrating out the heavy fields of the UV-completion.

A.2.4 Energy Scales

In order to understand the dynamics of a theory, it is important to identify the energy scales at which different phenomena become important.

Symmetry Breaking Scale

At low energies, the symmetry is spontaneously broken and a description of the physics in terms of weakly-coupled Goldstone boson is appropriate. At sufficiently high energies, the symmetry is restored and degrees of freedom other than the Goldstone modes become relevant (such as the field ρ in the example of the previous section). In this section we give a precise definition of the symmetry breaking scale Λ_b that divides these two regimes.

By definition, any theory with a continuous global symmetry has a conserved Noether current j^μ even if the symmetry is spontaneously broken. A signature of spontaneous symmetry breaking is the fact that the charge $Q = \int d^3x\, j^0$ does not exist. The existence of a well-defined charge requires that $j^0(x)$ vanishes at least as x^{-3} in the limit $x \to 0$. In momentum space this means that $j^0(p)$ scales at most like p^{-1} for $p \to 0$. We will use this criterium to identify the energy scale below which the symmetry is spontaneously broken.

We start with the current associated with the effective lagrangian (A.2.11),

$$j^\mu = -f_\pi \partial^\mu \pi + \mathcal{O}(f_\pi^0) \ . \tag{A.2.18}$$

The two-point function of the current is

$$\int d^4x e^{ipx} \langle 0 | T\{j^\mu(x) j^\nu(0)\} | \rangle \rangle = i(p^\mu p^\nu - \eta^{\mu\nu} p^2)\Pi(p^2) \ , \tag{A.2.19}$$

where

$$\Pi(p^2) \equiv \frac{f_\pi^2}{p^2} + \mathcal{O}(1) \ . \tag{A.2.20}$$

At low energies, $\omega < f_\pi$, the first term in (A.2.20) dominates, $j^0 \sim p^{-2}$ and the charge at infinity does not exist. As a result, the symmetry is spontaneously broken at energies below $\Lambda_b = f_\pi$. At energies above f_π the first term in (A.2.20) is sub-dominant and the symmetry is restored. The upshot of this slightly formal argument is that the symmetry breaking scale can be read of from the scale in the Noether current for the canonically-normalized field. We will use the same argument to determine the symmetry breaking scale for inflation.

Strong Coupling Scale

The regime of validity of an effective theory is not always obvious. Given a microscopic definition of the theory (i.e. a UV-completion such as the sigma model of the previous section), the regime of validity is determined by the scales at which additional modes were integrated out. Given only the effective description, these energy scales may not be transparent in the Lagrangian. A fairly reliable method to identify the cutoff of the effective theory is to determine the energy scale at which the theory becomes strongly coupled.

The Goldstone action (A.2.11) is an expansion in π/f_π which contains irrelevant operators of arbitrarily large dimensions. These interactions become important at high energies. By dimensional analysis, the effective coupling is ω/f_π, suggesting strong coupling near f_π. More formally we can define the strong coupling scale as the energy scale at which the loop expansion breaks down or perturbative unitary of Goldstone boson scattering is violated. Such analysis lead to the strong coupling scale $\Lambda_\star = 4\pi f_\pi$.

A.2.5 Gauge Symmetries and Decoupling

Since gravity is described by a *local* gauge symmetry, we will be interested in additional effects that arise when the symmetry breaking involves a local symmetry. The effective action for the Goldstone boson then becomes

$$\mathcal{L} = -\frac{f_\pi^2}{2}\nabla_\mu U\cdot\nabla^\mu U^\dagger + \cdots \ , \tag{A.2.21}$$

where $\nabla_\mu \equiv \partial_\mu + igA_\mu$. Here, A_μ is the gauge field associated with the broken gauge symmetry. The quadratic Lagrangian for the Goldstone boson and the gauge field becomes

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^2 - \frac{1}{2}(\partial_\mu\pi)^2 - \frac{1}{2}m^2A_\mu^2 + im\,\partial_\mu\pi\cdot A^\mu \ , \tag{A.2.22}$$

where $m^2 \equiv f_\pi^2 g^2$. Of course, we could always go to the gauge where $\pi = 0$ (unitary gauge) and the theory is describe purely in terms of a massive vector A_μ. However, describing the physics in terms of the Goldstone boson has the advantage that it makes the high-energy behavior of the theory manifest. Specifically, it tells us that at high energies, the scattering of the longitudinal mode of the gauge field is well-described by the scattering of the Goldstone bosons. (This is an example of the Goldstone boson equivalence theorem.) This is seen most easily by taking the *decoupling limit* where $g \to 0$ and $m \to 0$ for $f_\pi = m/g = const.$ In this limit, there is now *no* mixing between π and A^μ and the Goldstone action reduces to (A.2.10) (the local symmetry has effectively become a global symmetry). For energies $E > m$, the Goldstone bosons are the most convenient ways to describe the scattering of the massive vector fields. Restoring finite g and m, gives corrections to the results from pure Goldstone boson scattering that are perturbative in m/E and g^2.

This completes our review of spontaneous symmetry breaking in gauge theory. We are now ready to apply the same formalism to inflation.

A.3 Effective Theory of Inflation

A useful definition of inflation is as an FRW background with nearly constant expansion rate $H \approx const.$, or $|\dot{H}| \ll H^2$. However, inflation has to end, so the time-translation invariance

of the quasi-de Sitter background[6] has to be spontaneously broken, $H(t)$. This suggests that we can formulate inflation as an example of spontaneous symmetry breaking. We will use the classic treatment of the previous section as an analogy.[7]

A.3.1 Goldstone Action

To identify the Goldstone boson associated with the symmetry breaking, we perform a spacetime-dependent time shift

$$U \equiv t + \pi(x) \ . \tag{A.3.23}$$

We now construct the effective action for the Goldstone mode π. By definition, the Goldstone field transforms as $\pi \to \pi - \xi$ under time reparameterization $t \to t + \xi$, so that $U = t + \pi$ is invariant. The Goldstone mode π is related to the primordial curvature perturbation ζ by a simple rescaling, $\zeta = -H\pi$. Understanding the action for π will therefore allow us to compute correlation functions for cosmological observables. Following our gauge theory example, the effective action for the Goldstone boson is a general function of U

$$\mathcal{L} = F(U, (\partial_\mu U)^2, \Box U, \cdots) \ . \tag{A.3.24}$$

The low-energy expansion of this action unifies all known single-field models of inflation[8] and allows a systematic classification of interactions.

Slow-Roll Inflation

At $\mathcal{O}(E^2)$, the lagrangian is

$$\mathcal{L} = \Lambda^4(U) - f^4(U) g^{\mu\nu} \partial_\mu U \partial_\nu U \ , \tag{A.3.25}$$

where $\Lambda(U)$ and $f(U)$ are a priori free functions of the invariant time $U = t + \pi$. However, demanding *tadpole cancellation* determines the coefficients

$$\Lambda^4 \equiv -M_{\rm pl}^2 (3H^2 + \dot{H}) \qquad \text{and} \qquad f^4 \equiv M_{\rm pl}^2 \dot{H} \ . \tag{A.3.26}$$

Eq. (A.3.26) ensures that the action starts quadratic in π when the equations of motion of the FRW background are imposed. At leading order, the coefficients of the action are therefore completely fixed by the de Sitter background $H(t)$

$$\mathcal{L} = M_{\rm pl}^2 \dot{H} g^{\mu\nu} \partial_\mu U \partial_\nu U - M_{\rm pl}^2 (3H^2 + \dot{H}) \ . \tag{A.3.27}$$

This is nothing but *slow-roll inflation* in disguise:

$$\mathcal{L} = -\frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - V(\phi) \ , \tag{A.3.28}$$

where $\phi = \dot{\bar{\phi}}(t + \pi)$ and $V(\phi) = M_{\rm pl}^2 (3H^2 + \dot{H})$. The theory in (A.3.27) includes couplings between the Goldstone π and metric fluctuations $\delta g^{\mu\nu}$. This is analogous to the couplings

[6] The symmetry group G of de Sitter space includes both temporal and spatial diffeomorphisms. Spatial diffeomorphisms H will be preserved during inflation, but time diffeomorphisms, living in G/H, will be broken.

[7] Since inflation breaks a *spacetime* symmetry and not an internal symmetry, the analogy will not be perfect.

[8] In fact, the systematic treatment of the effective theory helped to identify single field theories that hadn't been previously discussed.

between π and A_μ in the gauge theory example. Just like in the gauge theory we can find a limit in which π alone controls the dynamics, i.e. we can define a *decoupling limit* $M_{\rm pl} \to \infty$ and $\dot{H} \to 0$, with $M_{\rm pl}^2 \dot{H}$ fixed. This limit is the same as in gauge theory under the following identifications: $g \to M_{\rm pl}^{-1}$, $m^2 \to \dot{H}$ and $E \to H$. At energies $E^2 \gg \dot{H}$, we can therefore ignore the mixing of π with metric perturbations $\delta g^{\mu\nu}$, i.e. we can evaluate the action for the Goldstone mode π in the unperturbed de Sitter background $\bar{g}^{\mu\nu}$

$$g^{\mu\nu}\partial_\mu U \partial_\nu U \;\to\; \bar{g}^{\mu\nu}\partial_\mu(t+\pi)\partial_\nu(t+\pi) \;=\; -1 - 2\dot{\pi} + (\partial_\mu \pi)^2 \; . \tag{A.3.29}$$

Since we care about correlation functions evaluated at freeze-out, $E^2 \sim H^2 \gg |\dot{H}|$, the decoupled π-lagrangian derived from (A.3.29) will give answers that are accurate up to fractional corrections of order $\frac{H^2}{M_{\rm pl}^2}$ and $\frac{\dot{H}}{H^2} = -\varepsilon$. For eq. (A.3.27), the decoupling limit (A.3.29) implies

$$\mathcal{L}_{\rm s.r.} = M_{\rm pl}^2 \dot{H} (\partial_\mu \pi)^2 \; . \tag{A.3.30}$$

We note that in the decoupling limit the Goldstone mode is precisely massless[9] and the fluctuations are perfectly Gaussian. This is consistent with our interpretation of (A.3.27) as the action for fluctuations around slow-roll inflationary backgrounds. The near-perfect Gaussianity won't be maintained when we consider higher orders in the derivative expansion.

DBI Inflation

At next-to-leading order, we can add the following single-derivative term to the effective lagrangian

$$\mathcal{L}_{c_s} = \frac{1}{2} M_2^4 \left(g^{\mu\nu}\partial_\mu U \partial_\nu U + 1 \right)^2 \; , \tag{A.3.31}$$

where '-1' was subtracted from $(\partial_\mu U)^2$ to cancel the tadpole, i.e. to ensure that (A.3.31) starts quadratic in π. In the decoupling limit (A.3.29) this adds the following terms to the Goldstone action

$$\mathcal{L}_{\rm c_s} = 2 M_2^4 \left(\dot{\pi}^2 + \dot{\pi}(\partial_\mu \pi)^2 \right) \; . \tag{A.3.32}$$

We note that a *non-linearly realized symmetry* relates *dispersion* to *interactions*. In other words, the size of the kinetic term $\dot{\pi}^2$ and the strength of the interaction $\dot{\pi}(\partial_\mu \pi)^2$ are related to the same coefficient M_2. In principle, $M_2(t+\pi)$ can depend on time, but in practice we are always interested in cases where any time-dependence is proportional to the small parameter $\varepsilon \ll 1$, or $\dot{M}_2 \ll H M_2$. Adding (A.3.30) and (A.3.31), we find

$$\mathcal{L}_{\rm s.r.} + \mathcal{L}_{c_s} = -\frac{M_{\rm pl}^2 \dot{H}}{c_s^2} \Big[\left(\dot{\pi}^2 - c_s^2 (\partial_i \pi)^2 \right) + \dot{\pi}(\partial_\mu \pi)^2 \Big] \; , \tag{A.3.33}$$

where we defined a sound speed

$$\frac{1}{c_s^2} \equiv 1 - \frac{2 M_2^2}{M_{\rm pl}^2 \dot{H}} \; . \tag{A.3.34}$$

[9] Including the mixing with gravity gives the (pseudo-)Goldstone a small mass $m_\pi^2 = -6\dot{H}$,

$$\mathcal{L}_{\rm s.r.} = M_{\rm pl}^2 \dot{H} \left((\partial_\mu \pi)^2 + 3\varepsilon H^2 \pi^2 \right) \; .$$

This mass for π is precise what is required so that $\zeta = -H\pi$ is massless

$$\mathcal{L}_{\rm s.r.} = M_{\rm pl}^2 \dot{H} (\partial_\mu \zeta)^2 \; ,$$

and hence freezes outside of the horizon.

The Goldstone formalism makes it apparent how a small sound is by symmetry related to large interactions and hence large non-Gaussianities.

Adding higher powers of $g^{\mu\nu}\partial_\mu U\partial_\nu U$ reproduces the so-called $P(X)$–theories, with $X \equiv -\frac{1}{2}(\partial_\mu\phi)^2$,

$$\mathcal{L}_{P(X)} = \sum_n \frac{1}{n!} M_n^4 \left(g^{\mu\nu}\partial_\mu U\partial_\nu U + 1\right)^n \ , \qquad \text{where} \quad M_n^4 = \bar{X}^n \frac{\partial^n P}{\partial \bar{X}^2} \ . \tag{A.3.35}$$

This includes DBI inflation as a special case

$$P(X,\phi) = f(\phi)^{-1}\sqrt{1 - f(\phi)X} - V(\phi) \ . \tag{A.3.36}$$

The DBI action (A.3.36) is special in that its form is protected against radiative corrections by a higher-dimensional boost symmetry.[10]

Higher Derivatives

To complete our discussion of the effective Goldstone action we consider contributions with more than one derivative acting on U. This leads to models of *ghost inflation* and *galieon inflation* (if the operators satisfy the Galilean symmetry $\pi \to \pi + b_\mu x^\mu + c$). A complete description of these higher-derivative theories is beyond the scope of these notes. We therefore restrict ourselves to citing all higher-derivative operators that contribute to the action up to cubic order in π,

$$\begin{aligned}\mathcal{L} \quad = \quad & \bar{M}_1^2\big[\Box U\big]^2 + \bar{M}_2\big[\Box U\big]^3 + \bar{M}_3^3\big[\Box U\big]\big[(\partial_\mu U)^2\big] + \bar{M}_4^3\big[\Box U\big]\big[(\partial_\mu U)^2\big]^2 \\ & + \bar{M}_5^2\big[\Box U\big]^2\big[(\partial_\mu U)^2\big] + \bar{M}_6^2\big[\nabla^\mu\nabla_\alpha U\big]\big[\nabla^\alpha\nabla_\mu U\big] + \bar{M}_7^2\big[(\partial_\mu U)^2\big]\big[\nabla^\mu\nabla_\alpha U\big]\big[\nabla^\alpha\nabla_\mu U\big] \\ & + \bar{M}_8^2\big[\Box U\big]\big[\nabla^\mu\nabla_\alpha U\big]\big[\nabla^\alpha\nabla_\mu U\big] + \bar{M}_9\big[\nabla^\mu\nabla_\alpha U\big]\big[\nabla^\alpha\nabla_\beta U\big]\big[\nabla^\beta\nabla_\mu U\big] + \cdots \ ,\end{aligned}$$

where the square brackets were introduced to denote terms with background values subtracted, such as

$$\big[(\partial_\mu U)^2\big] \equiv (\partial_\mu U)^2 + 1 \tag{A.3.37}$$

$$\big[\nabla_\mu\nabla_\nu U\big] \equiv \nabla_\mu\nabla_\nu U - H(g_{\mu\nu} + \nabla_\mu U\nabla_\nu U) \ . \tag{A.3.38}$$

This ensures that action doesn't have tadpoles. In unitary gauge the higher-derivative operators are related to the extrinsic curvature $\delta K_{\mu\nu}$. We will ignore them in the remainder and instead focus on the predictions from the single-derivative action.

A.3.2 Energy Scales

Arguably the most interesting limits of the effective theory of inflation is the limit of small sound speed, $c_s \ll 1$. In this limit, interactions are systematically enhanced and the non-Gaussianity of the fluctuations can be large. The Goldstone action in this limit is

$$\mathcal{L} = -\frac{M_{\rm pl}^2\dot{H}}{c_s^2}\Big[\left(\dot{\pi}^2 - c_s^2(\partial_i\pi)^2\right) + \dot{\pi}(\partial_i\pi)^2\Big] \ , \tag{A.3.39}$$

where we have dropped the $\dot{\pi}^3$ interaction since it is suppressed by a factor of c_s^2 relative to the $\dot{\pi}(\partial_i\pi)^2$ interaction.

To understand the dynamics of the theory we identify three fundamental energy scales:

[10] General $P(X)$ theories do not have this feature and are hence plagued by radiative instabilities.

- the *symmetry breaking scale*, $\Lambda_{\rm b}$, is the energy scale at which time translations are spontaneously broken and a description in terms of a Goldstone boson first becomes applicable;

- the *strong coupling scale*, $\Lambda_\star$, defines the energy scale at which the effective description breaks down and perturbative unitarity is lost;

- the *Hubble scale*, H, is the energy scale associated with the cosmological experiment.

In *slow-roll inflation*, these three energy scales are easily identified. Time-translation invariance is broken by the background $\bar{\phi}(t)$ at the scale $\Lambda_{\rm b}^4 = \dot{\bar{\phi}}^2 = 2M_{\rm pl}^2|\dot{H}|$. At energy scales above $\Lambda_{\rm b}$, the symmetry is restored and we should not integrate out the background. Because the theory is effectively Gaussian, the self-interactions of ϕ are weak up to very high energies. The theory only becomes strongly coupled at the Planck scale, so the UV-cutoff is $M_{\rm pl}$. Inflationary observables freeze out at horizon-crossing, or when their frequencies become equal to the expansion rate, $\omega \sim H$. Inflation therefore directly probes energies of order Hubble, i.e. the energy scale of the experiment is H. The freeze-out at the Hubble scale is universal, but the symmetry breaking scale and the strong coupling scale are model-dependent. We will now determine these scales for the lagrangian (A.3.39).

Symmetry Breaking

Although the inflationary background spontaneously breaks a gauge symmetry, in the decoupling limit the gauge symmetry becomes a global symmetry. As long as the decoupled π-lagrangian is a reliable description, the language of spontaneously broken global symmetries will therefore be useful.

We determine the symmetry breaking scale for inflation by considering the Noether current of time translations, i.e. the zero-component of the stress tensor, $j^\mu = T^{0\mu}$. For the lagrangian (A.3.39), we find

$$j^0 = T^{00} = \frac{2M_{\rm pl}^2\dot{H}}{c_s^2}\dot{\pi} + \cdots \tag{A.3.40}$$

From our gauge theory discussion we expect $2M_{\rm pl}^2\dot{H}c_s^{-2}$ to control the symmetry breaking (i.e. to control the energy scale at which the charge ceases to exist). However, for $c_s \neq 1$ Lorentz invariance is broken so we have to be careful to define an energy scale. In fact, by dimensional analysis we see that the coefficient in (A.3.40) is an energy *density* $[\omega][k^3]$, not an energy^4 $[\omega^4]$. To convert the symmetry breaking scale to a true energy scale, we use the dispersion relation $\omega = c_s k$. We conclude that the symmetry is spontaneously broken at

$$\Lambda_{\rm b}^4 \equiv 2\,M_{\rm pl}^2|\dot{H}|c_s\ . \tag{A.3.41}$$

Although the current gives a natural definition of the symmetry breaking scale, it is nice to check that it agrees with our intuition. First of all, in the case of slow-roll inflation (i.e. for $c_s = 1$), the symmetry breaking scale is given by $2\,M_{\rm pl}^2|\dot{H}| = \dot{\bar{\phi}}^2$. This matches the intuition that the time variation of the background is breaking the symmetry. Moreover, one may rewrite the (dimensionless) power spectrum of curvature fluctuations in terms of the symmetry breaking scale

$$\Delta_\zeta \equiv k^3 P_\zeta(k) = \frac{1}{4}\frac{H^2}{M_{\rm pl}^2\,\epsilon\,c_s} = \frac{1}{2}\Big(\frac{H}{\Lambda_{\rm b}}\Big)^4\ . \tag{A.3.42}$$

Hence, when $H \sim \Lambda_{\rm b}$, the size of quantum fluctuations is of the same order as the symmetry breaking scale. This is the regime of *eternal inflation*, which is again consistent with the interpretation of $\Lambda_{\rm b}^4 = \dot{\bar{\phi}}^2$ for slow-roll.

Strong Coupling

Next, we determine the strong coupling scale from the action for the Goldstone boson (A.3.39),

$$\mathcal{L} = -\frac{1}{2}\left(\dot{\pi}_c^2 - c_s^2(\partial_i \pi_c)^2\right) + \frac{1}{M_\star^2}\dot{\pi}_c(\partial_i \pi_c)^2 \ , \tag{A.3.43}$$

where we defined $\pi_c^2 = 2M_{\rm pl}^2|\dot{H}|c_s^{-2}\pi^2$ and

$$M_\star^4 \equiv M_{\rm pl}^2|\dot{H}|c_s^{-2}(1-c_s^2)^{-1} \ . \tag{A.3.44}$$

We expect that the theory becomes strongly coupled at an energy scale related to the coupling $M_\star$. For a relativistic theory the strong coupling scale would be equal to $M_\star$, but in the non-relativistic small-c_s theory we again have to careful to identify $\Lambda_\star^4$ with units $[\omega^4]$. By dimensional analysis, we find $[M_\star^4] = [k]^7[\omega]^{-3}$, and hence $\Lambda_\star^4 = M_\star^4 \times c_s^7$, or

$$\Lambda_\star^4 \equiv M_{\rm pl}^2|\dot{H}|c_s^5(1-c_s^2)^{-1} \ . \tag{A.3.45}$$

As in the gauge theory example, the effective coupling is given by $\omega/\Lambda_\star$. We expect that strong coupling arises at some order-one value of this coupling. More formally, we can define the strong coupling scale by the breakdown of perturbative unitarity of the Goldstone boson scattering. This is found to happen when $\omega^4 > 2\pi\Lambda_\star^4$. Note the large suppression of $\Lambda_\star^4$ by factors of $c_s \ll 1$.

The interactions which become strongly coupled are the same that give rise to measurable non-Gaussianity. As a result, we should be able to interpret the strong coupling scale in terms of the size of the non-Gaussianity. A simple estimate for the amplitude of the non-Gaussianity is

$$f_{\rm NL}\,\zeta \ \equiv \ \left.\frac{\mathcal{L}_3}{\mathcal{L}_2}\right|_{\omega=H} \sim \frac{M_{\rm pl}^2\dot{H}c_s^{-2}\,\dot{\pi}(\partial_i\pi)^2}{M_{\rm pl}^2\dot{H}\,\dot{\pi}^2} = c_s^{-2}\,\zeta \sim \left(\frac{\Lambda_{\rm b}}{\Lambda_\star}\right)^2\zeta \ . \tag{A.3.46}$$

Using the power spectrum (A.3.42) as an estimate for the size of $\zeta \sim \Delta_\zeta^{1/2}$, we find

$$\frac{\mathcal{L}_3}{\mathcal{L}_2} \sim \left(\frac{H}{\Lambda_\star}\right)^2 \ . \tag{A.3.47}$$

We see that $\mathcal{L}_3 \sim \mathcal{L}_2$, or $f_{\rm NL} \sim \zeta^{-1} \sim 10^4$, when $H \sim \Lambda_\star$. This indicates a breakdown of the perturbative description as $\Lambda_\star$ approaches H.

New Physics?

Summary. In the previous sections we derived two important energy scales in the effective theory of inflation: the symmetry breaking scale, $\Lambda_{\rm b}^4 = 2M_{\rm pl}^2|\dot{H}|c_s$, and the strong coupling scale, $\Lambda_\star^4 = 2M_{\rm pl}^2|\dot{H}|c_s^5(1-c_s^2)^{-1}$. In slow-roll inflation, $c_s \to 1$, the strong coupling scale is much larger that the symmetry breaking scale. However, in models with small speed of sound, $c_s \ll 1$, this hierarchy of scales is reversed,

$$\frac{\Lambda_{\rm b}^4}{\Lambda_\star^4} = (1-c_s^2)c_s^{-4} \simeq 16\,(f_{\rm NL}^{\rm equil.})^2 \ , \tag{A.3.48}$$

where we used (A.3.46) to relate c_s to $f_{\rm NL}$. Therefore, any measurable non-Gaussianity ($f_{\rm NL}^{\rm equil.} \gtrsim 10$) requires the strong coupling scale to appear parametrically below the scale at which the background is integrated out.

Implications. The inherently strongly-coupled nature of the above theories was a result of restricting the particle content of the model. However, just like in particle physics, one should take this as an indication that new degrees of freedom may become important at energies below the scale of strong coupling.[11] Therefore, there is an energy scale $\omega_{\rm new}$ at which 'new physics' becomes important. If we wish to maintain both weak coupling and the effective small c_s–description at Hubble, we require $H^2 < \omega_{\rm new}^2 \ll \sqrt{2\pi}\Lambda_\star^2$. Given our previous results, we find

$$\frac{H^4}{\Lambda_\star^4} = 32\,\Delta_\zeta (f_{\rm NL}^{\rm equil})^2 \ . \tag{A.3.49}$$

where $\frac{\Delta_\zeta}{2\pi^2} = 2.4\times 10^{-9}$ and $|f_{\rm NL}^{\rm equil.}| \lesssim 300$. This implies that the new physics must enter not far above the Hubble scale:

$$H^2 \ < \ \omega_{\rm new}^2 \ \ll \ \sqrt{2\pi}\Lambda_\star^2 \ \approx \ \mathcal{O}(20)\left(\frac{f_{\rm NL}^{\rm equil.}}{100}\right)^{-1} H^2 \ . \tag{A.3.50}$$

This range of energies is sufficiently small that the new physics is not obviously decoupled at the Hubble scale. The use of "$\ll$" in (A.3.50) is a reminder that our loop expansion is being controlled by the ratio $\omega^2/\sqrt{2\pi}\Lambda_\star^2$. When $\omega_{\rm new}^2 \to \sqrt{2\pi}\Lambda_\star^2$ the theory becomes strongly coupled. However, quantum corrections of any observable become increasingly important as one approaches this limit. Therefore, a useful perturbative description requires that we expand in small $\omega^2/\sqrt{2\pi}\Lambda_\star^2$, to ensure that our description is not dominated by strong dynamics. This ratio reflects the strength of a coupling in any UV-completion of the effective theory.

A.4 Non-Gaussianity

At the single-derivative level the effective Goldstone action up to cubic order is

$$\mathcal{L}_2 \ = \ M_{\rm pl}^2 \dot{H}\left[\dot\pi^2 - (\partial_i\pi)^2\right] + 2M_2^4\,\dot\pi^2 \ , \tag{A.4.51}$$

$$\mathcal{L}_3 \ = \ -2M_2^4\,\dot\pi(\partial_i\pi)^2 + 2\Big(M_2^4 + \frac{2}{3}M_3^4\Big)\,\dot\pi^3 \ . \tag{A.4.52}$$

The parameter M_2 corrects the quadratic slow-roll lagrangian by adding the kinetic term $\dot\pi^2$, while not modifying the gradient term $(\partial_i\pi)^2$. (The coefficient of the gradient term is completely fixed by the background.) For $M_2^4 \gg M_{\rm pl}^2|\dot H|$ this leads to a small sound speed. In the same limit we get enhanced interactions in the cubic lagrangian. Making this symmetry explicit is one of the main insights of the effective theory of inflation. Moreover, we have identified two unique interactions in the cubic lagrangian – $\dot\pi(\partial_i\pi)^2$ and $\dot\pi^3$ – with amplitudes set by the two free parameters M_2 and M_3. These parameters characterize the simplest deviations from vanilla slow-roll inflation. Measurements of the CMB anisotropies can be used to put constraints on M_2 and M_3 (see Fig. A.1). These measurements are the precise analog of electroweak precision tests of the Standard Model.

[11] The 'new physics' could also take the form of a change in the physical description of the existing degrees of freedom.

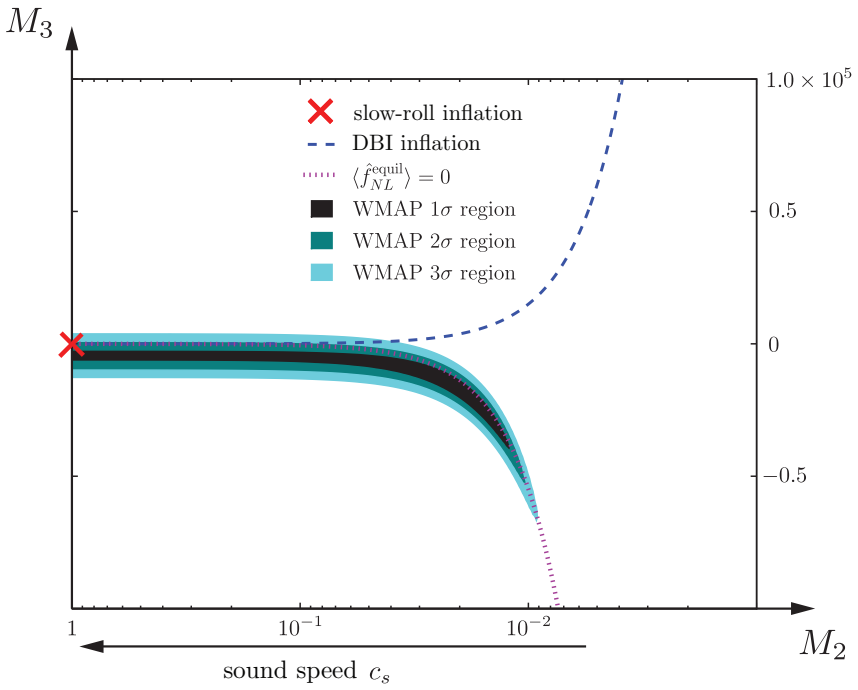


Figure A.1: *CMB Precision Tests. CMB data constrains the parameters M_2 and M_3 in the effective lagrangian: $\mathcal{L} = M_{\rm pl}^2 \dot{H}(\partial_\mu \pi)^2 + 2M_2^4(\dot{\pi}^2 - \dot{\pi}(\partial_\mu \pi)^2) + \frac{4}{3} M_3^4 \dot{\pi}^3$.*

The two operators $\dot{\pi}(\partial_i \pi)^2$ and $\dot{\pi}^3$ provide a physically motivated basis to define two unique shapes of non-Gaussianity. Being both purely derivative interactions, both $\dot{\pi}(\partial_i \pi)^2$ and $\dot{\pi}^3$ produce bispectra that peak in the *equilateral* momentum configuration. However, the bispectra are not identical, so we can find two linear combinations – called the *equilateral* and *orthogonal* shapes – that are mutually orthogonal is a well-defined sense.

Single-field bispectra. The bispectrum associated with the $\dot{\zeta}(\partial_i \zeta)^2 = H^3 \dot{\pi}(\partial_i \pi)^2$ interaction is

$$B_{\dot{\zeta}(\partial_i \zeta)^2} = -\frac{\Delta_\zeta^2}{4}\left(1 - \frac{1}{c_s^2}\right) \cdot \frac{24{K_3}^3 - 8{K_2}^2{K_3}^3 K_1 - 8{K_2}^4{K_1}^2 + 22{K_3}^3{K_1}^3 - 6{K_2}^2{K_1}^4 + 2{K_1}^6}{{K_3}^9{K_1}^3} \,,$$

where we defined

$$\begin{aligned} K_1 &= k_1 + k_2 + k_3 \,, \\ K_2 &= (k_1 k_2 + k_2 k_3 + k_3 k_1)^{1/2} \,, \\ K_3 &= (k_1 k_2 k_3)^{1/3} \,. \end{aligned}$$

The bispectrum associated with the $\dot{\zeta}^3$ interaction is

$$B_{\dot{\zeta}^3} = 4\Delta_\zeta^2 \left(\tilde{c}_3 + \frac{3}{2} c_s^2\right)\left(1 - \frac{1}{c_s^2}\right) \cdot \frac{1}{{K_3}^3{K_1}^3} \,,$$

where the parameter $\tilde{c}_3$ is defined via

$$M_3^4 \equiv \tilde{c}_3 \cdot \frac{M_2^4}{c_s^2} \,.$$

A.5 Conclusions

The effective theory of inflation is nice because:

- cosmologists can stop apologizing for using scalar fields;
- low-energy limit is constrained by symmetry;
- systematic classification of all single-field models of inflation;
- physical scales are readily identified and interpreted;
- Goldstone language explains decoupling from gravity;
- systematic classification of non-Gaussianities.

B Cosmological Perturbation Theory

In this appendix, we summarize basic facts of cosmological perturbation theory.

B.1 The Perturbed Universe

We consider perturbations to the homogeneous background spacetime and the stress-energy of the universe.

B.1.1 Metric Perturbations

The most general first-order perturbation to a spatially flat FRW metric is

$$\mathrm{d}s^2 = -(1+2\Phi)\mathrm{d}t^2 + 2a(t)B_i\mathrm{d}x^i\mathrm{d}t + a^2(t)[(1-2\Psi)\delta_{ij} + 2E_{ij}]\mathrm{d}x^i\mathrm{d}x^j \tag{B.1.1}$$

where Φ is a 3-scalar called the *lapse*, B_i is a 3-vector called the *shift*, Ψ is a 3-scalar called the spatial *curvature* perturbation, and E_{ij} is a spatial *shear* 3-tensor which is symmetric and traceless, $E^i_i = \delta^{ij}E_{ij} = 0$. 3-surfaces of constant time t are called *slices* and curves of constant spatial coordinates x^i but varying time t are called *threads*.

B.1.2 Matter Perturbations

The stress-energy tensor may be described by a density ρ, a pressure p, a 4-velocity u^μ (of the frame in which the 3-momentum density vanishes), and an anisotropic stress $\Sigma^{\mu\nu}$.

Density and pressure perturbations are defined in an obvious way

$$\delta\rho(t,x^i) \equiv \rho(t,x^i) - \bar{\rho}(t)\,, \qquad \text{and} \qquad \delta p(t,x^i) \equiv p(t,x^i) - \bar{p}(t)\,. \tag{B.1.2}$$

Here, the background values have been denoted by overbars. The 4-velocity has only three independent components (after the metric is fixed) since it has to satisfy the constraint $g_{\mu\nu}u^\mu u^\nu = -1$. In the perturbed metric (B.1.1) the perturbed 4-velocity is

$$u_\mu \equiv (-1-\Phi, v_i)\,, \qquad \text{or} \qquad u^\mu \equiv (1-\Phi, v^i + B^i)\,. \tag{B.1.3}$$

Here, u_0 is chosen so that the constraint $u_\mu u^\mu = -1$ is satisfied to first order in all perturbations. Anisotropic stress vanishes in the unperturbed FRW universe, so $\Sigma^{\mu\nu}$ is a first-order perturbation. Furthermore, $\Sigma^{\mu\nu}$ is constrained by

$$\Sigma^{\mu\nu}u_\nu = \Sigma^\mu_\mu = 0\,. \tag{B.1.4}$$

The orthogonality with u_μ implies $\Sigma^{00} = \Sigma^{0j} = 0$, *i.e.* only the spatial components Σ^{ij} are non-zero. The trace condition then implies $\Sigma^i_i = 0$. Anisotropic stress is therefore a traceless symmetric 3-tensor.

Finally, with these definitions the perturbed stress-tensor is

$$T^0_0 = -(\bar\rho + \delta\rho) \,, \tag{B.1.5}$$
$$T^0_i = (\bar\rho + \bar p) v_i \,, \tag{B.1.6}$$
$$T^i_0 = -(\bar\rho + \bar p)(v^i + B^i) \,, \tag{B.1.7}$$
$$T^i_j = \delta^i_j(\bar p + \delta p) + \Sigma^i_j \,. \tag{B.1.8}$$

If there are several contributions to the stress-energy tensor (*e.g.* photons, baryons, dark matter, etc.), they are added: $T_{\mu\nu} = \sum_I T^I_{\mu\nu}$. This implies

$$\delta\rho = \sum_I \delta\rho_I \,, \tag{B.1.9}$$
$$\delta p = \sum_I \delta p_I \,, \tag{B.1.10}$$
$$(\bar\rho + \bar p) v^i = \sum_I (\bar\rho_I + \bar p_I) v^i_I \,, \tag{B.1.11}$$
$$\Sigma^{ij} = \sum_I \Sigma^{ij}_I \,. \tag{B.1.12}$$

Density, pressure and anisotropic stress perturbations simply add. However, velocities do not add, which motivates defining the 3-momentum density

$$\delta q^i \equiv (\bar\rho + \bar p) v^i \,, \tag{B.1.13}$$

such that

$$\delta q^i = \sum_I \delta q^i_I \,. \tag{B.1.14}$$

B.2 Scalars, Vectors and Tensors

The Einstein Equations relate metric perturbations to the stress-energy perturbations. Einstein's Equations are both complicated (coupled second-order partial differential equations) and non-linear. Fortunately, the symmetries of the flat FRW background spacetime allow perturbations to be decomposed into independent scalar, vector and tensor components. This reduces the Einstein Equations to a set of uncoupled ordinary differential equations.

B.2.1 Helicity and SVT-Decomposition in Fourier Space

The decomposition into scalar, vector and tensor perturbations is most elegantly explained in Fourier space. We define the Fourier components of a general perturbation $\delta Q(t, \mathbf{x})$ as follows

$$\delta Q(t, \mathbf{k}) = \int \mathrm{d}^3\mathbf{x}\, \delta Q(t, \mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} \,. \tag{B.2.15}$$

First note that as a consequence of translation invariance different Fourier modes (different wavenumbers k) evolve independently.[1]

[1] The following proof was related to me by Uros Seljak and Chris Hirata.

Proof:

Consider the linear evolution of N perturbations δQ_I, $I = 1, \ldots, N$ from an initial time t_1 to a final time t_2

$$\delta Q_I(t_2, \mathbf{k}) = \sum_{J=1}^{N} \int \mathrm{d}^3\bar{\mathbf{k}}\; T_{IJ}(t_2, t_1, \mathbf{k}, \bar{\mathbf{k}}) \delta Q_J(t_1, \bar{\mathbf{k}})\,, \tag{B.2.16}$$

where the transfer matrix $T_{IJ}(t_2, t_1, \mathbf{k}, \bar{\mathbf{k}})$ follows from the Einstein Equations and we have allowed for the possibility of a mixing of k-modes. We now show that translation invariance in fact forbids such couplings. Consider the coordinate transformation

$$x^{i'} = x^i + \Delta x^i\,, \qquad \text{where} \quad \Delta x^i = const. \tag{B.2.17}$$

You may convince yourself that the Fourier amplitude gets shifted as follows

$$\delta Q'_I(t, \mathbf{k}) = e^{-ik_j \Delta x^j} \delta Q_I(t, \mathbf{k})\,. \tag{B.2.18}$$

Thus the evolution equation in the primed coordinate system becomes

$$\delta Q'_I(t_2, \mathbf{k}) = \sum_{J=1}^{N} \int \mathrm{d}^3\bar{\mathbf{k}}\; e^{-ik_j \Delta x^j} T_{IJ}(t_2, t_1, \mathbf{k}, \bar{\mathbf{k}}) e^{i\bar{k}_j \Delta x^j}\, \delta Q'_J(t_1, \bar{\mathbf{k}}) \tag{B.2.19}$$

$$\equiv \sum_{J=1}^{N} \int \mathrm{d}^3\bar{\mathbf{k}}\; T'_{IJ}(t_2, t_1, \mathbf{k}, \bar{\mathbf{k}}) \delta Q_J(t_1, \bar{\mathbf{k}})\,. \tag{B.2.20}$$

By translation invariance the equations of motion must be the same in both coordinate systems, *i.e.* the transfer matrices T_{IJ} and T'_{IJ} must be the same

$$T_{IJ}(t_2, t_1, \mathbf{k}, \bar{\mathbf{k}}) = e^{i(\bar{k}_j - k_j)\Delta x^j} T_{IJ}(t_2, t_1, \mathbf{k}, \bar{\mathbf{k}})\,. \tag{B.2.21}$$

This must hold for all Δx^j. Hence, either $\bar{\mathbf{k}} = \mathbf{k}$ or $T_{IJ}(t_2, t_1; \mathbf{k}, \bar{\mathbf{k}}) = 0$, *i.e.* the perturbation $\delta Q_I(t_2, \mathbf{k})$ of wavevector $\mathbf{k}$ depends only on the initial perturbations of wavevector $\mathbf{k}$. At linear order there is no coupling of different k-modes. QED.

Now consider rotations around the Fourier vector $\mathbf{k}$ by an angle ψ. We classify perturbations according to their *helicity* m: a perturbation of helicity m has its amplitude multiplied by $e^{im\psi}$ under the above rotation. We define scalar, vector and tensor perturbations as having helicities 0, ± 1, ± 2, respectively.

Consider a Fourier mode with wavevector $\mathbf{k}$. Without loss of generality we may assume that $\mathbf{k} = (0, 0, k)$ (or use rotational invariance of the background). The spatial dependence of any perturbation then is

$$\delta Q \propto e^{ikx^3}\,. \tag{B.2.22}$$

To study rotations around $\mathbf{k}$ it proves convenient to switch to the helicity basis

$$\mathbf{e}_\pm \equiv \frac{\mathbf{e}_1 \pm i\mathbf{e}_2}{\sqrt{2}}\,, \qquad \mathbf{e}_3\,, \tag{B.2.23}$$

where $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is the Cartesian basis. A rotation around the 3-axis by an angle ψ has the following effect

$$\begin{pmatrix} x^{1'} \\ x^{2'} \end{pmatrix} = \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix}\,, \qquad x^{3'} = x^3\,, \tag{B.2.24}$$

and

$$\mathbf{e}'_{\pm} = e^{\pm i\psi}\mathbf{e}_{\pm}\,, \qquad \mathbf{e}'_3 = \mathbf{e}_3\ . \tag{B.2.25}$$

The contravariant components of any tensor $T_{i_1 i_2 \ldots i_n}$ transform as

$$T'_{i_1 i_2 \ldots i_n} = e^{i(n_+ - n_-)\psi} T_{i_1 i_2 \ldots i_n} \equiv e^{im\psi} T_{i_1 i_2 \ldots i_n}\ , \tag{B.2.26}$$

where n_+ and n_- count the number of plus and minus indices in $i_1 \ldots i_n$, respectively. Helicity is defined as the difference $m \equiv n_+ - n_-$.

In the helicity basis $\{\mathbf{e}_{\pm}, \mathbf{e}_3\}$, a 3- scalar α has a single component with no indicies and is therefore obviously of helicity 0; a 3-vector β_i has 3 components $\beta_+, \beta_-, \beta_3$ of helicity ± 1 and 0; a symmetric and traceless 3-tensor γ_{ij} has 5 components $\gamma_{--}, \gamma_{++}, \gamma_{-3}, \gamma_{+3}, \gamma_{33}$ (the tracelessness condition makes γ_{-+} redundant), of helicity ± 2, ± 1 and 0.

Rotational invariance of the background implies that helicity scalars, vectors and tensors evolve independently.[2]

Proof:

Consider N perturbations δQ_I, $I = 1, \ldots, N$ of helicity m_I. The linear evolution is

$$\delta Q_I(t_2, \mathbf{k}) = \sum_{J=1}^{N} T_{IJ}(t_2, t_1, \mathbf{k}) \delta Q_J(t_1, \mathbf{k})\ , \tag{B.2.27}$$

where the transfer matrix $T_{IJ}(t_2, t_1, \mathbf{k})$ follows from the Einstein Equations. Under rotation the perturbations transform as

$$\delta Q'_I(t, \mathbf{k}) = e^{im_I\psi} \delta Q_I(t, \mathbf{k})\ , \tag{B.2.28}$$

and

$$\delta Q'_I(t_2, \mathbf{k}) = \sum_{J=1}^{N} e^{im_I\psi}\, T_{IJ}(t_2, t_1, \mathbf{k})\, e^{-im_J\psi} \delta Q'_J(t_1, \mathbf{k})\ . \tag{B.2.29}$$

By rotational invariance of the equations of motion

$$T_{IJ}(t_2, t_1, \mathbf{k}) = e^{im_I\psi}\, T_{IJ}(t_2, t_1, \mathbf{k})\, e^{-im_J\psi} = e^{i(m_I - m_J)\psi} T_{IJ}(t_2, t_1, \mathbf{k})\,, \tag{B.2.30}$$

which has to hold for any angle ψ; it follows that eithers $m_I = m_J$, *i.e.* δQ_I and δQ_J have the same helicity or $T_{IJ}(t_2, t_1, \mathbf{k}) = 0$. This proves that the equations of motion don't mix modes of different helicity. QED.

B.2.2 SVT-Decomposition in Real Space

In the last section we have seen that 3-scalars correspond to helicity scalars, 3-vectors decompose into helicity scalars and vectors, and 3-tensors decompose into helicity scalars, vectors and tensors. We now look at this from a different perspective.

A 3-scalar is obviously also a helicity scalar

$$\alpha = \alpha^S\,. \tag{B.2.31}$$

[2] The following proof was related to me by Uros Seljak and Chris Hirata.

Consider a 3-vector β_i. We argue that it can be decomposed as

$$\beta_i = \beta_i^S + \beta_i^V \,, \tag{B.2.32}$$

where

$$\beta_i^S = \nabla_i \hat{\beta} \,, \qquad \nabla^i \beta_i^V = 0 \,, \tag{B.2.33}$$

or, in Fourier space,

$$\beta_i^S = -\frac{ik_i}{k}\beta \,, \qquad k_i \beta_i^V = 0 \,. \tag{B.2.34}$$

Here, we have defined $\beta \equiv k\hat{\beta}$.

Exercise 1 (Helicity Vector) *Show that β_i^V is a helicity vector.*

Similarly, a traceless, symmetric 3-tensor can be written as

$$\gamma_{ij} = \gamma_{ij}^S + \gamma_{ij}^V + \gamma_{ij}^T \,, \tag{B.2.35}$$

where

$$\gamma_{ij}^S = \left(\nabla_i \nabla_j - \frac{1}{3}\delta_{ij}\nabla^2\right)\hat{\gamma} \tag{B.2.36}$$

$$\gamma_{ij}^V = \frac{1}{2}\left(\nabla_i \hat{\gamma}_j + \nabla_j \hat{\gamma}_i\right) \,, \quad \nabla_i \hat{\gamma}_i = 0 \tag{B.2.37}$$

$$\nabla_i \gamma_{ij}^T = 0 \,. \tag{B.2.38}$$

or

$$\gamma_{ij}^S = \left(-\frac{k_i k_j}{k^2} + \frac{1}{3}\delta_{ij}\right)\gamma \tag{B.2.39}$$

$$\gamma_{ij}^V = -\frac{i}{2k}\left(k_i \gamma_j + k_j \gamma_i\right) \,, \quad k_i \gamma_i = 0 \tag{B.2.40}$$

$$k_i \gamma_{ij}^T = 0 \,. \tag{B.2.41}$$

Here, we have defined $\gamma \equiv k^2\hat{\gamma}$ and $\gamma_i \equiv k\hat{\gamma}_i$.

Exercise 2 (Helicity Vectors and Tensors) *Show that γ_{ij}^V and γ_{ij}^T are a helicity vector and a helicity tensor, respectively.*

Choosing $\mathbf{k}$ along the 3-axis, *i.e.* $\mathbf{k} = (0, 0, k)$ we find

$$\gamma_{ij}^S = \frac{1}{3}\begin{pmatrix} \gamma & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 0 & -2\gamma \end{pmatrix} \,, \tag{B.2.42}$$

$$\gamma_{ij}^V = -\frac{i}{2}\begin{pmatrix} 0 & 0 & \gamma_1 \\ 0 & 0 & \gamma_2 \\ \gamma_1 & \gamma_2 & 0 \end{pmatrix} \,, \tag{B.2.43}$$

$$\gamma_{ij}^T = \begin{pmatrix} \gamma^\times & \gamma^+ & 0 \\ \gamma^+ & -\gamma^\times & 0 \\ 0 & 0 & 0 \end{pmatrix} \,. \tag{B.2.44}$$

B.3 Scalars

B.3.1 Metric Perturbations

Four scalar metric perturbations Φ, $B_{,i}$, $\Psi\delta_{ij}$ and $E_{,ij}$ may be constructed from 3-scalars, their derivatives and the background spatial metric, *i.e.*

$$\mathrm{d}s^2 = -(1+2\Phi)\mathrm{d}t^2 + 2a(t)B_{,i}\mathrm{d}x^i\mathrm{d}t + a^2(t)[(1-2\Psi)\delta_{ij} + 2E_{,ij}]\mathrm{d}x^i\mathrm{d}x^j \tag{B.3.45}$$

Here, we have absorbed the $\nabla^2 E\,\delta_{ij}$ part of the helicity scalar E^S_{ij} in $\Psi\,\delta_{ij}$.

The intrinsic Ricci scalar curvature of constant time hypersurfaces is

$$R_{(3)} = \frac{4}{a^2}\nabla^2\Psi\,. \tag{B.3.46}$$

This explains why Ψ is often referred to as the curvature perturbation.

There are two scalar gauge transformations

$$t \;\rightarrow\; t+\alpha\;, \tag{B.3.47}$$

$$x^i \;\rightarrow\; x^i + \delta^{ij}\beta_{,j}\;. \tag{B.3.48}$$

Under these coordinate transformations the scalar metric perturbations transform as

$$\Phi \;\rightarrow\; \Phi - \dot{\alpha}\;, \tag{B.3.49}$$

$$B \;\rightarrow\; B + a^{-1}\alpha - a\dot{\beta}\;, \tag{B.3.50}$$

$$E \;\rightarrow\; E - \beta\;, \tag{B.3.51}$$

$$\Psi \;\rightarrow\; \Psi + H\alpha\,. \tag{B.3.52}$$

Note that the combination $\dot{E} - B/a$ is independent of the spatial gauge and only depends on the temporal gauge. It is called the scalar potential for the anisotropic shear of world lines orthogonal to constant time hypersurfaces. To extract physical results it is useful to define gauge-invariant combinations of the scalar metric perturbations. Two important gauge-invariant quantities were introduced by Bardeen

$$\Phi_{\mathrm{B}} \;\equiv\; \Phi - \frac{d}{dt}[a^2(\dot{E} - B/a)]\;, \tag{B.3.53}$$

$$\Psi_{\mathrm{B}} \;\equiv\; \Psi + a^2 H(\dot{E} - B/a)\;. \tag{B.3.54}$$

B.3.2 Matter Perturbations

Matter perturbations are also gauge-dependent, *e.g.* density and pressure perturbations transform as follows under temporal gauge transformations

$$\delta\rho \rightarrow \delta\rho - \dot{\bar{\rho}}\,\alpha\,, \qquad \delta p \rightarrow \delta p - \dot{\bar{p}}\,\alpha\,. \tag{B.3.55}$$

Adiabatic pressure perturbations are defined as

$$\delta p_{ad} \equiv \frac{\dot{\bar{p}}}{\dot{\bar{\rho}}}\,\delta\rho\,. \tag{B.3.56}$$

The non-adiabiatic, or entropic, part of the pressure perturbations is then gauge-invariant

$$\delta p_{en} \equiv \delta p - \frac{\dot{\bar{p}}}{\dot{\bar{\rho}}}\,\delta\rho\,. \tag{B.3.57}$$

The scalar part of the 3-momentum density, $(\delta q)_{,i}$, transforms as

$$\delta q \to \delta q + (\bar{\rho} + \bar{p})\,\alpha\,. \tag{B.3.58}$$

We may then define the gauge-invariant comoving density perturbation

$$\delta\rho_m \equiv \delta\rho - 3H\delta q\,. \tag{B.3.59}$$

Finally, two important gauge-invariant quantities are formed from combinations of matter and metric perturbations. The *curvature perturbation on uniform density hypersurfaces* is

$$-\mathcal{R} \equiv \Psi + \frac{H}{\dot{\bar{\rho}}}\delta\rho\,. \tag{B.3.60}$$

The *comoving curvature perturbation* is

$$\zeta = \Psi - \frac{H}{\bar{\rho}+\bar{p}}\delta q\,. \tag{B.3.61}$$

We will show that ζ and $\mathcal{R}$ are equal on superhorizon scales, where they become time-independent. The computation of the inflationary perturbation spectrum is most clearly phrased in terms of ζ and $\mathcal{R}$.

B.3.3 Einstein Equations

To relate the metric and stress-energy perturbations, we consider the perturbed Einstein Equations

$$\delta G_{\mu\nu} = 8\pi G\,\delta T_{\mu\nu}\,. \tag{B.3.62}$$

We work at linear order. This leads to the *energy and momentum constraint equations*

$$3H(\dot{\Psi} + H\Phi) + \frac{k^2}{a^2}\left[\Psi + H(a^2\dot{E} - aB)\right] = -4\pi G\,\delta\rho \tag{B.3.63}$$

$$\dot{\Psi} + H\Phi = -4\pi G\,\delta q\,. \tag{B.3.64}$$

These can be combined into the gauge-invariant *Poisson Equation*

$$\frac{k^2}{a^2}\Psi_{\rm B} = -4\pi G\delta\rho_m\,. \tag{B.3.65}$$

The Einstein equation also yield two *evolution equations*

$$\ddot{\Psi} + 3H\dot{\Psi} + H\dot{\Phi} + (3H^2 + 2\dot{H})\Phi = 4\pi G\left(\delta p - \frac{2}{3}k^2\delta\Sigma\right) \tag{B.3.66}$$

$$(\partial_t + 3H)(\dot{E} - B/a) + \frac{\Psi - \Phi}{a^2} = 8\pi G\,\delta\Sigma\,. \tag{B.3.67}$$

The last equation may be written as

$$\Psi_{\rm B} - \Phi_{\rm B} = 8\pi G\,a^2\delta\Sigma\,. \tag{B.3.68}$$

In the absence of anisotropic stress this implies, $\Psi_{\rm B} = \Phi_{\rm B}$.

Energy-momentum conservation, $\nabla_\mu T^{\mu\nu} = 0$, gives the *continuity equation* and the *Euler equation*

$$\dot{\delta\rho} + 3H(\delta\rho + \delta p) = \frac{k^2}{a^2}\delta q + (\bar{\rho} + \bar{p})[3\dot{\Psi} + k^2(\dot{E} + B/a)]\,, \tag{B.3.69}$$

$$\dot{\delta q} + 3H\delta q = -\delta p + \frac{2}{3}k^2\delta\Sigma - (\bar{\rho} + \bar{p})\Phi\,. \tag{B.3.70}$$

Expressed in terms of the curvature perturbation on uniform-density hypersurfaces, ζ, (B.3.69) reads

$$\dot{\zeta} = -H\frac{\delta p_{en}}{\bar{\rho} + \bar{p}} - \Pi\,, \tag{B.3.71}$$

where δp_{en} is the non-adiabatic component of the pressure perturbation, and Π is the scalar shear along comoving worldlines

$$\frac{\Pi}{H} \equiv -\frac{k^2}{3H}\left[\dot{E} - B/a + \frac{\delta q}{a^2(\bar{\rho} + \bar{p})}\right] \tag{B.3.72}$$

$$= -\frac{k^2}{3a^2H^2}\left[\zeta - \Psi_{\rm B}\left(1 - \frac{2\bar{\rho}}{9(\bar{\rho} + \bar{p})}\frac{k^2}{a^2H^2}\right)\right]\,. \tag{B.3.73}$$

For adiabative perturbations, $\delta p_{en} = 0$ on superhorizon scales, $k/(aH) \ll 1$ (*i.e.* $\Pi/H \to 0$ for finite ζ and $\Psi_{\rm B}$), the curvature perturbation ζ is constant. This is a crucial result for our computation of the inflationary spectrum of ζ. It justifies computing ζ at horizon exit and ignoring superhorizon evolution.

B.3.4 Popular Gauges

For reference we now give the Einstein Equations and the conservation equations is various popular gauges:

- **Synchronous gauge**

  A popular gauge, especially for numerical implementation of the perturbation equations (*cf.* CMBFAST or CAMB), is synchronous gauge. It is defined by

  $$\Phi = B = 0\,. \tag{B.3.74}$$

  The Einstein equations become

  $$3H\dot{\Psi} + \frac{k^2}{a^2}\left[\Psi + Ha^2\dot{E}\right] = -4\pi G\,\delta\rho\ , \tag{B.3.75}$$

  $$\dot{\Psi} = -4\pi G\,\delta q\ , \tag{B.3.76}$$

  $$\ddot{\Psi} + 3H\dot{\Psi} = 4\pi G\left(\delta p - \frac{2}{3}k^2\delta\Sigma\right)\ , \tag{B.3.77}$$

  $$(\partial_t + 3H)\dot{E} + \frac{\Psi}{a^2} = 8\pi G\,\delta\Sigma\ . \tag{B.3.78}$$

  The conservation equation are

  $$\dot{\delta\rho} + 3H(\delta\rho + \delta p) = \frac{k^2}{a^2}\delta q + (\bar{\rho} + \bar{p})[3\dot{\Psi} + k^2\dot{E}]\ , \tag{B.3.79}$$

  $$\dot{\delta q} + 3H\delta q = -\delta p + \frac{2}{3}k^2\delta\Sigma\ . \tag{B.3.80}$$

- **Newtonian gauge**

  The Newtonian gauge has its name because it reduces to Newtonian gravity in the small-scale limit. It is popular for analytic work since it leads to algebraic relations between metric and stress-energy perturbations.

  Newtonian gauge is defined by

$$B = E = 0\,, \tag{B.3.81}$$

  and

$$\mathrm{d}s^2 - (1+2\Phi)\mathrm{d}t^2 + a^2(t)(1-2\Psi)\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j\,. \tag{B.3.82}$$

  The Einstein equations are

$$\begin{aligned} 3H(\dot{\Psi} + H\Phi) + \frac{k^2}{a^2}\Psi &= -4\pi G\,\delta\rho\,, && \text{(B.3.83)}\\ \dot{\Psi} + H\Phi &= -4\pi G\,\delta q\,, && \text{(B.3.84)}\\ \ddot{\Psi} + 3H\dot{\Psi} + H\dot{\Phi} + (3H^2 + 2\dot{H})\Phi &= 4\pi G\left(\delta p - \frac{2}{3}k^2\delta\Sigma\right)\,, && \text{(B.3.85)}\\ \frac{\Psi - \Phi}{a^2} &= 8\pi G\,\delta\Sigma\,. && \text{(B.3.86)} \end{aligned}$$

  The continuity equations are

$$\begin{aligned} \dot{\delta\rho} + 3H(\delta\rho + \delta p) &= \frac{k^2}{a^2}\delta q + 3(\bar{\rho} + \bar{p})\dot{\Psi}\,, && \text{(B.3.87)}\\ \dot{\delta q} + 3H\delta q &= -\delta p + \frac{2}{3}k^2\delta\Sigma - (\bar{\rho} + \bar{p})\Phi\,. && \text{(B.3.88)} \end{aligned}$$

- **Uniform density gauge**

  The uniform density gauge is useful for describing the evolution of perturbations on superhorizon scales. As its name suggests it is defined by

$$\delta\rho = 0\,. \tag{B.3.89}$$

  In addition, it is convenient to take

$$E = 0\,, \qquad -\Psi \equiv \mathcal{R}\,. \tag{B.3.90}$$

  The Einstein equations are

$$\begin{aligned} 3H(-\dot{\mathcal{R}} + H\Phi) - \frac{k^2}{a^2}\left[\mathcal{R} + aHB\right] &= 0 && \text{(B.3.91)}\\ -\dot{\mathcal{R}} + H\Phi &= -4\pi G\,\delta q\,, && \text{(B.3.92)}\\ -\ddot{\mathcal{R}} - 3H\dot{\mathcal{R}} + H\dot{\Phi} + (3H^2 + 2\dot{H})\Phi &= 4\pi G\left(\delta p - \frac{2}{3}k^2\delta\Sigma\right)\,, && \text{(B.3.93)}\\ (\partial_t + 3H)B/a + \frac{\mathcal{R} + \Phi}{a^2} &= -8\pi G\,\delta\Sigma\,. && \text{(B.3.94)} \end{aligned}$$

  The continuity equations are

$$\begin{aligned} 3H\delta p &= \frac{k^2}{a^2}\delta q + (\bar{\rho} + \bar{p})[-3\dot{\mathcal{R}} + k^2 B/a]\,, && \text{(B.3.95)}\\ \dot{\delta q} + 3H\delta q &= -\delta p + \frac{2}{3}k^2\delta\Sigma - (\bar{\rho} + \bar{p})\Phi\,. && \text{(B.3.96)} \end{aligned}$$

- **Comoving gauge**

  Comoving gauge is defined by the vanishing of the scalar momentum density,

$$\delta q = 0\,, \qquad E = 0\,. \tag{B.3.97}$$

  It is also conventional to set $-\Psi \equiv \zeta$ in this gauge.

  The Einstein equations are

$$\begin{aligned} 3H(-\dot{\zeta} + H\Phi) + \frac{k^2}{a^2}\left[-\zeta - aHB\right] &= -4\pi G\,\delta\rho & \text{(B.3.98)}\\ -\dot{\zeta} + H\Phi &= 0\,, & \text{(B.3.99)}\\ -\ddot{\zeta} - 3H\dot{\zeta} + H\dot{\Phi} + (3H^2 + 2\dot{H})\Phi &= 4\pi G\left(\delta p - \frac{2}{3}k^2\delta\Sigma\right)\,, & \text{(B.3.100)}\\ (\partial_t + 3H)B/a + \frac{\zeta + \Phi}{a^2} &= -8\pi G\,\delta\Sigma\,. & \text{(B.3.101)} \end{aligned}$$

  The continuity equations are

$$\begin{aligned} \dot{\delta\rho} + 3H(\delta\rho + \delta p) &= (\bar{\rho} + \bar{p})[-3\dot{\zeta} + k^2 B/a]\,, & \text{(B.3.102)}\\ 0 &= -\delta p + \frac{2}{3}k^2\delta\Sigma - (\bar{\rho} + \bar{p})\Phi\,. & \text{(B.3.103)} \end{aligned}$$

  Equations (B.3.103) and (B.3.99) may be combined into

$$\Phi = \frac{-\delta p + \frac{2}{3}\Sigma}{\bar{\rho} + \bar{p}}\,, \qquad kB = \frac{4\pi G a^2 \delta\rho - k^2\mathcal{R}}{aH}\,. \tag{B.3.104}$$

- **Spatially-flat gauge**

  A convenient gauge for computing inflationary perturbation is spatially-flat gauge

$$\Psi = E = 0\,. \tag{B.3.105}$$

  During inflation all scalar perturbations are then described by $\delta\phi$.

  The Einstein equations are

$$\begin{aligned} 3H^2\Phi + \frac{k^2}{a^2}\left[-aHB)\right] &= -4\pi G\,\delta\rho & \text{(B.3.106)}\\ H\Phi &= -4\pi G\,\delta q & \text{(B.3.107)}\\ H\dot{\Phi} + (3H^2 + 2\dot{H})\Phi &= 4\pi G\left(\delta p - \frac{2}{3}k^2\delta\Sigma\right) & \text{(B.3.108)}\\ (\partial_t + 3H)B/a + \frac{\Phi}{a^2} &= -8\pi G\,\delta\Sigma\,. & \text{(B.3.109)} \end{aligned}$$

  The continuity equations are

$$\begin{aligned} \dot{\delta\rho} + 3H(\delta\rho + \delta p) &= \frac{k^2}{a^2}\delta q + (\bar{\rho} + \bar{p})[k^2 B/a]\,, & \text{(B.3.110)}\\ \dot{\delta q} + 3H\delta q &= -\delta p + \frac{2}{3}k^2\delta\Sigma - (\bar{\rho} + \bar{p})\Phi\,. & \text{(B.3.111)} \end{aligned}$$

B.4 Vectors

B.4.1 Metric Perturbations

Vector type metric perturbations are defined as

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + 2a(t)S_i\mathrm{d}x^i\mathrm{d}t + a^2(t)[\delta_{ij} + 2F_{(i,j)}]\mathrm{d}x^i\mathrm{d}x^j\,, \tag{B.4.112}$$

where $S_{i,i} = F_{i,i} = 0$. The vector gauge transformation is

$$x^i \to x^i + \beta^i\,, \qquad \beta_{i,i} = 0\,. \tag{B.4.113}$$

They lead to the transformations

$$S_i \;\to\; S_i + a\dot{\beta}_i\,, \tag{B.4.114}$$

$$F_i \;\to\; F_i - \beta_i\,. \tag{B.4.115}$$

The combination $\dot{F}_i + S_i/a$ is called the gauge-invariant vector shear perturbation.

B.4.2 Matter Perturbations

We define the vector part of the anisotropic stress by

$$\delta\Sigma_{ij} = \partial_{(i}\Sigma_{j)}\,, \tag{B.4.116}$$

where Σ_i is divergence-free, $\Sigma_{i,i} = 0$.

B.4.3 Einstein Equations

For vector perturbations there are only two Einstein Equations,

$$\dot{\delta q_i} + 3H\delta q_i = k^2\delta\Sigma_i\,, \tag{B.4.117}$$

$$k^2(\dot{F}_i + S_i/a) = 16\pi G\,\delta q_i\,. \tag{B.4.118}$$

In the absence of anisotropic stress ($\delta\Sigma_i = 0$) the divergence-free momentum δq_i decays with the expansion of the universe; see Eqn. (B.4.117). The shear perturbation $\dot{F}_i + S_i/a$ then vanishes by Eqn. (B.4.118). Under most circumstances vector perturbations are therefore subdominant. They won't play an important role in these lectures. In particular, vector perturbations aren't created by inflation.

B.5 Tensors

B.5.1 Metric Perturbations

Tensor metric perturbations are defined as

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)[\delta_{ij} + h_{ij}]\mathrm{d}x^i\mathrm{d}x^j\,, \tag{B.5.119}$$

where $h_{ij,i} = h^i_i = 0$. Tensor perturbations are automatically gauge-invariant (at linear order). It is conventional to decompose tensor perturbations into eigenmodes of the spatial Laplacian, $\nabla^2 e_{ij} = -k^2 e_{ij}$, with comoving wavenumber k and scalar amplitude $h(t)$,

$$h_{ij} = h(t)e_{ij}^{(+,\times)}(x)\,. \tag{B.5.120}$$

Here, $+$ and $\times$ denote the two possible polarization states.

B.5.2 Matter Perturbations

Tensor perturbations are sourced by anisotropic stress Σ_{ij}, with $\Sigma_{ij,i} = \Sigma^i_i = 0$. It is typically a good approximation to assume that the anisotropic stress is negligible, although a small amplitude is induced by neutrino free-streaming.

B.5.3 Einstein Equations

For tensor perturbations there is only one Einstein Equation. In the absence of anisotropic stress this is

$$\ddot{h} + 3H\dot{h} + \frac{k^2}{a^2}h = 0\,. \tag{B.5.121}$$

This is a wave equation describing the evolution of gravitational waves in an expanding universe. Gravitational waves are produced by inflation, but then decay with the expansion of the universe. However, at recombination their amplitude may still be large enough to leave distinctive signatures in B-modes of CMB polarization.

C Exercises

Exercises for Chapter 1

Problem 1 (Horizon Problem)

- Consider a FRW metric in conformal coordinates, $\mathrm{d}s^2 = a(\tau)^2[-\mathrm{d}\tau^2 + \mathrm{d}\mathrm{x}^2]$. The scale factor in front of the whole metric does not affect the propagation of light rays and therefore does not affect causality. So why does the condition $\ddot{a} > 0$ solve the horizon problem?

- Suppose all matter fields — photons for example — are coupled not directly to the metric $g_{\mu\nu}$, but to $\tilde{g}_{\mu\nu} = h(\phi)g_{\mu\nu}$, where ϕ is a scalar field, evolving in time and h is a given function. We are assuming to be in 'Einstein frame', i.e. that the action for $g_{\mu\nu}$ is just the standard Einstein-Hilbert action. Is it enough to have acceleration of the "effective" scale factor $\tilde{a}$ of the metric $\tilde{g}$ to solve the horizon problem?

- Assuming instantaneous reheating after inflation at temperature T_{rh} show, using entropy conservation, that we need at least

$$N = 46 + \log \frac{T_{\mathrm{rh}}}{10^{10}\,\mathrm{GeV}} + \frac{1}{2}\log|\Omega_i - 1|$$

e-folds of inflation to solve the flatness problem. Here, Ω_i is the curvature parameter when inflation begins.

Problem 2 *In a bouncing model, the scale factor is initially contracting, then reaches a minimum (the bounce) and then starts a (decelerated) expansion. Compare the diagram $k^{-1}a$ vs. H^{-1} for inflation and for bouncing models. Discuss how to realize a bounce.*

Problem 3 *The ratio between pressure and energy density is usually called $w \equiv p/\rho$. What are the possible values of w for a scalar field, with standard kinetic term and potential, which evolves in time, assuming $\rho > 0$? And if we assume a positive definite potential?*

Problem 3 ($\lambda\phi^4$ Inflation) *Determine the predictions of an inflationary model with a quartic potential*

$$V(\phi) = \lambda\phi^4\,.$$

1. *Compute the slow-roll parameters ϵ and η in terms of ϕ.*

2. *Determine ϕ_{end}, the value of the field at which inflation ends.*

3. *To determine the spectrum, you will need to evaluate ϵ and η at horizon crossing, $k = aH$ (or $-k\tau = 1$). Choose the wavenumber k to be equal to $a_0 H_0$, roughly the horizon today. Show that the requirement $-k\tau = 1$ then corresponds to*

$$e^{60} = \int_0^N \mathrm{d}N' \frac{e^{N'}}{H(N')/H_{\mathrm{end}}} ,$$

where H_{end} is the Hubble rate at the end of inflation, and N is defined to be the number of e-folds before the end of inflation

$$N \equiv \ln\left(\frac{a_{\mathrm{end}}}{a}\right) .$$

4. *Take this Hubble rate to be a constant in the above with $H/H_{\mathrm{end}} = 1$. This implies that $N \approx 60$. Turn this into an expression for ϕ. This simplest way to do this is to note that $N = \int_t^{\mathrm{t_{end}}} \mathrm{d}t' H(t')$ and assume that H is dominated by potential energy. Show that this mode leaves the horizon when $\phi = 22 M_{\mathrm{pl}}$.*

5. *Determine the predicted values of n_{s}, r and n_{t}. Compare these predictions to the latest CMB data.*

6. *Estimate the scalar amplitude in terms of λ. Set $\Delta_{\mathrm{s}}^2 \approx 10^{-9}$. What value does this imply for λ?*

This model illustrates many of the features of generic inflationary models: (i) the field is of order – even greater than – the Planck scale, but (ii) the energy scale V is much smaller than Planckian because of (iii) the very small coupling constant.

Exercises for Chapter 2

Problem 1

- An harmonic oscillator of frequency ω_i is in its vacuum state. Its frequency is instantaneously changed to ω_f. Write the state of the system in terms of the new eigenstates and calculate its energy. This integral of Hermite polynomials may be useful:

$$\int_{-\infty}^{+\infty} e^{-x^2} H_{2m}(xy)\mathrm{d}x = \sqrt{\pi}\frac{(2m)!}{m!}(y^2-1)^m \ .$$

- An harmonic oscillator of frequency ω_i is in its first excited state and its frequency is instantaneously changed to $\omega_f \ll \omega_i$. What is its final energy? [No long calculation is needed.]

Problem 2 *Photons are massless, but they are not produced during inflation. Why?*

Problem 3 *Calculate the equal time 2-point function of a massless scalar in a fixed de Sitter background in real space. What is the physical meaning of the IR divergence?*

Problem 4 *Using symmetry and simple scaling arguments, calculate the tilt of the spectrum of a scalar with small mass, $m^2 \ll H^2$, in a fixed de Sitter background.*

Exercises for Chapter 3

Problem 1

The objective of this problem is to reproduce the results of Seljak 94 (S94). Starting with the equations in Ma & Bertschinger 95 derive equation (3) in S94. Derive the solutions for the evolution of the background quantities y, η as a function of x. Write a routine in Mathematica (using NDSolve) that solves the equations for the perturbations up to recombination for a given value of κ and cosmological parameters. Reproduce the top panel of figure 1. Derive equation (5) starting from the integral solution. Make a spline of the sources at recombination as a function of κ and integrate them to obtain C_ℓ. Reproduce the bottom panel of figure 2. Include Silk damping as done in S94. For the more ambitious you can include damping by adding shear viscosity directly to the equations (see Mukhanovs book).

Problem 2

Use the equations derived in the previous problem to determine the parameters of the cosmological model that govern the dynamics of the perturbations up to recombination. What is the effect of changing the distance to the last scattering surface? Assume you want to determine Ω_m, Ω_b, Ω_Λ and h using the CMB temperature power spectra. Argue that there is a degeneracy between parameters.

Problem 3

The objective of this problem is to get a sense of how the temperature power spectrum depends on the cosmological parameters and how well current data can determine these parameters. We will consider the following parameters to describe the matter content of the universe (Ω_Λ, $\Omega_b h^2$, $\Omega_b h^2$) and we will restrict ourselves to flat models, ($\Omega_k = 0$). To describe the power spectrum of initial curvature fluctuations we will use the amplitude A_s and the spectral index n_s. We will not consider gravity waves.

We will compare theoretical models with the latest WMAP. All the necessary ingredients can be found in LAMBDA http://lambda.gsfc.nasa.gov/*. Use CAMB online tool from LAMBDA to compute the temperature power spectra for the WMAP best fit model. You can choose the specific parameter from the WMAP best fit table also in LAMBDA. Produce three families of models with one parameter of (Ω_Λ, $\Omega_b h^2$, $\Omega_b h^2$) varying in each. Produce plots that show the C_ℓ for each family. In these plots also show the WMAP data with its corresponding error bars (you can download a table with the power spectra and its error bars from LAMBDA). For each family explain the physics behind the changes in the power spectra. Roughly estimate (by comparing the changes produced by each parameter with the error bars in the data) the range of values for each parameter seems acceptable. Compare with what is given in the table of best fit parameters.*

Exercises for Chapter 4

Problem 1 *Consider chaotic inflation with* $V(\phi) = \frac{1}{2}m^2\phi^2$ *and the coupling* $g^2\phi^2\chi^2$. *Broad resonance occurs for* $g > \frac{2m}{\Phi} \sim 10^{-6}$, *where we used* $\Phi \sim M_{\rm pl}$ *and* $m \sim 10^{-6}M_{\rm pl}$ *(from COBE normalization). Is such a large coupling consistent with naturalness of the inflationary potential? Consider the one-loop correction to the inflaton mass*

$$\delta m^2 = \frac{g^2}{16\pi^2}\Lambda_{\rm uv}^2 \sim \frac{g^2}{16\pi^2}M_{\rm pl}^2 \ .$$

Naturalness requires $\delta m < m \sim 10^{-6}M_{\rm pl}$, *or* $g < 10^{-5}$. *This seems to disallow the regime of broad parametric resonance for chaotic inflation. The reheating of this model is now predominantly through the elementary decays of* ϕ *and narrow resonances in* χ.

Exercises for Chapter 5

Problem 1 *Using symmetry arguments, show that the n-point function of ζ in Fourier space in a generic model of inflation is of the form*

$$\langle \zeta_{\mathbf{k}_1} \cdots \zeta_{\mathbf{k}_n} \rangle = (2\pi)^3 \delta(\mathbf{k}_1 + \cdots + \mathbf{k}_n) F(k_n) \ , \tag{C.0.1}$$

where F is an homogeneous function of the ks of degree $-3(n-1)$.

Problem 2 *Consider a massless scalar ϕ in de Sitter space with an interaction $M\phi^3$. Calculate the 3-point function $\langle \phi_{\mathbf{k}_1} \phi_{\mathbf{k}_2} \phi_{\mathbf{k}_3} \rangle$.*

Exercises for Chapter 9

Problem 1

Consider a probe D3-brane in an $AdS_5 \times X_5$ throat

$$\mathrm{d}s^2 = \frac{r^2}{R^2}(-\mathrm{d}t^2 + \mathrm{d}\mathbf{x}^2) + \frac{R^2}{r^2}\mathrm{d}r^2 + \mathrm{d}s^2_{X_5} \ .$$

Imagine the throat is cut off at $r = r_{\rm uv} \sim R$ and connected to a compactification. The brane can move in r, with canonically normalized field $\phi \sim r/\alpha'$. Compute the four-dimensional Planck mass as a function of the maximum value of the canonically normalized field $\phi_{\rm uv} \sim r_{\rm uv}/\alpha'$ and the geometry of the compact dimensions. Using the Lyth bound, convert this to a bound on the tensor to scalar ratio as a function of these quantities.

Problem 2

- Consider the DBI action

$$S = -\int \mathrm{d}^4x\sqrt{-g}\,\frac{\phi^4}{\lambda}\sqrt{1 - \lambda(\partial_\mu\phi)^2/\phi^4} - V(\phi) \ .$$

Derive the stress-energy tensor. Show explicitly that inflation can occur even on a steep potential that does not satisfy the slow-roll conditions.

- Generalize this to

$$S = \int \mathrm{d}^4x\sqrt{-g}\,P(X,\phi) \ , \qquad X \equiv -g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi \ .$$