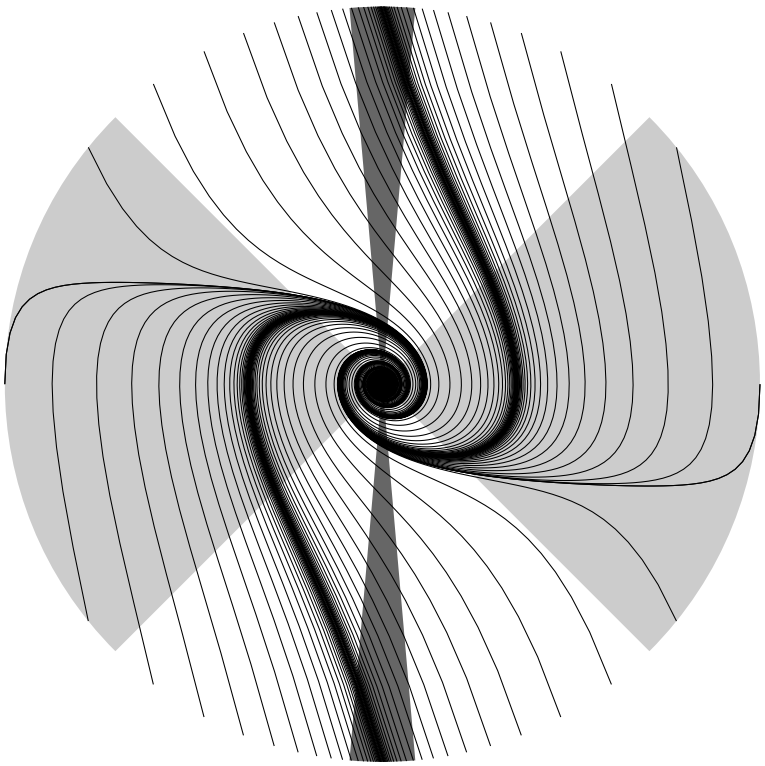
Kinetic initial conditions for inflation:
Theory, observations and methods

William James Handley
Cavendish Astrophysics Group
Gonville & Caius College

22nd April 2016



A dissertation submitted for the degree of Doctor of Philosophy at the University of Cambridge

Kinetic initial conditions for inflation:

Theory, observations and methods

William James Handley

Abstract

This thesis is concerned with the initial conditions for inflation, and the construction of methods to aid in the observational and theoretical analysis of the early universe. Chapter 1 outlines the context and content of the thesis. After this, the thesis is divided into two parts. The first chapters of each part (Chapters 2 and 7) are introductory, establishing basic theory and notation. The remainder of the thesis is entirely my own work, except where references explicitly state otherwise.

Part I contains theoretical and observational work in early-universe cosmology, and is divided into four chapters. Chapter 2 introduces the inflationary cosmological theory relevant to this thesis.

Chapter 3 is published in Handley et al. (2014), and proves the theoretical result that almost all classical universes begin at a finite time in the past in a generic kinetically dominated state. Classical kinetically dominated universes are examined in detail, and the possible observable consequences are postulated. Chapter 4 details further theoretical observations into the kinetically dominated universe that have arisen since the publication of Handley et al. (2014).

Chapter 5 was published as part of the Planck Collaboration et al. (2016b), and gives model-independent reconstructions of the primordial power spectrum of curvature perturbations. The results are consistent with the concordance ΛCDM cosmology, but show hints of possible effects of kinetic dominance.

Chapter 6 details theoretical work in the quantum mechanics of the early universe, and has been published in Handley et al. (2016a). A novel approach for defining the quantum vacuum is proposed via the renormalised stress-energy tensor of spacetime, and an application to the kinetically dominated universe is considered. The new theory makes potentially observable predictions.

Part II contains methods developed for the theoretical and observational analysis of the early universe. Chapter 7 provides an introduction to Bayesian methodologies and nested sampling.

Chapters 8 & 9 detail my contributions to the field of nested sampling, and have been published in Handley et al. (2015a,b). These demonstrate the effectiveness of the Bayesian PolyChord algorithm, a novel nested sampling methodology utilising slice sampling and semi-independent posterior mode analysis.

Chapter 10 demonstrates a new method for the efficient numerical solution of oscillatory ordinary differential equations, termed the Runge-Kutta-Wentzel-Kramers-Brillouin method, and has been submitted to the Journal of Computational Physics as Handley et al. (2016b).

At the end of each part, conclusions are given along with the direction of ongoing and potential further research.

Contents

Contents (detailed)

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the relevant Degree Committee.

Acknowledgements

Completing a PhD is no small task, and is certainly not performed in isolation. I am lucky to have been supported by a wide variety of people over the years, many of whom I will have inevitably forgotten. I apologise if you are unintentionally left out of these acknowledgements.

First, thanks go to all of my teachers over the past twenty-six years, but especially those at Alleyn's School: Ed Mann, Anne Poole, Neil Kinnear, Sara Hopley, Andy Skinnard & Margaret Hunnaball, and those at Cambridge: Rafi Blumenfeld, Derek Barnes, Adam Thorne, David Summers, Ulrich Kaiser, Steve Gull, Paul Alexander & John Ellis. Having now been a supervisor myself, I know how difficult a student I must have been, and with that in mind I am especially grateful.

Many thanks are due to the moderators and posters on the various technical fora throughout the internet, particularly on StackExchange. You are the unsung heroes and heroines of the modern internet age.

I am very grateful for the care of the Astrophysics administrator Karen Scrivener. Paul has often said that you are the most important member of the department, and I entirely agree.

To my office mates; Sonke Hee, Richard Wolstenhulme and Do Young Kim for the many amusing, interesting and fruitful conversations over the past few years.

To Professors Anthony Lasenby and Mike Hobson. One could not wish for more supportive and caring supervisors. I can only hope that one day I will be able to pass the favour on to the next generation of students with half as much skill and grace.

To my mother and father for their continuing love and support, be it financial, emotional or moral.

Last and most to my fiancée Sophie Lovick, for her patience and sacrifice, and above all for managing to live with a Physicist for five and a half years.

Conventions

- I make full use of symbolic overloading; using the same symbol in different contexts where their mathematical meaning is distinct, but their physical meaning is related. For example:
  - $f = f(x)$, where f on the right hand side refers to a function $f : X \to Y$, but on the left hand side $f \in Y$ refers to the image of $x \in X$ under the function f.
  - $\rho \to \rho + \delta\rho$, where on the right hand side ρ refers to the unperturbed solution, whilst on the left ρ refers to the perturbed solution.
  - $v_i \rightsquigarrow \partial_i v + v_i$, where on the left v_i refers to a generic vector field, but on the right v refers to the helicity scalar part of the field and v_i refers to the helicity vector part.
- I work using a metric with a positive signature $(+, -, -, -)$.
- Fourier transforms are defined so that Fourier synthesis carries the factors of 2π:

$$f(\mathbf{k}) = \int_{-\infty}^{\infty} \mathrm{d}^3\mathbf{x}\, e^{-i\mathbf{k}\cdot\mathbf{x}} f(\mathbf{x}) \qquad f(\mathbf{x}) = \int_{-\infty}^{\infty} \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3}\, e^{i\mathbf{x}\cdot\mathbf{k}} f(\mathbf{k}). \tag{1}$$

- Tensorial spacetime indices are denoted with Greek letters $\mu, \nu, \ldots$. The 0 index indicates the time component, and Latin letters $i, j, \ldots$ indicate spatial indices.
- Quantum mechanical operators are written without hats, unless their presence increases clarity.
- I work in natural units so that:

$$c = \hbar = G = k_B = 1, \tag{2}$$

  but retain the reduced Planck mass for clarity:

$$m_{\mathrm{p}}^2 = \frac{1}{8\pi G} = \frac{1}{8\pi}. \tag{3}$$

Chapter 1

Outline

1.1 The big picture

As cosmologists, nature has been incredibly kind to us. We have been given a near crystal clear snapshot of the universe a mere $380,000$ years after its birth. Maps created by the Planck satellite (Figure 1.1) allow us to determine the patterns of density in the early universe. For cosmologists these density distortions are interesting in two ways.

First, these perturbations in density are the beginnings of the formation of stars, galaxies and galaxy clusters. If one were to wind the clock forwards from this moment, cosmic structure would be seen coalescing around the regions of higher density.

Second, these distortions tell us a great deal about physics at much earlier times. Observations from particle physics experiments allow us to confidently wind the clock backwards to mere microseconds after the big bang. However, the expansion of the universe itself allows us to look even further back than this. We now have a wealth of observational evidence that early in its history, the universe underwent a rapid accelerated expansion. This expansion acts as a cosmic magnifying glass, allowing us to observe patterns $\sim 10^{-32}$ seconds after the big bang using the universe we see today. The upshot of this is that cosmologists effectively have access to the most powerful particle accelerator imaginable, reaching energies trillions of times greater than the Large Hadron Collider.

The canonical explanation for the early period of accelerated expansion is the theory of inflation, with quantum fields providing the necessary driving force. This thesis focusses on the initial conditions for inflation; i.e. what started this all off.

1.2 Kinetic initial conditions

Traditionally, cosmologists work under the assumption that at these early times the universe was in an effectively eternal inflating state, with no detectable beginning. Chapter 3 rigorously proves a result that suggests this picture may be somewhat incomplete. In fact, almost all classical universes begin at a finite time in the past. Moreover, this beginning is dominated by kinetic energy, and not inflating. This provides a novel and arguably simpler mechanism for

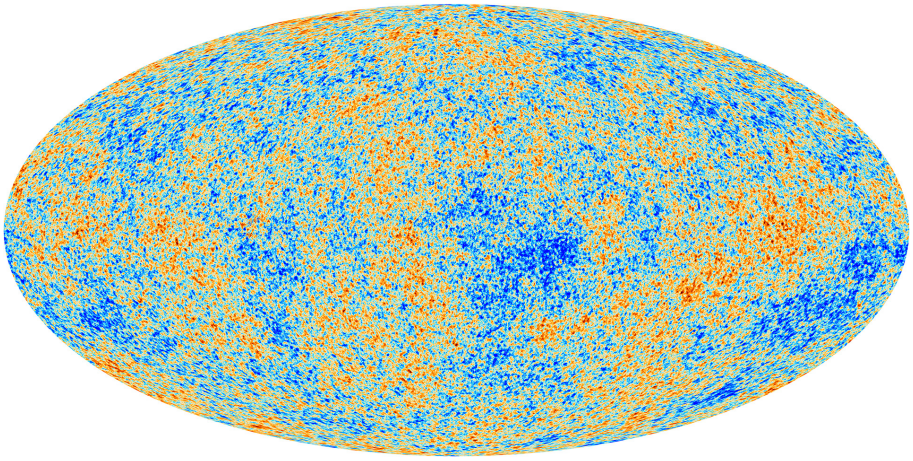


Figure 1.1: Temperature distortions in the cosmic microwave background.

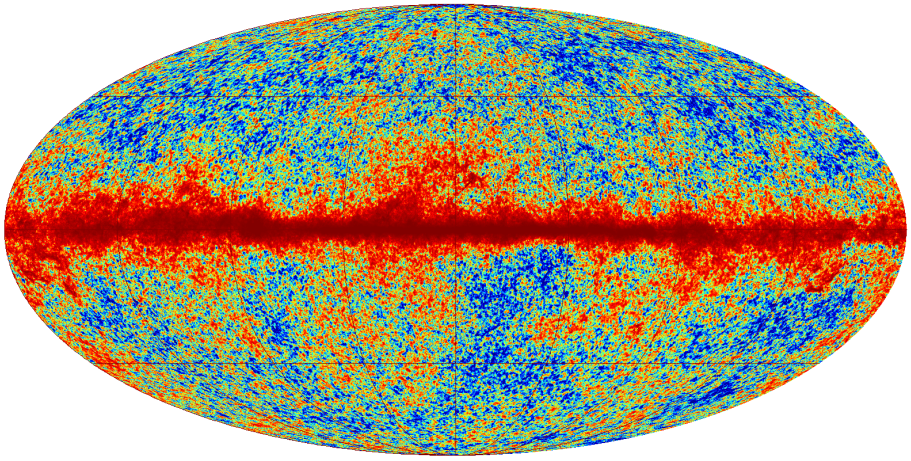


Figure 1.2: The microwave sky as seen by Planck at 143GHz.

setting the initial conditions of the universe. More importantly, I also show that this period could have produced a distinct observational signature in the primordial power spectrum of curvature perturbations. Chapter 4 details how this mathematical observation fits into more traditional approaches.

As a member of the Planck collaboration, I began to search for evidence of this pre-inflationary phase. Chapter 5 details a model-independent reconstruction of the primordial spectrum. Whilst not conclusive, there are tantalising hints of a signal consistent with a pre-inflationary epoch.

1.3 Observations in high dimensions

Our microwave sky does not look like Figure 1.1. The sky that Planck actually sees is more akin to Figure 1.2. The most notable difference between the two figures is the presence of a red band in the centre of the second image, which

are the microwaves emitted by our own Milky Way galaxy. In order to observe the signal generated by the beginning of the universe (Figure 1.1), we must first take into account the contaminating information of the Milky Way. This requires a sophisticated model of the galaxy, with many parameters that must be simultaneously determined and quantified.

Whilst attempting to reconstruct the primordial power spectrum (Chapter 5), it became apparent that there was an absence of Bayesian data analysis tools. The techniques available at the time were incapable of navigating and integrating the complicated high-dimensional likelihoods required for my reconstruction.

The Cavendish Astrophysics group has a long history of developing and applying novel Bayesian statistical approaches. With this in mind, I designed and implemented a novel nested sampling algorithm which was christened PolyChord (detailed in Chapter 9). This proved capable of scaling to the dimensionalities required, and reliably computing the Bayesian evidence, allowing me to produce model-independent reconstructions of the primordial power spectrum. As a result, PolyChord was rapidly adopted by many members of the team as their de-facto inference tool.

1.4 Quantum initial conditions

My latest work focusses on the quantum mechanical initial conditions of the early universe. A full theoretical treatment of this epoch requires a consideration of quantum fields in curved spacetime. One of the critical issues of this field is that the basic concepts we use to describe quantum particles are not designed to work in the context of gravity as a curved spacetime background. My latest research aims to resolve some of these issues, by building the quantum vacuum around the renormalised stress energy tensor. In essence, I hypothesise that empty space could be better defined as "lowest energy" rather than "particle-less". I also demonstrated that in the context of the early universe this alternative viewpoint makes detectable predictions which again differ from standard theory. This is detailed in Chapter 6.

Whilst examining this, I realised that I needed a better way of solving the differential equations of the early universe. I succeeded in developing a novel class of extremely efficient numerical methods for solving these, which I term Runge-Kutta-Wentzel-Kramers-Brillouin approaches. RKWKB is explained Chapter 10.

1.5 Thesis versus research

The theme of this thesis is the interplay between theory, observations and methods. Whilst the thesis divided into two parts, in reality both halves have strongly influenced each other in a manner not necessarily consistent with the sequence of the text. Figure 1.3 shows an approximate set of interactions between the various chapters. Whilst a thesis must be laid out sequentially, actual research is often very non-linear. The degree committee requires that a dissertation be a "connected account of research", and whilst this thesis is not sequentially connected, it does at least form a directed acyclic graph.

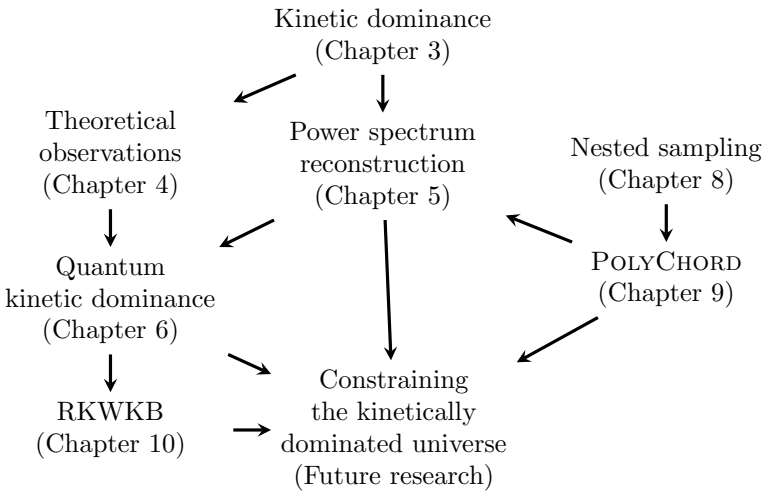


Figure 1.3: The "sequence" of my research.

Part I

Cosmology

Chapter 2

Inflationary Cosmology

2.1 Introduction

This chapter reviews the key concepts of inflationary cosmology, and mostly serves to establish notation. For further detail, excellent references can be found in Wald (1984), Hobson et al. (2006) & Dodelson (2008).

2.2 Einstein's gravity

> *Spacetime tells matter how to move;*
> *matter tells spacetime how to curve.* — John Wheeler

Einstein's theory of general relativity accounts for gravity by removing it as a fundamental force and considering gravitation as a property of spacetime itself. Objects and fields still interact with one another on a spacetime background via the usual forces (electromagnetic, strong and weak nuclear forces). The background spacetime can be thought of as curved, and the perceived effect of gravitation is due to objects moving on straight paths in a curved spacetime. Finally, the curvature (and thus gravitation) is generated by the matter content of the spacetime.

The formalism of Einstein's gravity can be effectively summarised using the Einstein-Hilbert action. An action S is written as a general relativistic integral over a Lagrangian density $\mathcal{L}$:

$$S = \int \mathrm{d}^4x \sqrt{|g|}\mathcal{L}, \tag{2.1}$$

where the factor of $\sqrt{g}$, $g = |\det(g_{\mu\nu})|$ ensures a relativistic volume element for integration. We typically decompose the Lagrangian $\mathcal{L}$ into a gravitational and matter part:

$$\mathcal{L} = \mathcal{L}_G + \mathcal{L}_M, \tag{2.2}$$

$$\mathcal{L}_G = \frac{1}{2} m_\mathrm{p}^2 R, \tag{2.3}$$

where R is the Ricci scalar and $\mathcal{L}_M$ is the portion of the Lagrangian pertaining to the material content of spacetime. Requiring that the action (2.1) is extremal

($\delta S = 0$) yields Einstein's equations:

$$G_{\mu\nu} = \frac{1}{m_{\rm p}^2} T_{\mu\nu}, \tag{2.4}$$

where:

$$T_{\mu\nu} = \frac{-2}{\sqrt{|g|}} \frac{\delta}{\delta g^{\mu\nu}} \left(\sqrt{|g|} \mathcal{L}_M \right), \tag{2.5}$$

is the stress energy tensor, and:

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R, \tag{2.6}$$

is the Einstein tensor. The symmetries of the Einstein tensor (2.6) mean that there are in fact only six independent equations in (2.4). Further, the fact that $\nabla^\mu G_{\mu\nu} = 0$ means that the stress energy tensor is conserved:

$$\nabla_\mu T^\mu{}_\nu = 0. \tag{2.7}$$

This conservation equation can provide a fast means for deriving alternative rearrangements of the Einstein equations.[1]

2.3 The smooth, expanding universe

On the largest scales, we observe the universe to be spatially *homogeneous* and *isotropic*. This justifies the philosophical *cosmological principle*, that the universe should look the same wherever you are, and that no place in the universe is special.

In this section, we examine the consequences that these observations have within the context of general relativity, by considering the solutions to the Einstein equations (2.4).

2.3.1 Metric

In any general relativistic analysis, it is helpful to restrict the form of the metric via the symmetries of the problem. Under the assumption of the cosmological principle, the metric may always be written in the Friedmann-Robertson-Walker form:

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a(t)^2 \, \mathrm{d}X^2, \tag{2.8}$$

where:

$$\mathrm{d}X^2 = \mathrm{d}\chi^2 + S_k^2(\chi) \, \mathrm{d}\Omega, \tag{2.9}$$

$$\mathrm{d}\Omega = \mathrm{d}\theta^2 + \sin^2\theta \, \mathrm{d}\phi^2, \tag{2.10}$$

$$S_k^2(\chi) = \begin{cases} \frac{1}{k} \sin^2\left(\chi\sqrt{k}\right) & : k > 0 \\ \chi^2 & : k = 0 \\ \frac{1}{|k|} \sinh^2\left(\chi\sqrt{|k|}\right) & : k < 0. \end{cases} \tag{2.11}$$

[1] Historical note: Einstein in fact derived this argument in reverse. He defined the Einstein tensor (2.6) by adjusting the Ricci tensor so that the stress energy tensor is conserved by construction. In the more modern approach presented here, the definition of the Einstein tensor arises naturally from variational principles.

Symbol	Definition
t	cosmic time
X	spatial coordinate
χ	comoving radial coordinate
θ	polar angle
ϕ	azimuthal angle
Ω	solid angle
a	cosmic scale factor
$k > 0$	closed universe
$k = 0$	flat universe
$k < 0$	open universe

Table 2.1: Definitions of terms in the FRW metric.

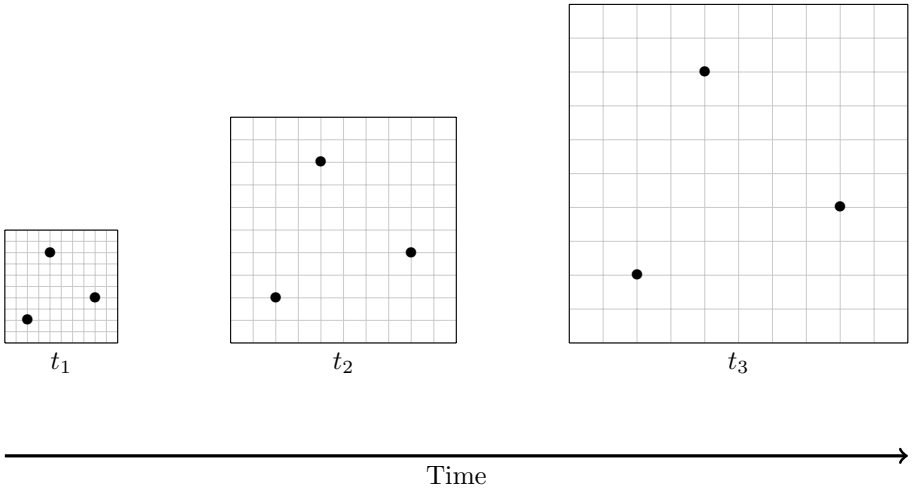


Figure 2.1: The expansion of the universe. As the universe evolves with time, the scale factor $a(t)$ changes. The scale factor $a(t)$ connects *comoving coordinates*, X with *physical coordinates*, $x = a(t)X$. Comoving variables can be thought of as a time-independent grid, which expands with the universe, physical variables are what observers would measure as distances. Hence, in an expanding universe, the physical distance between observers (dots in the diagram above) appears to increase over time, whilst their comoving distance remains the same.

The definitions of these terms can be found in Table 2.1.

The form of the metric (2.8) is close to Minkowski. Spatial slices at constant t are spaces with constant curvature, which may be positive, negative or zero (corresponding to a closed, open or flat universe). The time dependency of the spatial part is a scaling by a scale factor $a(t)$. As cosmic time t increases, $a(t)$ evolves, causing the spatial slice to expand or contract (Figure 2.1).

2.3.2 Dynamics

In order to obtain the equations governing $a(t)$ and thus the dynamics of the universe, we must make some assumptions about the universe's contents. For a smooth universe, one may model its contents as a collection of non-interacting, comoving, uniform, perfect fluids. A perfect fluid in thermodynamic equilibrium has stress-energy tensor:

$$T^{\mu\nu} = (P + \rho)u^\mu u^\nu - Pg^{\mu\nu} + \Sigma^{\mu\nu}, \tag{2.12}$$

where ρ is the energy density, P is the pressure, u^μ is the four velocity of the fluid, and $\Sigma^{\mu\nu}$ is a traceless, symmetric, anisotropic stress term. In accordance with the cosmological principle, we shall assume that in the comoving frame the fluid is stationary $u^\mu = [1, \mathbf{0}]$, and uniform $\rho = \rho(t), P = P(t)$, with no anisotropy $\Sigma = 0$.

Applying the metric (2.8) to the Einstein equations (2.4), with the stress-energy tensor (2.12) one finds:

$$\dot{H} + H^2 = -\frac{1}{6m_\mathrm{p}^2}\left(\rho + 3P\right), \tag{2.13}$$

$$H^2 = \frac{1}{3m_\mathrm{p}^2}\rho - \frac{k}{a^2}, \tag{2.14}$$

where $H = \dot{a}/a$ is the Hubble parameter and a dot denotes differentiation with respect to cosmic time, $\dot{f} \equiv \mathrm{d}f/\,\mathrm{d}t$. These are termed the *acceleration* and *Friedmann* equations respectively, and implicitly govern the dynamics of the scale factor $a(t)$. It should be noted that these equations are not complete, as additionally one requires an equation of state linking ρ and P.

2.3.3 Basic solutions

A reasonable model for the universe we observe today is to treat the matter as a multi-component fluid, with each component with its own equation of state:

$$\rho = \sum_i \rho_i, \qquad P = \sum_i P_i, \qquad P_i = w_i\rho_i, \tag{2.15}$$

where w_i is the equation of state parameter. In our universe, we observe matter ($w = 0$), radiation ($w = \frac{1}{3}$) and dark energy ($w \approx -1$). We can also notationally model the curvature's scale-factor contribution of $-\frac{k}{a^2}$ as a cosmological fluid with $w = -\frac{1}{3}$. Applying these equations of states, equations (2.13) and (2.14) may be re-cast as an equation purely in a:

$$\left(\frac{H}{H_0}\right)^2 \equiv \frac{1}{H_0^2}\left(\frac{\dot{a}}{a}\right)^2 = \Omega_0^{(\mathrm{rad})}a^{-4} + \Omega_0^{(\mathrm{mat})}a^{-3} + \Omega_0^{(\mathrm{curv})}a^{-2} + \Omega_0^{(\mathrm{de})}, \tag{2.16}$$

where the present-day ($t = t_0$) scale factor is chosen to be unity ($a(t_0) = 1$), quantities subscripted with 0 indicate the present-day value of the parameter, and the density parameter Ω is defined as:

$$\Omega = \frac{\rho}{\rho_\mathrm{c}}, \qquad \rho_\mathrm{c} = 3m_\mathrm{p}^2H^2, \tag{2.17}$$

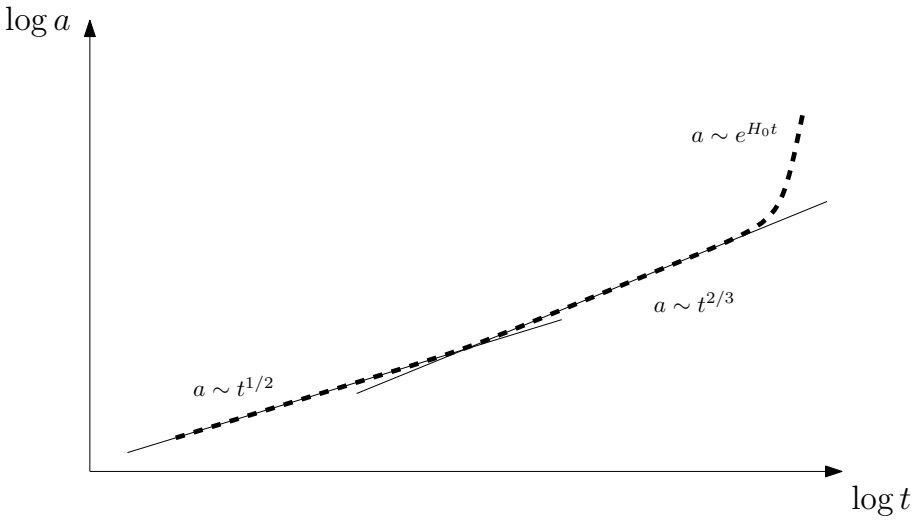


Figure 2.2: The approximate expansion history of our universe. Initially, the universe is dominated by radiation $a \sim t^{1/2}$. As the universe expands, the photons lose energy as their wavelengths are stretched. Thus, the universe transitions to a matter dominated $a \sim t^{2/3}$ regime. Finally, the dark energy of the universe overtakes the matter content, causing an exponential expansion $a \sim e^{H_0 t}$.

which measures the fraction of the critical density ρ_c taken up by a given component. Equation (2.16) cannot be analytically solved in general, but if one assumes that one component is dominant, and the rest negligible, then one recovers the solution:

$$a \propto \begin{cases} t^{2/3(w+1)} & : w \neq -1 \\ e^{H_0 t} & : w = -1. \end{cases} \tag{2.18}$$

We can see immediately that in a universe dominated by dark energy ($w \approx -1$) the expansion is exponential, and in one dominated by radiation ($w = 1/3$), that $a \propto t^{1/2}$, which is a slower expansion than one dominated by matter $a \propto t^{2/3}$. For our universe, we observe that it is approximately flat $\Omega_0^{(\mathrm{curv})} \approx 0$, and that matter today is of the same order of magnitude as dark energy but vastly outweighs radiation $\Omega_0^{(\mathrm{de})} \approx \Omega_0^{(\mathrm{mat})} \gg \Omega_0^{(\mathrm{rad})}$. We thus expect an expansion history of the form shown in Figure 2.2.

2.4 Conformal time and redshift

2.4.1 Redshift z

The expansion of the universe causes the wavelengths of photons to increase: Photons have a four-momentum proportional to their wavevector $p^\mu \propto k^\mu$. Since the FRW metric has no explicit χ-dependence for radially travelling photons, p_χ is conserved: $p_\chi(t_1) = p_\chi(t_2)$. Using the metric to raise the indices, one finds that $p^\chi(t_2)a(t_2)^2 = p^\chi(t_1)a(t_1)^2$. Identifying the physical momentum

Epoch	Redshift
Matter-radiation equality	$z \sim 3400$
Recombination	$z \sim 1089$
Dark ages	$20 < z < 1089$
First stars	$z \sim 20$
Reionisation	$6 < z < 20$
Dark energy-matter equality	$z = 0.4$
Now	$z = 0$

Table 2.2: Recent history of the universe. As redshift is a directly observable quantity, it provides a natural measure of cosmic epoch.

$ap^\chi \propto k^r \propto \lambda^{-1}$ where λ is the wavelength of the photon, one finds:

$$\frac{\lambda_2}{\lambda_1} = \frac{a_2}{a_1}, \tag{2.19}$$

and thus the wavelengths of photons increase with the expansion of the universe. Physically, the stretching of spacetime stretches the wavelengths of photons. The redshift z of the photon from some early time t_1, relative to the current epoch t_0, is defined as usual as:

$$z = \frac{\lambda_0 - \lambda_1}{\lambda_1}, \tag{2.20}$$

which gives a relation between the redshift of a photon and the scale factor:

$$a = \frac{1}{1+z}. \tag{2.21}$$

Since redshift is a physically observable quantity, it provides a cosmology-independent measure of the epoch of the universe (Table 2.2).

2.4.2 Conformal time η

It is convenient to define conformal time as:

$$\eta = \int \frac{\mathrm{d}t}{a}, \tag{2.22}$$

so that the line element becomes:

$$\mathrm{d}s^2 = a(\eta)^2 \left(\mathrm{d}\eta^2 - \mathrm{d}X^2\right). \tag{2.23}$$

This is often analytically useful, as it demonstrates the metric is conformally equivalent to Minkowski space[2]. Physically conformal time corresponds to a temporal coordinate in which photons appear as if they were in flat space: If we consider (without loss of generality) radially travelling photons $\mathrm{d}\Omega = 0$ the line element is:

$$\mathrm{d}s^2 = a(\eta) \left(\mathrm{d}\eta^2 - \mathrm{d}\chi^2\right). \tag{2.24}$$

[2]Hence the name "conformal time".

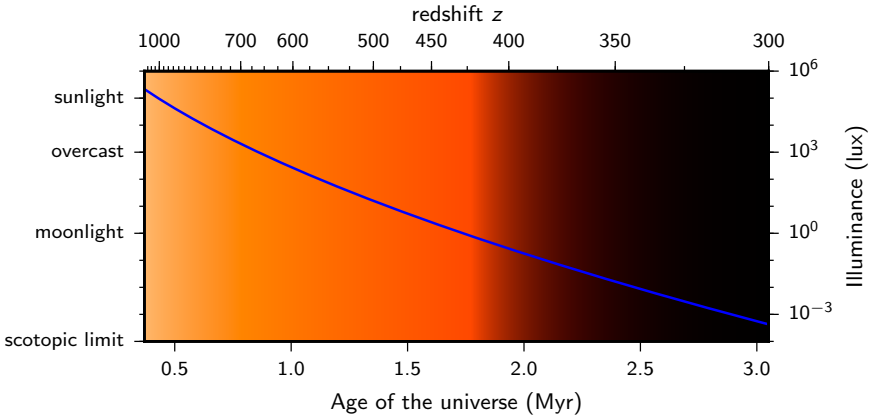


Figure 2.3: The colour and illuminance of the radiation background, as a function of cosmic history. As the universe ages, the initially optical background gradually redshifts all the way down to the microwaves that are observed today. This plot shows the colour and the perceived irradiance (illuminance) that the CMB would appear to be in the first few million years after last scattering. For example, for an observer (un)fortunate enough to be born $500,000$ years after the big bang, the universe would still be bathed in orange light from all directions.

We have thus removed all the complexities of curvature and comoving coordinates. Since photons have a null trajectory, $ds^2 = 0 \Rightarrow \mathrm{d}\eta = \pm\,\mathrm{d}\chi$, and therefore travel in straight lines on spacetime diagrams with η and χ as axes. This considerably simplifies most pictures.[3] It can therefore be thought of as a "comoving" time, in analogy with comoving spatial coordinates.[4]

2.5 The cosmic microwave background

As we turn telescopes on objects further away from earth, we begin to look appreciably back in time. The radiation from these objects has taken so long to reach us that we can observe the universe in a much younger state than it is now. The furthest galaxies imaged by the Hubble space telescope are more than 13 billion years old. Many of these objects are so far away that they have redshifted out of the visible spectrum and into the infra-red. If we look beyond, we enter the dark ages of the universe, before the first stars had turned on. This would appear to be the end of the observational story.

However, from behind the dark ages, there is a background of microwaves. These would originally have been emitted as optical light, but have redshifted all the way down to the microwave end of the spectrum (Figure 2.3). This uniform backdrop of radiation is our direct image of the universe as it was at redshift $z \sim 1000$, when the universe was a mere $380,000$ years old.

[3] For example, later in the chapter Figures 2.5 & 2.6 both use this construction.

[4] Alternatively, conformal time is the time measured by a small light clock that expands comovingly with the universe.

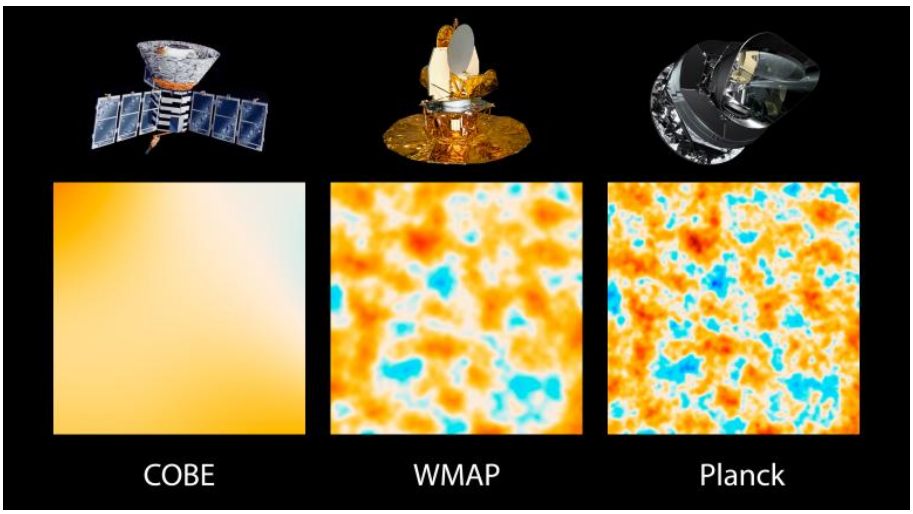


Figure 2.4: Microwave satellite images of the minute CMB anisotropies. COBE ran from 1989–1993, WMAP from 2001–2010 and Planck from 2009–2013.

Since its first (accidental) detection in 1964 by Penzias and Wilson (1965), a succession of microwave telescopes, both on the ground and in space, have sharpened our image of this (Figure 2.4). The cosmic microwave background (CMB) is found to be:

1. A near-perfect blackbody spectrum, with $T_0 = 2.7254(5)$K.
2. Isotropic to one part in 10^5. The minute anisotropies have non-trivial power spectrum, and contain a wealth of cosmological information.

2.6 Problems in the CMB

2.6.1 Flatness problem

The degree to which the universe is not flat can be measured by re-writing the Friedmann equation (2.14) in terms of Ω as in (2.17):

$$1 - \Omega = -\frac{k}{(aH)^2} \equiv \Omega_k. \tag{2.25}$$

In terms of these kind of variables, the acceleration equation (2.13) is:

$$\frac{\mathrm{dlog}\, aH}{\mathrm{dlog}\, a} = -\frac{1}{2}\Omega(1 + 3w), \tag{2.26}$$

where here $w = P/\rho$ is not necessarily constant. Taking absolute logarithms of (2.25), differentiating and applying (2.26) yields:

$$\frac{\mathrm{dlog}\,|\Omega - 1|}{\mathrm{dlog}\, a} = \Omega(1 + 3w). \tag{2.27}$$

One can see from this that a flat universe ($\Omega = 1$) is an unstable point of equilibrium of these equations, provided that $1 + 3w > 0$. For both matter ($w = 0$) and radiation ($w = \frac{1}{3}$), the universe is rapidly driven away from flatness as the universe expands. This is natural, as the effect of spatial curvature on the energy and dynamics of the universe has a slower decay rate in comparison with ordinary matter (equation 2.16).

This presents a problem. The universe we see today is measured to be flat to at least one part in 10^{-2}, which means that at earlier times it would have been even flatter still. Given that within the context of cosmology there is no a-priori reason to believe the universe should be exactly flat, this requires a unreasonable quantity of fine tuning. It would be more satisfactory if we had a dynamical reason to explain why the universe is as flat as it is.

As well as revealing the problem, equation (2.27) also provides a solution. If for some period $w < -1/3$ at some earlier time, then (2.27) says that $\Omega = 1$ is an attractor state. The acceleration equation shows that:

$$\frac{\ddot{a}}{a} = -(1 + 3w)\rho. \tag{2.28}$$

Thus, the condition that $w < -1/3$ is equivalent to requiring that at some point early in its history the universe was *accelerating* $\ddot{a} > 0$. It is also equivalent to having a form of matter which satisfies:

$$P < -\frac{1}{3}\rho, \tag{2.29}$$

which can be recognised as a violation of the *strong energy condition*. For more detail, see Visser and Barceló (2000).

2.6.2 Horizon problem

The near-perfect isotropy of the CMB also presents a problem. Our image of the CMB represents the universe as it was some 300,000 years after it was born. In cosmologies with traditional forms of matter, there is quantitatively not enough time before the emission of the CMB for *any* dynamical process to allow the universe to reach a homogeneous state. Indeed, the CMB can be seen to be made up of some 10^5 causally disconnected patches, with the causal patch size approximately one thumbs-width on the sky. This can be seen graphically in Figure 2.5.

Note that we may resolve this problem with an accelerated phase as well (Figure 2.6). An accelerated phase pushes any singularity far back into the conformal past. This means that there is more than enough time for the universe to come dynamically to equilibrium. In fact, a detailed analysis shows that this accelerated epoch acts as a "smoothing" effect on a given causal patch, yielding a homogeneous universe from any generic initial conditions. We should therefore expect universes with accelerated early epochs to have generically smooth cosmic microwave backgrounds.

2.7 Inflation

The canonical way to explain this early accelerated phase is via the phenomenon of *inflation*. This early accelerated phase will have occurred when the universe

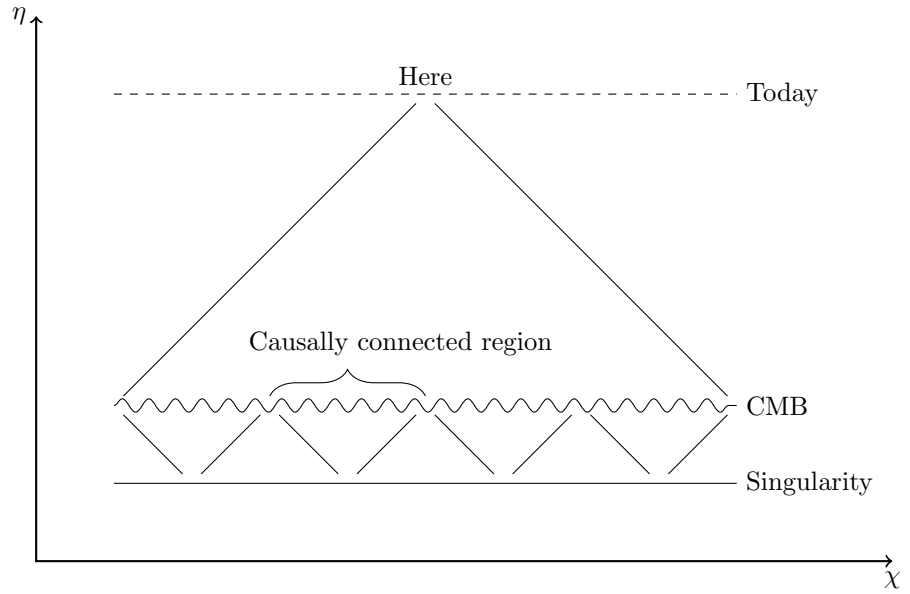


Figure 2.5: The horizon problem. From our position, we observe the CMB as a surface at some earlier time. The homogeneity of the CMB suggests that there must have been some physical mechanism to smooth it out. However, in cosmologies with traditional matter, there is not enough time before the emission of the CMB to allow this to occur. In fact, the CMB appears to be made up of causally disconnected regions. There is thus no dynamical reason as to why these disconnected regions should show such similar physical conditions. As a rough guideline, a causal patch is roughly a "thumbs width at arms length" on the sky.

was at extremely high energy, which suggests that we should turn to particle physics phenomenology. It turns out if we consider even the most simple particle physics models in the context of general relativity, these are capable of generating a sustained accelerated phase.

2.7.1 Basic theory

Consider the Lagrangian:

$$\mathcal{L}_\phi = \frac{1}{2}\nabla^\mu\phi\nabla_\mu\phi - V(\phi). \tag{2.30}$$

This represents a scalar field ϕ, minimally coupled to gravity with some unspecified potential V. Inserting this into (2.5) yields a stress energy tensor of:

$$T^\mu_\nu = \nabla^\mu\phi\nabla_\nu\phi - \left(\frac{1}{2}\nabla^\alpha\phi\nabla_\alpha\phi - V(\phi)\right)\delta^\mu_\nu. \tag{2.31}$$

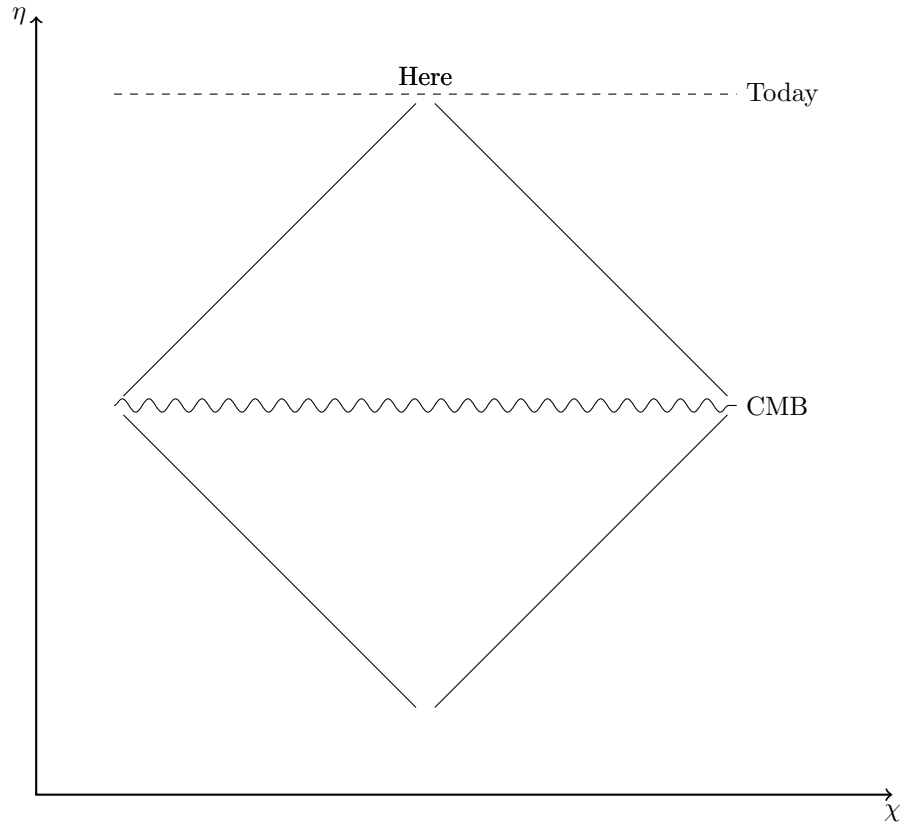


Figure 2.6: Horizon problem resolved. An early accelerating phase pushes the singularity far into the conformal past, meaning there is more than enough time for the CMB to come to equilibrium.

If we initially assume in accordance with the cosmological principle that the field has no spatial dependence $\phi = \phi(t)$, then the stress energy tensor becomes:

$$T_0^0 = \frac{1}{2}\dot{\phi}^2 + V(\phi) = \rho, \tag{2.32}$$

$$T_j^i = -\left[\frac{1}{2}\dot{\phi}^2 - V(\phi)\right]\delta_j^i = -P\delta_j^i. \tag{2.33}$$

Thus, a homogeneous scalar field acts as a perfect fluid with pressure and density as shown above. In order to derive the non-trivial and time dependent equation of state, we need to generate an equation of motion for ϕ. This can be done either by extremising $S_\phi = \int \mathrm{d}^4x\sqrt{|g|}\mathcal{L}_\phi$, or by applying the continuity equation (2.7) to the stress energy tensor (2.31):

$$0 = \ddot{\phi} + 3H\dot{\phi} + \frac{\mathrm{d}}{\mathrm{d}\phi}V(\phi). \tag{2.34}$$

The middle term involving H is the term that arises as a result of including the effects of gravity (i.e. an expanding universe). The homogeneous value of

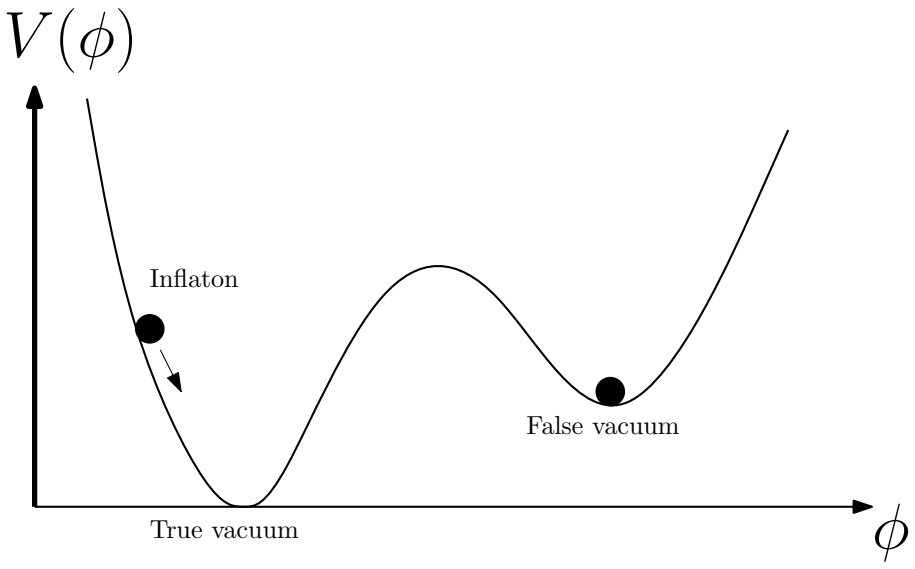


Figure 2.7: An example inflationary potential.

the scalar field ϕ satisfies the equation of a particle in a potential $V(\phi)$ with a frictional term proportional to H. To get an equation for H, we use the acceleration equation (2.13) along with our formulae (2.32) & (2.33) for ρ and P, to find:

$$\dot{H} + H^2 = -\frac{1}{3m_{\mathrm{p}}^2}\left(\dot{\phi}^2 - V(\phi)\right). \tag{2.35}$$

Equations (2.34) & (2.35) are tightly coupled, and result in far less trivial behaviour compared with a simple perfect fluid.

2.7.2 Phenomenology

One way to trigger an accelerated phase is to set the particle of the field as trapped in a false vacuum (Figure 2.7). This sets $\dot{\phi} = 0$, $V(\phi) = V_0$ and therefore $H = H_0 = V_0/3m_{\mathrm{p}}^2$ and $a \sim e^{H_0 t}$, which is an exponential and therefore accelerated expansion. However, to end this accelerated phase, the particle would have to tunnel quantum mechanically out of its false vacuum. It turns out that the false vacuum is too stable, and results in unphysical predictions.

In fact, one does not need a false vacuum, one merely needs the inflaton to be rolling slowly down the potential with its speed small in comparison to the potential energy:

$$\dot{\phi}^2 \ll V(\phi). \tag{2.36}$$

In this case, one still has the Hubble parameter $H \approx H_0$ being approximately constant and therefore exponential expansion. Since the frictional term in (2.34) robs the particle of energy, one finds that in fact these slowly rolling phases are generic attractors for most inflationary potentials. All one requires is that the inflaton begins a reasonable way from the potential minimum.

Thus, a very general scalar particle is capable of triggering a generic exponential accelerated expansion. This epoch of the universe is termed inflation, and the particle of the field ϕ the inflaton.

2.7.3 Reheating

Inflation finishes when the inflaton reaches its true vacuum, and oscillates about its minimum. One then imagines that the inflaton then decays into more traditional matter in a period called reheating. This is theoretically interesting and an area of active research. As we shall see however, for the purposes of setting initial conditions on the remainder of the contents of the universe, reheating can often be ignored.

2.8 The perturbed universe

Observationally, the real universe is not perfectly smooth.[5] We may treat the smooth FRW metric as a 0^{th} order approximation, to the universe, and expand about this solution using perturbation theory. In general then, we write each quantity as:

$$X(t, \mathbf{x}) \to X(t) + \delta X(t, \mathbf{x}), \tag{2.37}$$

where "$\to$" in this context should be read as "is perturbed as". Since in the early universe the perturbations were small $\delta X \ll X$, we may expand all equations to linear order with very high accuracy. One cannot hope to cover the full subtlety of cosmological perturbation theory in a short introductory chapter. For more detail, I highly recommend the expositions by Mukhanov et al. (1992) and Baumann (2009). We shall work with the perturbations to the flat FRW metric, but the analysis can be extended to the open and closed cases as well.

2.8.1 Metric perturbation

A general perturbation of the flat FRW metric will take the form:

$$\mathrm{d}s^2 = (1 + 2\Phi)\,\mathrm{d}t^2 - 2aB_i\,\mathrm{d}x^i\,\mathrm{d}t - a^2\left[(1 - 2\Psi)\,\delta_{ij} + 2E_{ij}\right]\mathrm{d}x^i\,\mathrm{d}x^j, \tag{2.38}$$

where the various terms in the above expression are defined in Table 2.3, and $\Phi, B_i, \Psi, E_{ij} \ll 1$ are small, and the tensor E_{ij} is symmetric and traceless.

In order to arrive at (2.38), begin by considering a perturbation to the FRW metric tensor (2.8): $g_{\mu\nu} \to g_{\mu\nu} + \delta g_{\mu\nu}$. Since the background form is split into spatial and temporal parts, one may split the perturbed metric into $\mathrm{d}t^2$, $\mathrm{d}t\,\mathrm{d}x^i$ and $\mathrm{d}x^i\,\mathrm{d}x^j$ terms. We may choose to define $\delta g_{00} = 2\Phi$, $\delta g_{0i} = \delta g_{i0} = -aB_i$, $\delta g_{ij} = -2a^2F_{ij}$. The factors of two and a are introduced for later notational convenience, and $F_{ij} = F_{ji}$ is symmetric. Finally, we split $F_{ij} = -\Psi\delta_{ij} + E_{ij}$ into a symmetric, traceless tensor E_{ij} and a scalar Ψ. Putting this all together, one arrives at (2.38) above.

[5] For example, the fact that you are reading this thesis indicates that there must be some departure from homogeneity.

Symbol	Definition
Φ	lapse
B_i	shift
Ψ	(spatial) curvature perturbation
E_{ij}	(spatial) shear (3-tensor)

Table 2.3: Definitions of terms in the perturbed FRW metric. Although E_{ij} is termed the shear tensor, it is in fact its time derivative $\dot{E}_{ij}$ which determines the shear of the worldlines of constant coordinate position.

2.8.2 Matter perturbation

Perturbations of the matter content will take the form:

$$\rho \to \rho + \delta\rho, \qquad P \to P + \delta P, \qquad u_\mu \to [1 + \Phi, -\delta q_i/(\rho + P)]\,. \tag{2.39}$$

The density and pressure perturbations are self-explanatory. The perturbation to the fluid velocity is first chosen so that $u^\mu u_\mu = 1$ to first order, and δq_i is the perturbation to the momentum density. The advantage of this is that for non-interacting fluids, $\delta\rho$, δP and δq_i are all additive, as in equation (2.15). We have neglected a perturbation to the anisotropic stress. Anisotropic stress may be included if desired (for example in the case of neutrinos) but will add complexity unnecessary for this exposition.

At some point one must also assume an equation of state for the fluid which provides equations to constrain P and ρ. One can remain general by splitting the pressure perturbation into adiabatic and entropic parts:

$$\delta P = \delta P_{\rm ad} + \delta P_{\rm en}, \tag{2.40}$$

where:

$$\frac{\delta P_{\rm ad}}{\dot{P}} = \frac{\delta\rho}{\dot{\rho}}. \tag{2.41}$$

For more detail, see Baumann (2009, p. 48). In most cases, we limit ourselves to adiabatic matter $\delta P_{\rm en} = 0$. This is not particularly restrictive, and is naturally satisfied by the equations of state (2.15), although it is not in general satisfied by the pressure perturbations generated by scalar fields.

The scalar field perturbation is similarly simple:

$$\phi \to \phi + \delta\phi. \tag{2.42}$$

2.8.3 Fourier modes and scalar-vector-tensor decomposition

In general, the Einstein equations (2.6) are non-linear, second-order, partial differential equations and are therefore extremely challenging to solve. Perturbing about a background as in the previous section yields linearised versions of the Einstein (2.4) and conservation equations (2.7):

$$\delta G_{\mu\nu} = \frac{1}{m_{\rm p}^2}\delta T_{\mu\nu}, \tag{2.43}$$

$$\delta(\nabla_\mu T^\mu_\nu) = 0, \tag{2.44}$$

which greatly simplifies the analysis. We may go further and exploit the symmetries of the background unperturbed universe to make life even easier.

First, a general field $\varphi(t, \mathbf{x})$ can be written in Fourier modes as:

$$\varphi(t, \mathbf{k}) = \int \mathrm{d}^3\mathbf{x}\, \varphi(t, \mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}}, \tag{2.45}$$

$$\varphi(t, \mathbf{x}) = \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3}\, \varphi(t, \mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}}. \tag{2.46}$$

The translational invariance of the unperturbed universe means that the Fourier modes of the perturbations decouple, turning spatial derivatives into multiples of wavevectors. It is worth remarking that the above decomposition is specifically for a flat universe. There are analogous approaches for the open and closed cases, but for simplicity we shall keep to the flat case.

Second, a general spatial vector field v_i and a symmetric tensor field T_{ij} can be decomposed into helicity modes as:

$$\begin{aligned} v_i \rightsquigarrow & \partial_i v + v_i, \\ & (\partial^k v_k = 0), \end{aligned} \tag{2.47}$$

$$\begin{aligned} T_{ij} \rightsquigarrow & (\partial_i\partial_j - \frac{\delta_{ij}}{3}\partial^k\partial_k)T + \frac{1}{2}(\partial_i T_j + \partial_j T_i) + T_{ij}, \\ & (\partial^k T_k = \partial^k T_{ki} = T^k_k = 0), \end{aligned} \tag{2.48}$$

where "$\rightsquigarrow$" in this context should be read as "is decomposed as". Thus v and T are helicity scalars, v_i and T_i are divergenceless 3-vectors and T_{ij} is a divergenceless, symmetric, traceless 3-tensor[6]. The rotational invariance of the unperturbed universe means that the spatial vector and 3-tensor perturbations also decouple.

We may therefore decompose all the perturbations into Fourier components and scalar, vector and tensor parts, decoupling the modes into three parallel analyses without spatial derivatives.

For simplicity, we shall focus on the scalar part of the analysis, since vectors generically decay under an accelerated expansion, and tensors have extremely simple dynamics. Finally, one can absorb the Laplacian term of the shear tensor $\delta_{ij}\partial^k\partial_k E$ into the definition of the curvature perturbation Ψ, and henceforth we assume that $\nabla^2 E = 0$.

2.8.4 Gauge choice

Observant readers will have noted that there are too many perturbation variables and not enough constraints. This lack of constraint arises from the fact that the split into background and perturbation $X \to X + \delta X$ implied by (2.37) is more subtle than first appears.

In addition to perturbing the dynamical variables, one can also perturb the coordinate system:

$$t \to t + \delta t, \tag{2.49}$$

$$x^i \to x^i + \delta x^i, \tag{2.50}$$

[6]Note the overloading: v and T have different meanings on either side of each relation.

where in general the small perturbations δt and δx^i are functions of time and space. For a generic scalar field $\varphi = \varphi(t, x^i)$, this coordinate perturbation will cause an alteration of the field value:

$$\varphi \to \varphi - \dot{\varphi}\delta t - \partial_i \varphi \delta x^i. \tag{2.51}$$

This means that it is easy to conflate "true" perturbations in the variables with coordinate perturbations. In the extreme limit of no dynamical perturbation, one can see that the coordinate transformation (2.49) & (2.50) will in fact generate a "false" perturbation in φ of $\delta\varphi = -\dot{\varphi}\delta t - \partial_i\varphi\delta x^i$.

Careful calculation will show that the transformation of the dynamical scalar variables under (2.49) & (2.50) is in fact:

$$\begin{aligned} \Phi &\to \Phi - \delta\dot{t}, & \Psi &\to \Psi + H\delta t, \\ B &\to B + \delta t/a - a\delta\dot{x}, & E &\to E - \delta x, \\ \delta\rho &\to \delta\rho - \dot{\rho}\delta t, & \delta P &\to \delta P - \dot{P}\delta t, \\ \delta q &\to \delta q + (\rho + P)\delta t, & \delta\phi &\to \delta\phi - \dot{\phi}\delta t, \end{aligned} \tag{2.52}$$

where E and B are the helicity scalar parts of the shear and shift, and δx is the scalar part of the spatial coordinate perturbation (2.50). By choosing an appropriate coordinate transformation, one can remove the additional dynamical variables that are unconstrained by the Einstein equations. This is typically done by setting some of the perturbation variables equal to zero. This procedure is termed a *gauge choice*.[7] Examples of popular gauge choices can be found in Table 2.4.

Whilst choosing a gauge will give one a set of soluble equations, the question still remains whether the perturbations really are "true", or in some-way coordinate-dependent. A more robust approach proposed by Bardeen (1980) is to avoid the issue entirely and define the gauge independent variables:

$$\begin{aligned} \Phi^{(B)} &= \Phi - \dot{T}, & \Psi^{(B)} &= \Psi + HT, \\ \delta\rho^{(B)} &= \delta\rho - \dot{\rho}T, & \delta P^{(B)} &= \delta P - \dot{P}T, \\ \delta q^{(B)} &= \delta q + (\rho + P)T, & \delta\phi^{(B)} &= \delta\phi - \dot{\phi}T, \end{aligned} \tag{2.53}$$

where $T = a^2[\dot{E} - B/a]$, and the superscript (B) is for "Bardeen". The variables in (2.53) remain unchanged under gauge transformations, and are independent of coordinate perturbations. Any perturbation in these variables cannot be "gauged away". Of course, these are not unique, as any linear combination of them is also gauge-invariant, but the above set results in a reasonable basis.

Procedurally, one of the most powerful approaches is to choose the Newtonian gauge $B = E = B_i = 0$. In this gauge, the Bardeen variables (2.53) are equivalent to the Newtonian ones $X^{(B)} = X^{(\mathrm{Newt.})}$. At the end of the computation, we may note that any equations in $X^{(\mathrm{Newt.})}$ will be manifestly gauge-invariant, so we may re-promote all variables $X^{(\mathrm{Newt.})}$ to the gauge-invariant ones $X^{(B)}$.

[7] δX is defined as the difference between the value X has in the physical (perturbed) spacetime, and the value X has in the background (unperturbed) spacetime. This can only be done if there is a prescription for identifying points between the two spacetimes, and in the language of differential geometry, this is termed a *gauge choice*.

Name	Definition
Synchronous	$\Phi = B = 0$
Newtonian	$B = E = 0$
Uniform density	$\delta\rho = 0$ and e.g. $E = 0$
Comoving	$\delta q = E = 0$
Comoving orthogonal	$\delta q = B = 0$
Spatially-flat	$\Psi = E = 0$

Table 2.4: Popular gauge choices for the scalar perturbation equations.

2.9 Comoving curvature perturbation

2.9.1 Classical behaviour

It is convenient to examine the gauge-invariant variable:

$$\mathcal{R} = \Psi - \frac{H}{\rho + P}\delta q. \tag{2.54}$$

This is termed the *comoving curvature perturbation*. This has the powerful conservation property that adiabatic ($\delta P_{\rm en} = 0$) modes outside the horizon ($k \ll aH$) are conserved. Since one of the crucial properties of inflation is a shrinking horizon (Section 2.6.2), this means that an inflating universe acts to "freeze-out" extremely early-time primordial fluctuations. These re-enter the horizon at a later time where our understanding of physics is much more concrete, allowing us to observe the effects of extremely high energy physics on the universe we see today.

To show this conservation property, it is convenient to work in the comoving gauge ($\delta q = E = 0$), since in this gauge, $\mathcal{R} = \Psi$. We begin by stating the relevent equations. The time-space components of the perturbed Einstein equations ($\delta G^0_i = \frac{1}{m_{\rm p}^2}\delta T^0_i$) become:

$$\dot{\mathcal{R}} + H\Phi = 0. \tag{2.55}$$

The time-time component of the perturbed Einstein equation ($\delta G^0_0 = \frac{1}{m_{\rm p}^2}\delta T^0_0$) coupled with above equation shows:

$$\frac{k^2}{a^2}\left(aHB - \mathcal{R}\right) = \frac{\delta\rho}{2m_{\rm p}^2}. \tag{2.56}$$

The spatial components of the perturbed conservation equation $\delta(\nabla_\mu T^\mu_i) = 0$ become:

$$\delta P + (P + \rho)\Phi = 0. \tag{2.57}$$

The spatial components of the perturbed Einstein equations ($\delta G^i_j = \frac{1}{m_{\rm p}^2}\delta T^i_j$) read:

$$\Phi - \mathcal{R} + 2HaB + a\dot{B} = 0. \tag{2.58}$$

We now derive two equations that demonstrate $\mathcal{R}$ is conserved outside of the horizon. First, eliminating Φ from (2.58) using (2.56) and rearranging yields:

$$\frac{\mathrm{d}}{\mathrm{d}t}(a\mathcal{R}) = H\frac{\mathrm{d}}{\mathrm{d}t}(a^2 B). \tag{2.59}$$

Second, combining equations (2.55), (2.56) & (2.57), along with definition (2.40) for the entropic part of the pressure perturbation, we find:

$$\frac{\mathrm{d}\mathcal{R}}{\mathrm{d}\log a} = \frac{2m_{\mathrm{p}}^2 k^2}{a^2(\rho+P)}\frac{\dot{P}}{\dot{\rho}}(aHB - \mathcal{R}) + \frac{\delta P_{\mathrm{en}}}{\rho+P}. \tag{2.60}$$

where B is evaluated in the comoving gauge for both (2.59) and (2.60).

As the universe inflates, H is approximately constant and perturbations move far outside the horizon ($k \ll aH$). Equation (2.59) therefore shows that $\mathcal{R} - aHB \sim 1/a$, and since $\rho + P \sim H^2$, under these conditions equation (2.60) becomes:

$$\lim_{k \ll aH} \frac{\mathrm{d}\mathcal{R}}{\mathrm{d}\log a} = \frac{\delta P_{\mathrm{en}}}{\rho+P}, \tag{2.61}$$

which is a gauge-invariant expression. Outside the horizon, perturbations in $\mathcal{R}$ become driven purely by entropic pressure fluctuations. Thus for adiabatic perturbations ($\delta P_{\mathrm{en}} = 0$) $\mathcal{R}$ is conserved outside the horizon.

2.9.2 Quantum behaviour

Using the Lagrangians as defined by equations (2.3) & (2.30), we may construct the action:

$$S = \int \mathrm{d}^4x\sqrt{|g|}\mathcal{L}_G + \mathcal{L}_\phi. \tag{2.62}$$

If we expand this action to second-order in $\mathcal{R}$, we find (after some effort):

$$S^{(2)} = \int \mathrm{d}^4x\, a^3 \frac{\dot{\phi}^2}{H^2}\left[\dot{\mathcal{R}}^2 - a^{-2}(\partial_i\mathcal{R})^2\right]. \tag{2.63}$$

Switching to conformal time and defining:

$$v = z\mathcal{R}, \qquad z = \frac{a\dot{\phi}}{H}, \tag{2.64}$$

the action becomes (Baumann, 2009, App B):

$$S^{(2)} = \int \mathrm{d}^4x\, \left[(v')^2 - (\partial_i v)^2 + \frac{z''}{z}v^2\right], \tag{2.65}$$

where primes denote derivatives with respect to conformal time.

To quantise, we promote the field variable v to a quantum operator, and write it as a Fourier superposition:

$$v = \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3}\left[\chi_k(\eta)a_{\mathbf{k}}e^{i\mathbf{k}\cdot\mathbf{x}} + \chi_k^*(\eta)a_{\mathbf{k}}^\dagger e^{-i\mathbf{k}\cdot\mathbf{x}}\right], \tag{2.66}$$

of mode functions $v_{\mathbf{k}} = \chi_k a_{\mathbf{k}}$. Requiring that the canonical commutator relation:

$$[a_{\mathbf{k}}\,, a_{\mathbf{k}'}^\dagger] = (2\pi)^3\delta^{(3)}\left(\mathbf{k} - \mathbf{k}'\right), \tag{2.67}$$

holds true, then the wavefunction $\chi_k(\eta)$ must satisfy:

$$\chi_k'' + \left(k^2 - \frac{z''}{z}\right)\chi_k = 0, \tag{2.68}$$

$$\chi_k{}'\chi_k^* - \chi_k^{*\prime}\chi_k = -i. \tag{2.69}$$

2.9.3 de Sitter limit

A perfectly inflating universe is a de Sitter space with:

$$H = \text{const.} \Rightarrow \quad a = \frac{1}{H(\eta_{\text{end}} - \eta)} \propto e^{Ht}, \quad -\infty < \eta < \eta_{\text{end}}. \tag{2.70}$$

As discussed in Section 2.7.2, pure de Sitter space may be achieved in an inflationary scenario if $H, \phi = \text{const.}$, and is approximated under the conditions of slow roll (2.36). Thus, in a space where H and $\dot{\phi}$ are both constant, $z''/z = a''/a = 2(\eta_{\text{end}} - \eta)^{-2}$. Taking the limit as $\dot{\phi} \to 0$ creates a pure de Sitter scenario, and the differential equation (2.68) has the general solution:

$$\begin{aligned} \chi_k =& A_k \frac{1}{\sqrt{2k}} \left(1 - \frac{i}{k(\eta_{\text{end}} - \eta)}\right) e^{-ik(\eta_{\text{end}} - \eta)}, \\ &+ B_k \frac{1}{\sqrt{2k}} \left(1 + \frac{i}{k(\eta_{\text{end}} - \eta)}\right) e^{ik(\eta_{\text{end}} - \eta)}, \end{aligned} \tag{2.71}$$

whilst the mode normalisation condition (2.69) is:

$$|B_k|^2 - |A_k|^2 = 1. \tag{2.72}$$

The overall phase in the solution (2.71) is unimportant, so along with (2.72), there are two degrees of freedom remaining. Fixing these degrees of freedom amounts to choosing the vacuum, which is easy in de Sitter space: As $\eta \to -\infty$, the mode equations become those of Minkowski spacetime ($\chi_k'' + k^2 \chi_k = 0$). In this case, there is a well-defined choice: $A_k = 0$, $B_k = 1$, termed Bunch-Davies initial conditions. These conditions diagonalise the Hamiltonian and select the lowest energy vacuum state,[8] and will be discussed in greater detail in Chapter 6.

2.9.4 Power spectra

The two point correlation function of a general spatial field $\varphi(\mathbf{x})$ is defined as:

$$\xi_\varphi(\mathbf{r}) = \langle \varphi(\mathbf{x}) \varphi(\mathbf{x} + \mathbf{r}) \rangle. \tag{2.73}$$

Taking the Fourier transform of both sides with respect to $\mathbf{x}$ and $\mathbf{r}$ yields:

$$\langle \varphi(\mathbf{k}) \varphi(\mathbf{k}') \rangle = (2\pi)^3 \delta^{(3)}(\mathbf{k} + \mathbf{k}') P_\varphi(\mathbf{k}), \tag{2.74}$$

where the power spectrum:

$$P_\varphi(\mathbf{k}) = \int \mathrm{d}^3 r \, \xi_\varphi(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}}, \tag{2.75}$$

is the Fourier transform of the two-point correlation function.

For an isotropic and homogeneous space, the quantities $\xi_\varphi(\mathbf{r})$ and $P_\varphi(\mathbf{k})$ are all a function of spatial distance $\xi(\mathbf{r}) = \xi(r)$ or wavevector magnitude $P_\varphi(\mathbf{k}) = P_\varphi(k)$. It is also conventional in this case to define a "normalised" power spectrum as:

$$\mathcal{P}_\varphi(k) = \frac{k^3}{2\pi^2} P_\varphi(k), \tag{2.76}$$

[8] In the general case, these statements are not equivalent.

as this has the property that:

$$\int \mathrm{d}(\log k)\mathcal{P}_\varphi(k) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\, P_\varphi(k) = \big\langle \varphi^2(x)\big\rangle, \tag{2.77}$$

i.e. $\mathcal{P}_\varphi(k)$ may be thought of as the logarithmic spectral density of the pointwise spatial variance of the field.

2.9.5 Power spectrum of comoving curvature perturbations

To compute the power spectrum of comoving curvature perturbations $\mathcal{R}$, one therefore needs to compute the average on the left hand side of equation (2.74). The two-point spatial correlation is:

$$\xi_\mathcal{R}(\eta, \mathbf{r}) = \langle \mathcal{R}(\eta, \mathbf{x}+\mathbf{r})\mathcal{R}(\eta, \mathbf{x})\rangle = z^2\langle v(\eta, \mathbf{x}+\mathbf{r})v(\eta, \mathbf{x})\rangle. \tag{2.78}$$

Decomposing v into Fourier modes via equation (2.66), identifying the average $\langle\cdot\rangle$ above as a quantum average $\langle 0|\cdot|0\rangle$, and applying the usual properties of creation and annihilation operators along with the commutator relation (2.67) yields:

$$\xi_\mathcal{R}(\eta, \mathbf{r}) = \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3}\left|\frac{\chi_k}{z}\right|^2 e^{i\mathbf{k}\cdot\mathbf{r}}, \tag{2.79}$$

which allows us to read off:

$$P_\mathcal{R}(k) = \left|\frac{\chi_k}{z}\right|^2. \tag{2.80}$$

As show in Section 2.9.1, the value of $\mathcal{R}_k = \chi_k/z$ freezes out at the horizon crossing time t_* when $k \approx a_*H_*$. Under the assumption of slow-roll inflation, one can assume that the universe was effectively de Sitter from the moment of setting the Bunch-Davies initial conditions up until horizon crossing. For pure de Sitter space (2.70), as $\eta \to \eta_\mathrm{end}$,

$$|\chi_k|^2 \to 2k^3(\eta_\mathrm{end} - \eta)^{-2}, \qquad a \to H_*(\eta_\mathrm{end} - \eta)^{-1}. \tag{2.81}$$

Evaluating $z = a\dot\phi/H$ at horizon crossing under the above approximation and computing the dimensionless form (2.75) of the power spectrum (2.80) yields:

$$\mathcal{P}_\mathcal{R}(k) = \frac{H_*^2}{(2\pi)^2}\left(\frac{H_*}{\dot\phi_*}\right)^2. \tag{2.82}$$

This therefore predicts a slowly varying power spectrum, since different k modes exit at different times, H_* and $\dot\phi_*$ are k-dependent. The power spectrum is normally parameterised as:

$$\mathcal{P}_\mathcal{R}(k) = A_s\left(\frac{k}{k_s}\right)^{n_s-1}, \tag{2.83}$$

where k_s is some pre-defined pivot scale. Different inflationary models predict alternative values of A_s and n_s, with n_s typically less than 1, in contrast to the Harrison-Zel'dovich spectrum.

2.10 Statistics of the CMB

We write a general perturbation to the temperature of the photon field of the universe as:

$$T(t, \mathbf{x}, \mathbf{p}) = T(t) \left[1 + \Theta(t, \mathbf{x}, \mathbf{p})\right], \tag{2.84}$$

where $\mathbf{p}$ is a unit vector denoting the direction of the incoming photons as viewed at position $\mathbf{x}$. Note that this therefore encodes the inhomogeneities and anisotropies present in the photon field by virtue of having $\mathbf{x}$ and $\mathbf{p}$ dependence respectively. We may separate the angular dependence by expanding Θ in spherical harmonics:

$$\Theta(t, \mathbf{x}, \mathbf{p}) = \sum_{\ell=1}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m}(t, \mathbf{x}) Y_{\ell m}(\mathbf{p}), \tag{2.85}$$

$$a_{\ell m}(t, \mathbf{x}) = \int \mathrm{d}\Omega Y^*_{\ell m}(\mathbf{p}) \Theta(t, \mathbf{x}, \mathbf{p}). \tag{2.86}$$

We only observe the temperature field here (at $\mathbf{x}_0$) and now (at t_0). By taking sufficiently many measurements in different directions of $\mathbf{p}$ of the temperature field Θ we can compute the integral (2.86), to obtain $a_{\ell m}(t_0, \mathbf{x}_0)$ up to some $\ell_{\max}$.[9]

In general, theories do not predict specific values of $a_{\ell m}$, but merely tell us about the statistical distribution from which they are drawn. At each value of ℓ, the $(2\ell + 1)$ observations $\{a_{\ell m} : m = -\ell \cdots \ell\}$ are typically independent realisations of the same random variable. Their mean value is zero, but they will have some non-zero variance:

$$\langle a_{\ell m} \rangle = 0, \qquad \langle a_{\ell m} a^*_{\ell' m'} \rangle = \delta_{\ell \ell'} \delta_{m m'} C_\ell. \tag{2.87}$$

Note that the error in the sample variance obeys:

$$\left(\frac{\Delta C_\ell}{C_\ell} \right) = \sqrt{\frac{2}{2\ell + 1}}. \tag{2.88}$$

This means that on larger angular scales, we have fewer $a_{\ell m}$'s to use to compute the sample variance, and hence have a larger sample error. This is a manifestation of *cosmic variance*, resulting from the fact that we only have one universe to observe.

Computing the theoretical predictions of these C_ℓ's from a given cosmology is complicated, but amounts to an integral:

$$C_\ell^{XY} = \frac{2}{\pi} \int k^2 \, \mathrm{d}k \, P_{\mathcal{R}}(k) \Delta_{X\ell}(k) \Delta_{Y\ell}(k), \tag{2.89}$$

where $\Delta_{X\ell}(k)$ are transfer functions. These transfer functions are computed from line of sight integrals, requiring one to evolve the contents of the universe from the k-mode's horizon re-entry point through the hot big bang era, past the surface of last scattering at recombination and all the way to the current epoch before projecting the photon field onto the sky we see today. $X \in \{T, E, B\}$

[9] As a rough guide, if one has N_{pix} independent pixels, there are $(\ell_{\max} + 1)^2$ different $a_{\ell m}$'s to measure, so $\ell_{\max} \sim \sqrt{N_{\mathrm{pix}}}$.

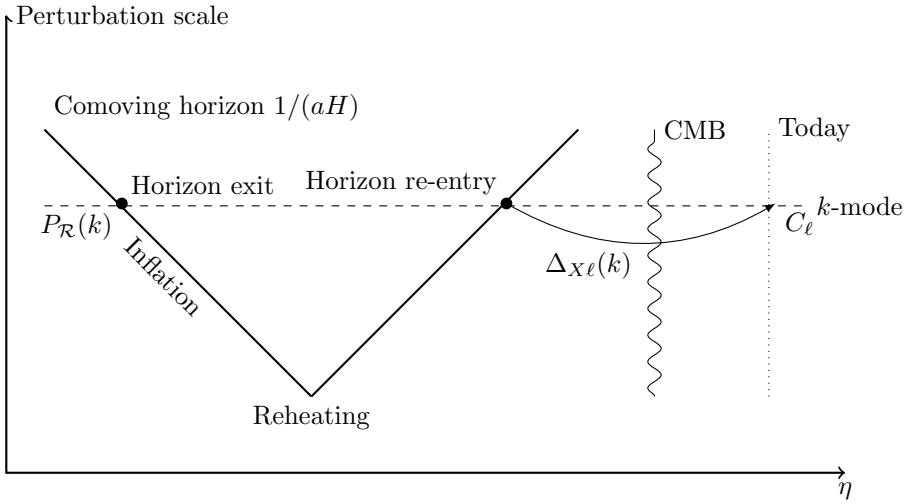


Figure 2.8: History of the early universe.

labels the components of the photon field corresponding to the temperature anisotropy, and the E and B polarisation modes respectively. Treating this calculation fully is beyond the scope of this thesis, but it suffices to say that this is now extremely well understood physics, and a full set of C_ℓ's can be computed for a given cosmology using Boltzmann codes such as CAMB (Lewis et al., 2000) in a matter of seconds.

2.11 Conclusions

We have now covered all the necessary background theory. The picture is summarised in Figure 2.8. Inflation's fundamental contribution to our model of the universe is that it predicts the primordial power spectrum $P_{\mathcal{R}}(k)$ at some early time. The universe's accelerated expansion then freezes these perturbations out until much later on in cosmic history, when we understand the physics in a lot more detail. We can then use standard physics to evolve these initial conditions set by inflation to make predictions about the sky that we see today.

Chapter 3

Kinetic dominance in the early universe

3.1 Introduction

Cosmological inflation was first introduced by Starobinsky (1979), Guth (1981) and others, and extended by Linde (1982) and several other workers to create modern inflationary theory. It is able to solve long-standing problems with the paradigm of big bang cosmology. In addition to solving the monopole, flatness and horizon problems, inflation provides a mechanism for generating superhorizon-scale cosmological perturbations from quantum fluctuations of the inflaton field (see, for example Mukhanov et al., 1992). Inflation thus predicts that large-scale structures in the universe are the result of quantum-mechanical fluctuations occurring during the inflationary epoch. Inflationary perturbations of this type are consistent with the anisotropy power spectrum of the cosmic microwave background (CMB) (Hinshaw et al., 2013, Planck Collaboration et al., 2013a).

In this chapter, we focus primarily on the background dynamics of single-field inflationary models, as determined by the evolution of the scalar field $\phi(t)$ and the Hubble parameter $H(t)$ as functions of cosmic time. This cosmological evolution can generally only be determined numerically, which requires initial conditions for the numerical integration. We therefore consider the limiting forms of the coupled dynamical equations for $\phi(t)$ and $H(t)$ as one evolves backwards in time and the universal scale factor $a \to 0$. We work under the extremely broad assumption that there exists a time prior to which $|\dot{\phi}| > \xi > 0$, for some positive constant ξ, as $a \to 0$. With this assumption, we show that as $a \to 0$ the kinetic energy of the inflaton comes to dominate the potential energy: $\dot{\phi}^2 \gg V(\phi)$. We call this condition *kinetic dominance* (KD). This is generically true, except perhaps for a single special solution for each potential $V(\phi)$.

Kinetically dominated universes emerge from a singularity at a finite time in the past and in a non-inflating state. This statement is true even if additional auxiliary fluids are present such as radiation, matter or curvature. In the kinetically dominated regime, the coupled equations of motion admit simple analytical solutions for ϕ and H, which do not depend on the form of the inflaton potential $V(\phi)$. These solutions therefore provide a simple way of

setting the initial conditions for such inflation models.

With these initial conditions in hand, we then analyse (numerically) the evolution of $\phi(t)$ and $H(t)$ in the flat case through to the end of inflation, thereby determining the background evolution, and also calculate the spectrum of scalar perturbations produced. We find that the latter generically has a cutoff at large spatial scales, which could provide an explanation for the recently observed low-ℓ falloff in the CMB power spectrum (Hinshaw et al., 2013, Planck Collaboration et al., 2013a).

Throughout this chapter we work many Planck times away from the singularity: $t \gg t_{\rm p}$ so we do not expect quantum gravitational effects to be present. In addition, the homogeneity and scale of the inflaton field indicates that the number of inflaton particles $n \gg 1$, so quantum field theoretic effects will not be present. Thus, although inflationary dynamics requires a rigorous quantum treatment, it is possible to adopt a classical phenomenological approach to setting initial conditions for the background dynamical variables.

The structure of this chapter is as follows. In Section 3.2, we will briefly introduce the dynamics of inflationary models based on a scalar field with the possibility of additional 'auxiliary' fluids. In Section 3.3 we prove the generic nature of kinetic dominance. We explore the consequences of kinetic dominance in Section 3.4 and present simple analytical solutions in this regime. We then illustrate the utility of the kinetically dominated phase in Section 3.5 by application to a spatially flat universe with polynomial and exponential potentials. We enumerate the solutions that do not obey our broad assumptions in Section 3.6. We conclude in Section 3.7. Appendix 3.A proves a uniqueness result crucial to the final step in the proof of kinetic dominance.

3.2 Scalar field inflation models

A universe comprised of multiple components with densities $\{\rho_i\}$ and pressures $\{P_i\}$ has the evolution equations:

$$\dot{H} + H^2 = -\frac{1}{6m_{\rm p}^2}\sum_i \left(\rho_i + 3P_i\right), \tag{3.1}$$

$$H^2 = \frac{1}{3m_{\rm p}^2}\sum_i \rho_i, \tag{3.2}$$

$$\dot{\rho}_i = -3(\rho_i + P_i)H, \tag{3.3}$$

where $H = \dot{a}/a$ is the Hubble parameter, a is the normalized scale factor and a dot denotes differentiation with respect to cosmic time, $\dot{f} \equiv \mathrm{d}f/\mathrm{d}t$. The first equation is the *acceleration* equation, and is derived from the trace of the Einstein equations. The second is the *Friedmann* equation and represents the conservation of energy. The third is the *continuity* equation for the fluid ρ_i. It should be noted that these equations are not independent, and that the acceleration equation may be straightforwardly derived from the Friedmann and continuity equations. For convenience, we use Planck units ($G = c = \hbar = 1$) throughout, but for clarity retain the reduced Planck mass:

$$m_{\rm p} = \sqrt{\frac{\hbar c}{8\pi G}} = (8\pi)^{-1/2}.$$

Type of fluid	w
Scalar field during KD	1
Radiation	1/3
Matter	0
Spatial curvature	$-1/3$
Missing matter	$-2/3$
Dark energy (cosmological constant)	-1

Table 3.1: Commonly assumed cosmological fluids and their equation-of-state parameters w, defined by equation equation (3.5). For more information on "missing matter", see Vazquez et al. (2012).

The simplest way to create a homogeneous and isotropic cosmological background model which undergoes an inflationary phase is by assuming that one of the fields is a real, time-dependent and homogeneous scalar field $\phi(t)$. The energy density and pressure of such a field is given by:

$$\rho_\phi = \tfrac{1}{2}\dot{\phi}^2 + V(\phi), \qquad P_\phi = \tfrac{1}{2}\dot{\phi}^2 - V(\phi). \tag{3.4}$$

In addition to the scalar field, we shall allow the possibility of including a collection of additional non-interacting fluids with densities $\{\rho_i\}$ and pressures $\{P_i\}$ defined by their equation-of-state parameters:

$$w_i = \frac{P_i}{\rho_i}, \tag{3.5}$$

where w_i are a set of constants determining the type of each fluid. Some commonly assumed cosmological fluids are listed in Table 3.1 along with their w-values. Note that we are accommodating the possibility of spatially curved universes implicitly by including the case $w_i = -1/3$. We shall term all of these *auxiliary fluids*.

Using the notation in (3.5) and defining the present-day densities $\{\rho_{i,0}\}$, the evolution equations (3.1)–(3.3) take the form:

$$\dot{H} + H^2 = -\frac{1}{3m_\mathrm{p}^2}\left[\dot{\phi}^2 - V(\phi) + \sum_i \tfrac{1}{2}(1+3w_i)\rho_i\right], \tag{3.6}$$

$$H^2 = \frac{1}{3m_\mathrm{p}^2}\left[\tfrac{1}{2}\dot{\phi}^2 + V(\phi) + \sum_i \rho_i\right], \tag{3.7}$$

$$\rho_i = \rho_{i,0}\, a^{-3(1+w_i)}, \tag{3.8}$$

$$0 = \ddot{\phi} + 3\dot{\phi}H + V'(\phi). \tag{3.9}$$

Inflation is defined as $\ddot{a} > 0$, or equivalently as $\dot{H} + H^2 > 0$. In the case when only an inflaton is present, this condition can be recast in terms of the scalar field using the acceleration equation (3.6) as:

$$\dot{\phi}^2 < V(\phi). \tag{3.10}$$

The slow-roll inflation regime satisfies:

$$\dot{\phi}^2 \ll V(\phi). \tag{3.11}$$

The amount of inflation is measured by the number of e-folds $N \propto \log a$, which is related to the Hubble parameter H by:

$$\dot{N} = H. \tag{3.12}$$

For a generic potential $V(\phi)$, there is no analytic solution for the dynamics of a scalar field inflation model, even if no other fluids are present. Hence, even in this simple case, the evolution equations (3.6) and (3.9) have to be integrated numerically using suitable "initial" conditions at some time $t = t_{\rm i}$. In principle, $t = t_{\rm i}$ may be *any* cosmic time, although numerical stability of the solution usually requires that the conditions be specified prior to the onset of inflation. Once any two of $\phi_{\rm i} \equiv \phi(t_{\rm i})$, $\dot{\phi}_{\rm i} \equiv \dot{\phi}(t_{\rm i})$ and $H_{\rm i} \equiv H(t_{\rm i})$ have been specified, the Friedmann equation (3.7) yields the third. The quantities $H_{\rm i}$, $\phi_{\rm i}$ and $\dot{\phi}_{\rm i}$ then provide the necessary initial conditions for the integration of the coupled dynamical equations (3.6) and (3.9) for $\phi(t)$ and $H(t)$.

3.3 Generic nature of kinetic dominance

As has been previously observed (Linde, 1985, Belinsky et al., 1985, Alimi et al., 1990), if one assumes that at some point early in the universe's history the opposite of the slow-roll condition (3.11) were true,

$$\dot{\phi}^2 \gg V(\phi), \tag{3.13}$$

then the evolution equations are analytically solvable. We call this condition kinetic dominance, since the potential energy of the field $V(\phi)$ is negligible in comparison to its kinetic energy $\frac{1}{2}\dot{\phi}^2$.

We shall restrict our attention to the very broad class of cosmological models that satisfy:

$$|\dot{\phi}| > \xi > 0 \qquad \text{as} \qquad a \to 0, \tag{3.14}$$

for some positive constant ξ. This condition demands that there be some epoch before which the inflaton evolves in a purely monotonic manner, which we shall refer to as a *steadily moving inflaton*. In this case, we find that the kinetic dominance condition (3.13) is entirely generic as $a \to 0$, and holds independently of the form of the potential $V(\phi)$.

From (3.13) it is then possible to show that the universe emerges from a singularity at a finite time in the past, which can be set to $t = 0$. In addition, kinetic dominance also implies that the (kinetic) energy density of the inflaton dominates the energy densities of all of the other components $\{\rho_i\}$ at early times, provided that $w_i < 1$. We shall leave the proof of these statements until Section 3.4.

We shall now prove the generic nature of kinetic dominance, i.e. that (3.14) implies (3.13). The proof runs as follows:

A. A new variable $N_{\rm e}$, termed the effective e-folds is introduced. This new variable enables one to assume without loss of generality that the potential $V(\phi)$ is positive.

B. The time coordinate t is rescaled to a new timelike coordinate τ, termed *Halliwell* time. This removes the majority of the potential dependence, and the two equations condense into a single equation in a new variable u.

C. The Hamilton-Jacobi representation is then utilized, exchanging Halliwell time for the field ϕ. The field ϕ is then rescaled to a new variable ψ, absorbing the Hubble parameter and implicitly all of the $\{\rho_i\}$ dependence. One final monotonic transformation of the dependent variable u is made, leaving a single differential equation for a function y with ψ as the independent variable.

D. The resulting equation has the property that, for any given potential $V(\phi)$, there is at most a single solution $f(\psi)$ that is both finite and positive. All other positive solutions $y(\psi)$ are divergent. When interpreted, a positive $y(\psi)$ corresponds to a steadily moving inflaton, and a diverging $y(\psi)$ represents a kinetically dominated universe.

3.3.1 Effective e-folds

We define a new function $\dot{N}_{\rm e}$ by the relation:

$$\dot{N}_{\rm e}^2 = \frac{1}{3m_{\rm p}^2}\left(\tfrac{1}{2}\dot{\phi}^2 + V_1 + V(\phi)\right), \tag{3.15}$$

where V_1 is a positive constant, the value of which will be discussed shortly. One may regard this as a modified Friedmann equation, where the explicit dependence on the auxiliary fields has been absorbed into a new parameter $\dot{N}_{\rm e}$. The variable $N_{\rm e}$ can be interpreted as the "effective e-folds." In the case where the fluids may be neglected, one finds that $\dot{N}_{\rm e} \approx \dot{N}$.

By differentiating the above definition, it is simple to show using the Klein-Gordon equation (3.9) that:

$$\frac{\dot{N}_{\rm e}\ddot{N}_{\rm e}}{\dot{N}} = -\frac{\dot{\phi}^2}{2m_{\rm p}^2}, \tag{3.16}$$

where $\dot{N} = H$ from equation (3.12). The Hubble parameter $\dot{N}$ can be related to the effective e-folds by combining the Friedmann equation (3.7) with equation (3.15):

$$\dot{N}^2 = \dot{N}_{\rm e}^2 - \frac{V_1}{3m_{\rm p}^2} + \frac{1}{3m_{\rm p}^2}\sum_i \rho_i. \tag{3.17}$$

Equations (3.8), (3.15)–(3.17) may now be regarded as the evolution equations for the system in the variables $\{N, N_{\rm e}, \phi, \rho_i\}$. The potential $V(\phi)$ now only arises in (3.15) in combination with V_1. Since all physical potentials are bounded below, one can choose V_1 such that $V_1 + V(\phi)$ is always positive. One can therefore treat $V_1 + V(\phi)$ as the new "effective" potential and hence we drop the V_1 part from (3.15) and assume $V(\phi)$ is positive.

3.3.2 Halliwell time

We now define a new time coordinate τ, such that:

$$\frac{{\rm d}\tau}{{\rm d}t} = \sqrt{V(\phi)} \quad \Leftrightarrow \quad \tau = \int^t \sqrt{V(\phi)}\ {\rm d}t. \tag{3.18}$$

This relation is well defined (up to a constant) and one-to-one as from the above section one may assume that $V(\phi)$ is finite and positive. Physically, (3.18)

corresponds to choosing a measure of time in which the inflaton "sees" a near-constant potential. This approach is analogous to the method used by Halliwell (1987) in his work with exponential potentials. We shall thus term this new timelike coordinate *Halliwell time*.

Under this rescaling of time, the modified evolution equations (3.15) and (3.16) take the form:

$$N_{\mathrm{e}}{}'^2 = \frac{1}{3m_{\mathrm{p}}^2}\left(\tfrac{1}{2}\phi'^2 + 1\right), \tag{3.19}$$

$$-\frac{\phi'^2}{2m_{\mathrm{p}}^2} = \frac{N_{\mathrm{e}}{}'}{N'}\left(N_{\mathrm{e}}{}'' + \frac{1}{2}\phi' N_{\mathrm{e}}{}' \frac{\mathrm{d}}{\mathrm{d}\phi}\log V\right), \tag{3.20}$$

where a prime denotes differentiation with respect to τ. Equation (3.19) states that the dynamical variables $N_{\mathrm{e}}{}'$ and ϕ' lie on a hyperbola with asymptotic ratio $\left(\sqrt{6}m_{\mathrm{p}}\right)^{-1}$, as illustrated in Figure 3.1. Since one may take $N_{\mathrm{e}}{}' > 0$, a sensible parametrization therefore is in terms of a hyperbolic angle u:

$$N_{\mathrm{e}}{}' = \frac{1}{m_{\mathrm{p}}\sqrt{3}}\cosh u, \tag{3.21}$$

$$\phi' = -\sqrt{2}\sinh u. \tag{3.22}$$

Applying the transformation above to the Halliwell-time evolution equations, we find that equation (3.19) is trivially satisfied, and equation (3.20) takes the form:

$$\frac{N_{\mathrm{e}}{}'}{N'}\frac{m_{\mathrm{p}}}{\sqrt{6}}\left(\frac{\sqrt{2}}{\sinh u}\frac{\mathrm{d}u}{\mathrm{d}\tau} - \frac{1}{\tanh u}\frac{\mathrm{d}}{\mathrm{d}\phi}\log V\right) = -1. \tag{3.23}$$

3.3.3 Hamilton-Jacobi representation

We reformulate the equation using the Hamilton-Jacobi representation; instead of considering the variables as functions of time τ, one uses the field ϕ as the independent variable. Since we are considering universes with a monotonic inflaton ($\dot{\phi} \neq 0$), the transformation from t to ϕ is monotonic, and hence so too is that from τ to ϕ.

One can switch to the Hamilton-Jacobi representation by changing the variables in the derivatives using the relation:

$$\frac{\mathrm{d}}{\mathrm{d}\tau} = \frac{\mathrm{d}\phi}{\mathrm{d}\tau}\frac{\mathrm{d}}{\mathrm{d}\phi} = \phi'\frac{\mathrm{d}}{\mathrm{d}\phi} = -\sqrt{2}\sinh u\frac{\mathrm{d}}{\mathrm{d}\phi}, \tag{3.24}$$

which on applying to equation (3.23) yields:

$$m_{\mathrm{p}}\frac{N_{\mathrm{e}}{}'}{N'}\sqrt{\frac{2}{3}}\left(\frac{\mathrm{d}u}{\mathrm{d}\phi} + \frac{1}{2\tanh u}\frac{\mathrm{d}}{\mathrm{d}\phi}\log V\right) = 1. \tag{3.25}$$

We now rescale the ϕ field into a new field ψ via the relation:

$$\frac{\mathrm{d}}{\mathrm{d}\psi} = m_{\mathrm{p}}\frac{N_{\mathrm{e}}{}'}{N'}\sqrt{\frac{2}{3}}\frac{\mathrm{d}}{\mathrm{d}\phi}. \tag{3.26}$$

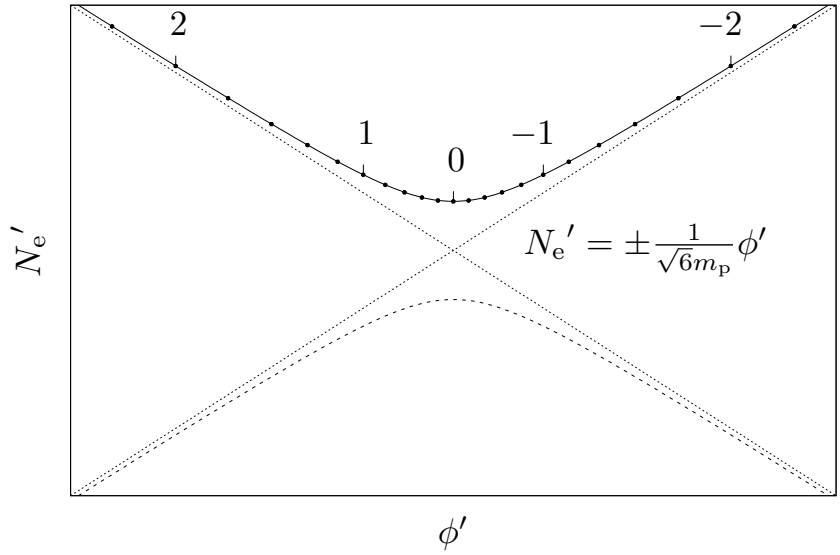


Figure 3.1: The constraint provided by the definition of N_{e} in Halliwell time. The dynamical variables $N_{\mathrm{e}}{}'$ and ϕ' lie on a hyperbola according to (3.19). A natural parametrization uses a hyperbolic angle u detailed in (3.21) and (3.22). Note that only the upper half is parametrized, as the lower half of the hyperbola suggests a collapsing universe ($N_{\mathrm{e}}{}' \propto H \propto \dot{a} < 0$). The points corresponding to $u \in \{-2, -1, 0, 1, 2\}$ have been labelled to guide the eye.

More explicitly, ψ is defined up to a constant by the monotonic transformation:

$$\frac{\mathrm{d}\psi}{\mathrm{d}\phi} = \sqrt{\frac{3}{2}}\frac{N'}{m_{\mathrm{p}}N_{\mathrm{e}}{}'} \quad \Leftrightarrow \quad \psi = \sqrt{\frac{3}{2}}\frac{1}{m_{\mathrm{p}}}\int^{\phi}\frac{N'}{N_{\mathrm{e}}{}'}\,d\phi, \tag{3.27}$$

and this relationship is well defined since:

$$\frac{N'}{N_{\mathrm{e}}{}'} = \frac{H}{\dot{N}_{\mathrm{e}}} \geq 0. \tag{3.28}$$

This rescaling absorbs all of the dependence on N' and thus $\{\rho_i\}$ via (3.17) into the definition of ψ. Under the transformation (3.27), the master equation (3.25) takes the form:

$$\frac{\mathrm{d}u}{\mathrm{d}\psi} = 1 - \frac{1}{\tanh u}\frac{\mathrm{d}}{\mathrm{d}\psi}\log\sqrt{V}. \tag{3.29}$$

We have transformed the evolution equations (3.6)–(3.9) into a single "master equation" in one variable, where all of the potential dependence is kept in a single term.

We may rearrange this slightly by making the monotonic transformation:

$$u \mapsto y = \log\cosh(u), \tag{3.30}$$

under which the master equation takes the form:

$$\frac{\mathrm{d}y}{\mathrm{d}\psi} = \sqrt{1 - e^{-2y}} - \frac{\mathrm{d}}{\mathrm{d}\psi}\log\sqrt{V}. \tag{3.31}$$

3.3.4 Interpreting the master equation

We now prove kinetic dominance by considering the asymptotics of the master equation (3.31) in the limit $a \to 0$. We prove that we may assume wlog that $|\psi| \to \infty$ as $a \to 0$. Given this, we prove that there is at most a single solution $f(\psi)$ which is finite, with all of the rest diverging as $a \to 0$. We finish by showing that a diverging solution $y(\psi)$ of the master equation is equivalent to kinetic dominance.

We begin by examining the behaviour of ψ as $a \to 0$. Through elementary derivative transformations with the chain rule and the definitions of various variables, we find:

$$\begin{aligned}
\frac{\phi'}{{N_\mathrm{e}}'} &= \frac{\psi' \frac{\mathrm{d}\phi}{\mathrm{d}\psi}}{{N_\mathrm{e}}'}, && \text{(chain rule)} \\
&= m_\mathrm{p}\sqrt{\frac{2}{3}}\frac{\psi'}{N'}, && \text{(from equation 3.27)} \\
&= m_\mathrm{p}\sqrt{\frac{2}{3}}\frac{\dot{\psi}}{\dot{N}}, && \text{(chain rule)} \\
&= m_\mathrm{p}\sqrt{\frac{2}{3}}\frac{\mathrm{d}\psi}{\mathrm{d}\log a}. && \text{(since } H = \frac{\mathrm{d}}{\mathrm{d}t}\log a\text{)} \qquad (3.32)
\end{aligned}$$

Given the definition of Halliwell time (3.18), on the left hand side of the above expression, we have $\phi' = \dot{\phi}/\sqrt{V}$. Since ${N_\mathrm{e}}' > 0$, and by assumption (3.14) $|\dot{\phi}| > \xi > 0$ we thus find that ψ is monotonic in a. The direction of the monotonicity of ψ can be found by considering the transformations we have made:

$$t \xrightarrow{(3.18)} \tau \xrightarrow{(3.22)} \phi \xrightarrow{(3.27)} \psi. \qquad (3.33)$$

By considering the equations denoted above, one can see that $\frac{\mathrm{d}\tau}{\mathrm{d}t} > 0$, $\frac{\mathrm{d}\psi}{\mathrm{d}\phi} > 0$, and $\frac{\mathrm{d}\phi}{\mathrm{d}\tau} = -\sqrt{2}\sinh u$. Thus as a and t decrease, one finds that if $u > 0$, then ψ is monotonically *increasing*.[1] As we are considering universes where $\dot{\phi} \neq 0$, and wlog $V(\phi) > 0$, this places the constraint that:

$$u = \sinh^{-1}\left(\frac{\phi'}{\sqrt{2}}\right) = \sinh^{-1}\left(\frac{\dot{\phi}}{\sqrt{2}V(\phi)}\right) \neq 0. \qquad (3.34)$$

The problem therefore breaks down into two possibilities: $u > 0$ and ψ increasing, or $u < 0$ and ψ decreasing. We will consider the first possibility; the second may be treated in exactly the same way, with a couple of sign changes.

If ψ is monotonically increasing, then as $a \to 0$, either $\psi \to \psi_\mathrm{max}$ or it diverges $\psi \to \infty$. In the first of these possibilities, it follows that:

$$\psi \to \psi_\mathrm{max} \quad \text{as} \quad a \to 0 \qquad \Rightarrow \qquad \frac{\mathrm{d}\psi}{\mathrm{d}\log a} \to 0. \qquad (3.35)$$

Now, since:

$$\frac{\mathrm{d}\psi}{\mathrm{d}\log a} \propto \frac{\phi'}{{N_\mathrm{e}}'} \propto \frac{\dot{\phi}}{\sqrt{\frac{1}{2}\dot{\phi}^2 + V(\phi)}}, \qquad (3.36)$$

[1] This explains the choice of sign in the parametrization (3.22).

the assumption of a steadily moving inflaton (3.14) (i.e. $|\dot{\phi}| > \xi > 0$ as $a \to 0$) means that if the left hand side of the above tends to zero, then $V(\phi)$ must diverge. We now show that this set of statements contradict our initial assumptions. Integrating the master equation (3.31) from some start point up to $\psi_{\max}$ yields:

$$y(\psi) = \int^{\psi_{\max}} \sqrt{1 - e^{-2y}}\, \mathrm{d}\psi - \log\sqrt{V} + c. \tag{3.37}$$

Given that the integrand on the right hand side is bounded, and that the integral is over a finite range, the first term remains finite. However, we know that when $\psi \to \psi_{\max}$, the potential V and hence $\log\sqrt{V}$ must diverge. Thus, the solution y in the above equation must become negative as $\psi \to \psi_{\max}$, which contradicts the definition of y from (3.30). Thus, since $\psi \not\to \psi_{\max}$, we may assume that $\psi \to \infty$ as $a \to 0$.

We now show that as $\psi \to \infty$ there is at most one *finite* solution y satisfying the master equation (3.31) while remaining non-zero. Let us assume that there exists a solution $f(\psi)$ of (3.31) that is positive and finite ($0 < f < f_{\max}$ for some finite $f_{\max}$):

$$\frac{\mathrm{d}f}{\mathrm{d}\psi} = \sqrt{1 - e^{-2f}} - \frac{\mathrm{d}}{\mathrm{d}\psi}\log\sqrt{V}, \tag{3.38}$$

where:

$$0 < f(\psi) < f_{\max}, \tag{3.39}$$

and assume some initial condition on f at some finite value $\psi = \psi_0$:

$$f(\psi_0) = f_0. \tag{3.40}$$

Now consider a solution $h(\psi)$ with some larger initial value:

$$h(\psi_0) = h_0 > f_0. \tag{3.41}$$

Since h is also a solution of the master equation (3.31) it satisfies:

$$\frac{\mathrm{d}h}{\mathrm{d}\psi} = \sqrt{1 - e^{-2h}} - \frac{\mathrm{d}}{\mathrm{d}\psi}\log\sqrt{V}. \tag{3.42}$$

Taking the difference of (3.38) and (3.42) gives:

$$\frac{\mathrm{d}}{\mathrm{d}\psi}(h - f) = \sqrt{1 - e^{-2h}} - \sqrt{1 - e^{-2f}}. \tag{3.43}$$

By a uniqueness theorem, discussed in Appendix 3.A, if $h(\psi_0) > f(\psi_0)$, then $h(\psi_1) > f(\psi_1)$ (for $\psi_1 \neq \psi_0$). Therefore one can see from (3.43) that the difference between h and f is monotonically increasing in any finite interval $[\psi_0, \psi_1]$. One can thus conclude that:

$$h - f > h_0 - f_0 = \Delta_0 > 0, \tag{3.44}$$

where we have defined $\Delta_0 = h_0 - f_0$, and h and f are evaluated at the end of the interval ψ_1. Since $h > \Delta_0 + f$, it is easy to see using (3.43) that:

$$\frac{\mathrm{d}}{\mathrm{d}\psi}(h - f) > \sqrt{1 - e^{-2(f+\Delta_0)}} - \sqrt{1 - e^{-2f}}. \tag{3.45}$$

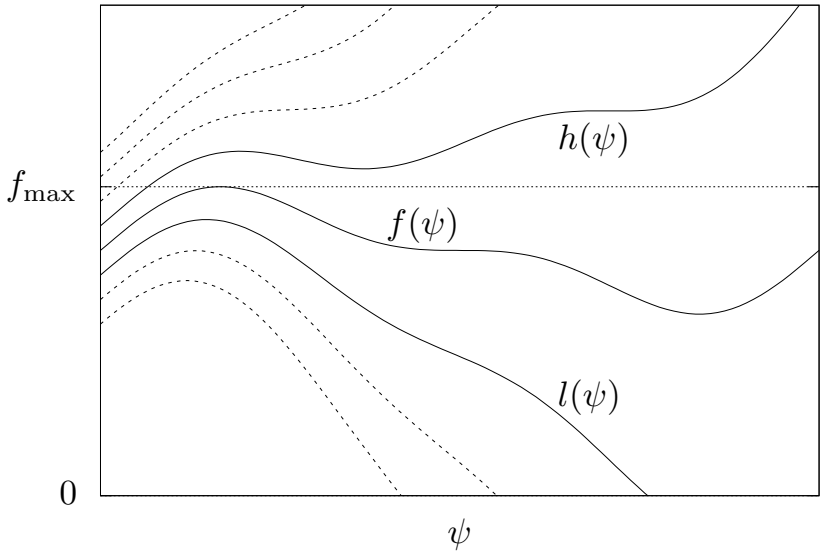


Figure 3.2: In general there is at most a single solution $f(\psi)$ to the master equation (3.31) that is positive and finite $0 < f(\psi) < f_{\max}$. Any solution $h(\psi)$ that begins higher than $f(\psi)$ must diverge. Consequently, any solution $l(\psi)$ that begins lower than $f(\psi)$ cannot remain above 0. If it were able to, then $l(\psi)$ would also be another finite solution lying between 0 and some $l_{\max}$. By the previous argument, this would mean f could not be finite, contradicting our initial assumption.

This is a monotonically decreasing function of f, and hence attains its minimum at $f_{\max}$; thus,

$$\frac{\mathrm{d}}{\mathrm{d}\psi}(h-f) > \sqrt{1-e^{-2(f_{\max}+\Delta_0)}} - \sqrt{1-e^{-2f_{\max}}} > 0. \tag{3.46}$$

The difference between h and f is therefore monotonically increasing at a rate greater than some positive number. Since f is positive, it is bounded below and as ψ_1 is made arbitrarily large, h grows without bound.

We therefore find ourselves in a situation demonstrated in Figure 3.2. For any given potential $V(\phi)$, there is at most a single solution $f(\psi)$ that is finite and positive $[0 < f(\psi) < f_{\max}]$: any solution that is larger than $f(\psi)$ at some point $\psi = \psi_0$ diverges as $\psi \to \infty$ $(a \to 0)$. Further, any solution $l(\psi)$ that starts out less than $f(\psi)$ must fall to a value less than 0. If $l(\psi)$ did not fall below 0, but were another example of a finite positive solution, then by the argument above, this would imply $f(\psi) > l(\psi)$ must diverge, contradicting our initial assumptions on $f(\psi)$.

One therefore expects universes with a steadily moving inflaton to have a generically diverging y as $a \to 0$, except perhaps for a single special case for a given potential $V(\phi)$. The consequences of a generically divergent y shall now be examined. If y diverges, then so does u by equation (3.30). If u diverges,

then we find that ϕ' diverges by (3.22):

$$\frac{\mathrm{d}\phi}{\mathrm{d}\tau} = \phi' = -\sqrt{2}\sinh u \to -\infty. \tag{3.47}$$

Converting back from Halliwell time to cosmic time using equation (3.18) shows:

$$\phi'^2 = \frac{\dot{\phi}^2}{V(\phi)}. \tag{3.48}$$

One can thus see that the divergence of y, u and ϕ' therefore requires that:

$$\lim_{y\to\pm\infty} \frac{\dot{\phi}^2}{V(\phi)} = \lim_{a\to 0} \frac{\dot{\phi}^2}{V(\phi)} = \infty, \tag{3.49}$$

which is equivalent to saying that $\dot{\phi}^2 \gg V(\phi)$ as $a \to 0$. The early universe is generically kinetically dominated.

3.4 Consequences of kinetic dominance

The condition $\dot{\phi}^2 \gg V(\phi)$ for kinetic dominance allows one to derive several results. First, the kinetic energy of the inflaton dominates over the other fluids as $a \to 0$, allowing us at early times to neglect any additional effects such as curvature, radiation, matter or a cosmological constant. Second, the universe emerges from an initial singularity at a finite coordinate time, which may be taken as $t = 0$. Finally, one is able to determine exact analytic expressions for the solutions in coordinate or conformal time for each of the cases where curvature, radiation, matter or a cosmological constant are present.

3.4.1 Dominance of $\dot{\phi}^2$ over other fluids

In the limit that $a \to 0$, the auxiliary fluid with the largest value of w dominates over all of the others. Along with kinetic dominance, we can therefore assume that the acceleration (3.6) and Friedmann (3.7) equations take the form:

$$\dot{H} + H^2 = -\frac{\dot{\phi}^2}{3m_\mathrm{p}^2} - \frac{1}{6m_\mathrm{p}^2}(1+3w)\rho_w, \tag{3.50}$$

$$H^2 = \frac{\dot{\phi}^2}{6m_\mathrm{p}^2} + \frac{1}{3m_\mathrm{p}^2}\rho_w, \tag{3.51}$$

where ρ_w is the density of the auxiliary fluid with the largest w. It is not difficult to show using equation (3.8), along with $H = \frac{\mathrm{d}}{\mathrm{d}t}\log a$ that these equations solve to give:

$$H^2 = \frac{1}{3m_\mathrm{p}^2}\left(\frac{\beta^2}{a^6} + \rho_w\right), \tag{3.52}$$

$$\dot{\phi}^2 = 2\frac{\beta^2}{a^6}, \tag{3.53}$$

where β is an integration constant. From this, if $w < 1$ then as $a \to 0$, one finds:

$$\dot{\phi}^2 \propto a^{-6} \gg \rho_w \propto a^{-3(1+w)}. \tag{3.54}$$

Thus, the kinetic term of the inflaton dominates over all other fluids with $w < 1$.

Historically, inflationary potentials were considered in the context of grand unified theories (Albrecht and Steinhardt, 1982, Linde, 1982) which resulted in an effective potential $V(\phi, T)$ depending on the value of the field ϕ and a temperature. This was then developed (Berera and Fang, 1995, Berera, 1995) into a theory in which the inflaton remains in thermal equilibrium with an auxiliary radiation fluid.

More recent work (Powell and Kinney, 2007) typically assumes that the inflaton is decoupled from the auxiliary fluids in the preinflationary phase, and that the universe is *radiation dominated* at this early stage. Given the above result (3.54), such assumptions may now need revisiting.

3.4.2 Finite time singularity

Since $H = \dot{a}/a$, we can express the coordinate time t as an integral:

$$t = \int \frac{\mathrm{d}a}{aH}. \tag{3.55}$$

From equation (3.52), one can see that $(aH)^{-1}$ is finite as $a \to 0$. By the above integral, this shows that the universe emerges at a finite time in the past, which can be taken as $t = 0$. Moreover, from (3.54), the (dominant) energy density of the universe scales as a^{-6} as $a \to 0$, showing that $t = 0$ is a singularity.

3.4.3 Analytic solutions for the kinetically dominated universe

If one considers the solutions of the acceleration and Friedmann equations (3.6) and (3.7) in the limit that $a \to 0$, then one can neglect the potential term $V(\phi)$ as it is suppressed by $\dot{\phi}^2$. In addition, the other fluid terms are negligible in comparison to the term with the largest w. When these considerations are taken into account, the evolution equations take the form shown in (3.50) and (3.51). One can find solutions for $\phi(a), H(a)$ and $\rho_w(a)$ parametrically in terms of a. In addition, one can find coordinate time $t(a)$ in terms of a using the relation:

$$t = \int \frac{\mathrm{d}a}{aH(a)}. \tag{3.56}$$

Conformal time η is defined by the equation $\dot{\eta} = a^{-1}$, and can be found in terms of a using:

$$\eta = \int \frac{\mathrm{d}a}{a^2 H(a)}. \tag{3.57}$$

The solutions are:

$$\rho_w(a) \propto a^{-3(1+w)}, \tag{3.58}$$

$$H(a)^2 = \frac{1}{3m_{\rm p}^2}\left(\frac{\beta^2}{a^6} + \rho_w\right), \tag{3.59}$$

$$\dot{\phi}(a)^2 = 2\frac{\beta^2}{a^6}, \tag{3.60}$$

$$\phi(a) = c \pm \sqrt{\frac{2}{3}}\frac{m_{\rm p}}{1-w}\log\left[\frac{a^{3(1-w)}}{\left(\sqrt{1+\frac{\rho_w a^6}{\beta^2}}+1\right)^2}\right], \tag{3.61}$$

$$t(a) = a^3\frac{m_{\rm p}\sqrt{3}}{3\beta}\ {}_2F_1\left(\frac{1}{2}, \frac{1}{1-w}; \frac{2-w}{1-w}, -\frac{a^6\rho_w}{\beta^2}\right), \tag{3.62}$$

$$\eta(a) = a^2\frac{m_{\rm p}\sqrt{3}}{2\beta}\ {}_2F_1\left(\frac{1}{2}, \frac{2}{3(1-w)}; \frac{5-3w}{3(1-w)}, -\frac{a^6\rho_w}{\beta^2}\right), \tag{3.63}$$

where β and c are constants of integration which will be redefined shortly. We have chosen t, η such that $a \to 0$ as $t, \eta \to 0$.

For specific values of w, the hypergeometric functions ${}_2F_1$ take simple forms. If $w = -1$ or 0, then equation (3.62) is expressible in closed form, in terms of trigonometric and algebraic functions in a respectively. If $w = -1/3$ or $1/3$, then (3.63) may be expressed in closed form. In each of these cases, these equations are invertible giving an expression for $a(t)$ or $a(\eta)$. We note that, except for the case $w = -1/3$, the solutions (3.59)–(3.63) correspond to a spatially flat universe.

We shall examine each of the above cases in turn, after first looking at the case in which there are no auxiliary fields, $\rho_w = 0$. In so doing, it will prove useful to define the functions:

$$\begin{matrix} S_k(x), & S_k^{-1}(x) \\ C_k(x), & C_k^{-1}(x) \\ T_k(x), & T_k^{-1}(x) \end{matrix} = \begin{cases} \begin{matrix} \sin(x), & \arcsin(x) \\ \cos(x), & \arccos(x) \\ \tan(x), & \arctan(x) \end{matrix} & : k > 0 \\ \begin{matrix} x, & x \\ 1, & 1 \\ x, & x \end{matrix} & : k = 0 \\ \begin{matrix} \sinh(x), & \operatorname{arcsinh}(x) \\ \cosh(x), & \operatorname{arccosh}(x) \\ \tanh(x), & \operatorname{arctanh}(x) \end{matrix} & : k < 0. \end{cases} \tag{3.64}$$

No auxiliary fields, $\rho_w = 0$

If $\rho_w = 0$, then equation (3.62) becomes:

$$t = a^3\frac{m_{\rm p}\sqrt{3}}{3\beta}. \tag{3.65}$$

One can rearrange this to find a as a function of t,

$$a(t) = t^{1/3}\left(\frac{3\beta}{m_{\rm p}\sqrt{3}}\right)^{1/3}, \tag{3.66}$$

and then substitute this into equations (3.59) and (3.60) to find:

$$H(t) = \frac{1}{3t}, \tag{3.67}$$

$$\dot{\phi}(t) = \pm\sqrt{\frac{2}{3}}\frac{m_{\rm p}}{t}. \tag{3.68}$$

The latter integrates to give:

$$\phi(t) = \phi_{\rm p} \pm \sqrt{\frac{2}{3}} m_{\rm p} \log\left(\frac{t}{t_{\rm p}}\right), \tag{3.69}$$

where $\phi_{\rm p}$ is an integration constant chosen such that $\phi(t_{\rm p}) = \phi_{\rm p}$, where $t_{\rm p}$ is some time. It is more appropriate to redefine the integration constant β as:

$$\beta \equiv \frac{a_{\rm p}^3 m_{\rm p}}{t_{\rm p}\sqrt{3}}, \tag{3.70}$$

since then, equation (3.66) for $a(t)$ becomes:

$$a(t) = a_{\rm p}\left(\frac{t}{t_{\rm p}}\right)^{1/3}, \tag{3.71}$$

which is more in keeping with equation (3.69).

For this case, one can also obtain analytical solutions in terms of conformal time. If $\rho_w = 0$ and β is defined by (3.70), equation (3.63) becomes:

$$\eta(t) = a^2 \frac{m_{\rm p}\sqrt{3}}{2\beta} = \frac{3t_{\rm p}}{2a_{\rm p}^3} a^2. \tag{3.72}$$

Using equation (3.71), we can show:

$$\eta = \eta_{\rm p}\left(\frac{t}{t_{\rm p}}\right)^{2/3}, \tag{3.73}$$

where we have defined $\eta_{\rm p}$ as:

$$\eta_{\rm p} = \frac{3t_{\rm p}}{2a_{\rm p}}. \tag{3.74}$$

Now we have $\eta(t)$ in equation (3.72), we can change equations (3.67), (3.68), (3.69) and (3.71) to:

$$a(\eta) = a_{\rm p}\left(\frac{\eta}{\eta_{\rm p}}\right)^{1/2}, \tag{3.75}$$

$$H(\eta) = \frac{1}{3t_{\rm p}}\left(\frac{\eta}{\eta_{\rm p}}\right)^{-3/2}, \tag{3.76}$$

$$\dot{\phi}(\eta) = \pm\sqrt{\frac{2}{3}}\frac{m_{\rm p}}{t_{\rm p}}\left(\frac{\eta}{\eta_{\rm p}}\right)^{-3/2}, \tag{3.77}$$

$$\phi(\eta) = \phi_{\rm p} \pm \sqrt{\frac{3}{2}} m_{\rm p} \log\left(\frac{\eta}{\eta_{\rm p}}\right). \tag{3.78}$$

It should be noted that since $\dot{\phi}^2 \gg \rho_w$ at sufficiently early times, all solutions reduce to the above forms for small enough t or η. We can thus fix the form of solutions with nonzero ρ_w by matching onto the above solutions for sufficiently small t or η.

Dark energy, $w = -1$

For dark energy in the form of a cosmological constant, we find that the energy density in standard notation is:

$$\rho_w = m_{\rm p}^2 \Lambda. \tag{3.79}$$

For $w = -1$, equation (3.62) is expressible in terms of trigonometric functions, and may be rearranged to express the scale factor a in terms of coordinate time. Once $a(t)$ is obtained, the remaining equations (3.59), (3.60) and (3.61) can be used to express the rest of the variables in terms of t. Using our definition of β (equation 3.70), and defining the new time scale,

$$t_\Lambda = \frac{1}{\sqrt{3\Lambda}}, \tag{3.80}$$

the solutions are:

$$a(t) = a_{\rm p}\left[\frac{S_{-\Lambda}(t/t_\Lambda)}{t_{\rm p}/t_\Lambda}\right]^{1/3}, \tag{3.81}$$

$$H(t) = \frac{1}{3t_\Lambda}\frac{1}{T_{-\Lambda}(t/t_\Lambda)}, \tag{3.82}$$

$$\dot{\phi}(t) = \sqrt{\frac{2}{3}}\frac{m_{\rm p}}{t_\Lambda}\frac{1}{S_{-\Lambda}(t/t_\Lambda)}, \tag{3.83}$$

$$\phi(t) = \phi_{\rm p} \pm \sqrt{\frac{2}{3}} m_{\rm p} \log\left[\frac{t_\Lambda}{t_{\rm p}}\frac{2\,S_{-\Lambda}\left(t/t_\Lambda\right)}{1 + C_{-\Lambda}\left(t/t_\Lambda\right)}\right]. \tag{3.84}$$

Spatial curvature, $w = -1/3$

Spatial curvature is equivalent to a fluid with equation-of-state parameter $w = -1/3$ and density:

$$\rho_w = -3m_{\rm p}^2\frac{\kappa}{a^2}. \tag{3.85}$$

For $w = -1/3$, equation (3.63) is expressible in terms of trigonometric functions, and may be rearranged to express the scale factor a in terms of conformal time. Once $a(\eta)$ is obtained, the remaining equations (3.59), (3.60) and (3.61) can be used to express the rest of the variables in terms of η. Using our definitions of β (equation 3.70) and $\eta_{\rm p}$ (equation 3.74), and defining the new time scale,

$$\eta_\kappa = \frac{1}{2\sqrt{\kappa}}, \tag{3.86}$$

the solutions are:

$$a(\eta) = a_{\rm p}\left[\frac{S_\kappa\left(\eta/\eta_\kappa\right)}{\eta_{\rm p}/\eta_\kappa}\right]^{1/2}, \tag{3.87}$$

$$H(\eta) = \frac{1}{3t_{\rm p}}\frac{\left(\eta_{\rm p}/\eta_\kappa\right)^{3/2}}{T_\kappa(\eta/\eta_\kappa)\sqrt{S_\kappa(\eta/\eta_\kappa)}}, \tag{3.88}$$

$$\dot{\phi}(\eta) = \pm\sqrt{\frac{2}{3}}\frac{m_{\rm p}}{t_{\rm p}}\left[\frac{\eta_{\rm p}/\eta_\kappa}{S_\kappa(\eta/\eta_\kappa)}\right]^{3/2}, \tag{3.89}$$

$$\phi(\eta) = \phi_{\rm p} \pm \sqrt{\frac{3}{2}} m_{\rm p} \log\left[\frac{\eta_\kappa}{\eta_{\rm p}}\frac{2\,S_\kappa\left(\eta/\eta_\kappa\right)}{1 + C_\kappa\left(\eta/\eta_\kappa\right)}\right]. \tag{3.90}$$

Matter, $w = 0$

For matter with zero pressure, one has $w = 0$ and so:

$$\rho_w = \rho_{\rm p}^{\rm m}\left(\frac{a}{a_{\rm p}}\right)^{-3}, \tag{3.91}$$

where $\rho_{\rm p}^{\rm m}$ is an integration constant, labelling the energy density of matter at the epoch $a_{\rm p}$. For $w = 0$, equation (3.62) is expressible as an algebraic function, and may be rearranged to express the scale factor a in terms of coordinate time. Once $a(t)$ is obtained, the remaining equations (3.59), (3.60) & (3.61) can be used to express the rest of the variables in terms of t. Using our definition of β (equation 3.70), and defining the new time scale,

$$t_{\rm m} = \frac{4m_{\rm p}^2}{3t_{\rm p}\rho_{\rm p}^{\rm m}}, \tag{3.92}$$

the solutions are:

$$a(t) = a_{\rm p}\left(\frac{t}{t_{\rm p}}\right)^{1/3}\left(1+\frac{t}{t_{\rm m}}\right)^{1/3}, \tag{3.93}$$

$$H(t) = \frac{1+2\frac{t}{t_{\rm m}}}{3t\left(1+\frac{t}{t_{\rm m}}\right)}, \tag{3.94}$$

$$\dot{\phi}(t) = \pm\sqrt{\frac{2}{3}}m_{\rm p}\frac{1}{t\left(1+\frac{t}{t_{\rm m}}\right)}, \tag{3.95}$$

$$\phi(t) = \phi_{\rm p} \pm \sqrt{\frac{2}{3}}m_{\rm p}\log\left[\left(\frac{t}{t_{\rm p}}\right)\frac{1}{1+\frac{t}{t_{\rm m}}}\right]. \tag{3.96}$$

Radiation, $w = 1/3$

For radiation one has $w = 1/3$, and so:

$$\rho_w = \rho_{\rm p}^{\rm r}\left(\frac{a}{a_{\rm p}}\right)^{-4}, \tag{3.97}$$

where $\rho_{\rm p}^{\rm r}$ is an integration constant, labelling the energy density of matter at the epoch $a_{\rm p}$. For $w = 1/3$, equation (3.63) is expressible as an algebraic function, and may be rearranged to express the scale factor a in terms of coordinate time. Once $a(\eta)$ is obtained, the remaining equations (3.59), (3.60) and (3.61) can be used to express the rest of the variables in terms of η. Using our definition of β (3.70), and $\eta_{\rm p}$ (3.74), and defining the new time scale,

$$\eta_{\rm r} = \frac{3m_{\rm p}^2}{a_{\rm p}^2\eta_{\rm p}\rho_{\rm p}^{\rm m}}, \tag{3.98}$$

the solutions are:

$$a(\eta) = a_{\rm p}\left(\frac{\eta}{\eta_{\rm p}}\right)^{1/2}\left(1+\frac{\eta}{\eta_{\rm r}}\right)^{1/2}, \tag{3.99}$$

$$H(\eta) = \frac{1}{3t_{\rm p}}\left(\frac{\eta}{\eta_{\rm p}}\right)^{-3/2}\frac{1+2\frac{\eta}{\eta_{\rm r}}}{\left(1+\frac{\eta}{\eta_{\rm r}}\right)^{3/2}}, \tag{3.100}$$

$$\dot{\phi}(\eta) = \sqrt{\frac{2}{3}}\frac{m_{\rm p}}{t_{\rm p}}\left(\frac{\eta}{\eta_{\rm p}}\right)^{-3/2}\frac{1}{\left(1+\frac{\eta}{\eta_{\rm r}}\right)^{3/2}}, \tag{3.101}$$

$$\phi(\eta) = \phi_{\rm p} + \sqrt{\frac{3}{2}}m_{\rm p}\log\left[\left(\frac{\eta}{\eta_{\rm p}}\right)\frac{1}{\left(1+\frac{\eta}{\eta_{\rm r}}\right)}\right]. \tag{3.102}$$

3.4.4 The constants of integration

In the previous section, several constants arose, which we shall now review. For the system of equations (3.50)–(3.51), one would expect four constants of integration. The first is chosen by setting $a = 0$ at $t = 0$. For the case $\rho_w = 0$, the second and third are chosen by choosing a later time $t_{\rm p} > 0$ and fixing:

$$\phi(t_{\rm p}) = \phi_{\rm p},$$
$$a(t_{\rm p}) = a_{\rm p}.$$

We can also determine conformal time in this case, which involved defining a new constant $\eta_{\rm p}$ in terms of $a_{\rm p}$ and $\phi_{\rm p}$ via equation (3.74), and the solutions may be determined in conformal time. For the remaining cases, there is an additional integration constant determined by the value of ρ_w when the scale factor is $a_{\rm p}$. The solutions for these cases are then determined by matching them onto the case $\rho_w = 0$ at early times.

In addition, we defined a relevant time scale for each of the $\rho_w \neq 0$ cases: t_Λ, η_κ, $t_{\rm m}$, $\eta_{\rm r}$. These are expressed in terms of the previous integration constants in equations (3.80), (3.86), (3.92) and (3.98). If one chooses $t_{\rm p}$ or $\eta_{\rm p}$ to be much less than this second time scale, then the universe is in a fully kinetically dominated regime at $t_{\rm p}$, $\eta_{\rm p}$, with solutions very close to the $\rho_w = 0$ case. When η is of the order of the second time scale, then the effects of the universe's (dominant) additional component can be seen.

Although we have determined these equations in terms of three integration constants, there are in fact only two. The evolution equations (3.6) & (3.7) possess a two-parameter symmetry corresponding to a rescaling of a and t. More precisely, the form of the equations does not change under the transformation:

$$a \mapsto \alpha a, \qquad t \mapsto \sigma^{-1}t, \tag{3.103}$$

provided that ρ_i and the potential transform as:

$$\rho_i \mapsto \alpha^{3(1+w_i)}\sigma^2\rho_i, \qquad V(\phi) \mapsto \sigma^2 V(\phi). \tag{3.104}$$

This symmetry can be used effectively to remove some of the remaining integration constants. In practice this means that one may set $t_{\rm p} = 1$ to be the

Planck time and the scale factor $a(t_{\rm p}) = a_{\rm p} = 1$. Usually one takes the scale factor to be unity at the present epoch $a_0 \equiv a(t_0) = 1$, but this requirement is complicated by the uncertainties of reheating, so we do not follow that convention here. One may "physically" interpret $\phi_{\rm p}$ as the value of the field at $t = t_p$. This requires extrapolating the classical equations far beyond their validity, so it is more of a mnemonic aid than a physical interpretation. As we will see below, $\phi_{\rm p}$ controls the total number of e-folds of inflation.

3.5 Kinetic dominance in action

We shall now demonstrate the utility of kinetic initial conditions in the analysis of inflationary models. Even without integrating the evolution equations for $H(t)$ and $\phi(t)$, one sees that the basic scenario entails the universe emerging from an initial singularity at $t = 0$ in a regime where the kinetic energy of the inflaton dominates its potential energy along with any curvature or additional fluids. The evolution of $H(t)$ and $\phi(t)$ in this regime are given by (3.67) and (3.69) at sufficiently early times. $H(t)$ and $|\phi(t)|$ and their time derivatives decrease during this period of kinetic dominance, which concludes when there is approximate equipartition $\dot{\phi}^2 \sim V(\phi)$ between the kinetic and potential energies of the inflaton. This marks the onset of a (typically brief) period of fast-roll inflation (Linde, 2001), which must eventually become slow-roll inflation, with $\dot{\phi}^2 \ll V(\phi)$, since the latter is a generic attractor solution for inflation models (Belinsky et al., 1985). We will see that integration of the equations of motion in some illustrative cases does indeed verify these expectations.

For simplicity we shall work in the case with no other additional fields, $\rho_w = 0$, although our methods apply equally well to more complicated solutions. After discussing the validity of the initial conditions and numerical techniques, we shall consider two forms of potential: polynomial and exponential.

3.5.1 Initial conditions and scaling

For $\rho_w = 0$ the universe is spatially flat and contains only the inflaton field. The evolution equations then take the form:

$$H^2 = \frac{1}{3m_{\rm p}^2}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right), \tag{3.105}$$

$$0 = \ddot{\phi} + 3\dot{\phi}H + V'(\phi). \tag{3.106}$$

The general solution to the evolution equations has the asymptotic form given in equations (3.67) and (3.69). As discussed in Section 3.4.4, we may choose $t_{\rm p} = 1$ to be the Planck time. Given this, we set the initial conditions at an initial time $t_{\rm i}$ as:

$$\phi(t_{\rm i}) \equiv \phi_{\rm i} = \phi_{\rm p} - \sqrt{\frac{2}{3}} m_{\rm p} \log t_{\rm i}, \tag{3.107}$$

$$\dot{\phi}(t_{\rm i}) \equiv \dot{\phi}_{\rm i} = -\sqrt{\frac{2}{3}}\frac{m_{\rm p}}{t_{\rm i}}, \tag{3.108}$$

$$H(t_{\rm i}) \equiv H_{\rm i} = \frac{1}{3t_{\rm i}}. \tag{3.109}$$

There is a single constant of integration $\phi_{\rm p}$, which directly controls the number of e-folds during inflation. The number of e-folds N_* between the pivot scale k_* exiting the Hubble radius and the end of inflation is typically 50–60 (Planck Collaboration et al., 2013b). For the rest of this chapter $\phi_{\rm p}$ will be chosen so that the total number of e-folds $N_{\rm tot} = 65$. This will be discussed in greater detail in Section 3.5.5.

Throughout this chapter we work in the classical regime. In order for the conditions above to be valid, the initial conditions must be set at a time greater than the Planck time, $t_{\rm i} > t_{\rm p} = 1$, but within the kinetic dominated regime, for which $V(\phi_{\rm i}) \ll \dot{\phi}_{\rm i}^2$. Setting $t_{\rm i} = t_{\rm p} = 1$ in the above, one sees that kinetic dominance will endure beyond the Planck time provided:

$$V(\phi_{\rm p}) \ll m_{\rm p}^2. \tag{3.110}$$

The above requirement typically holds for potentials that give physically reasonable inflation models. For example, in the case of a free inflaton with mass m, one has

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

In order to generate the correct amplitude of curvature perturbations, the mass must be of the order $m \sim 10^{-5}m_{\rm p}$, whereas to generate the correct number of e-folds one requires $\phi_{\rm p} \sim \mathcal{O}(10)$, in which case $V(\phi_{\rm p}) \sim 10^{-8}m_{\rm p}^2$. Thus, there is no need to advocate trans-Planckian physics, since kinetic dominance lasts well beyond the Planck time, so one can set $t_{\rm i} \gg t_{\rm p}$.

We note that the evolution equations (3.105) & (3.106) are invariant under the simultaneous rescaling of the time coordinate, Hubble parameter and inflaton potential:

$$t \mapsto \sigma^{-1}t, \tag{3.111}$$

$$H \mapsto \sigma H, \tag{3.112}$$

$$V(\phi) \mapsto \sigma^2 V(\phi). \tag{3.113}$$

The advantage of this for numerical work is that a multiplicative scaling parameter from the potentials can be removed without loss of generality.

3.5.2 Polynomial potentials

We begin by analysing examples of polynomial potentials of the form:

$$V_n^{\rm pol}(\phi) = \mu^2\phi^n. \tag{3.114}$$

To obtain results we integrate the evolution equations (3.105) & (3.106) numerically. Our kinetic initial conditions are chosen using equation (3.107)–(3.109) with an initial time $t_{\rm i}$ small enough such that the inflaton is in the kinetic regime $V(\phi_{\rm i}) \ll \dot{\phi}_{\rm i}^{\,2}$. For the purposes of numerics the scaling parameter μ can be removed by rescaling the time coordinate (setting $\sigma = \mu^{-1}$). $\phi_{\rm p}$ is set by requiring that there be 55 e-folds during inflation.

The evolution of the Hubble parameter is shown in Figure 3.3 for polynomials with $n = 2, 4, 6$. It is helpful to define the variable:

$$\mathcal{K} \equiv \frac{\frac{1}{2}\dot{\phi}^2}{\rho} = \frac{\frac{1}{2}\dot{\phi}^2}{\frac{1}{2}\dot{\phi}^2 + V(\phi)}, \tag{3.115}$$

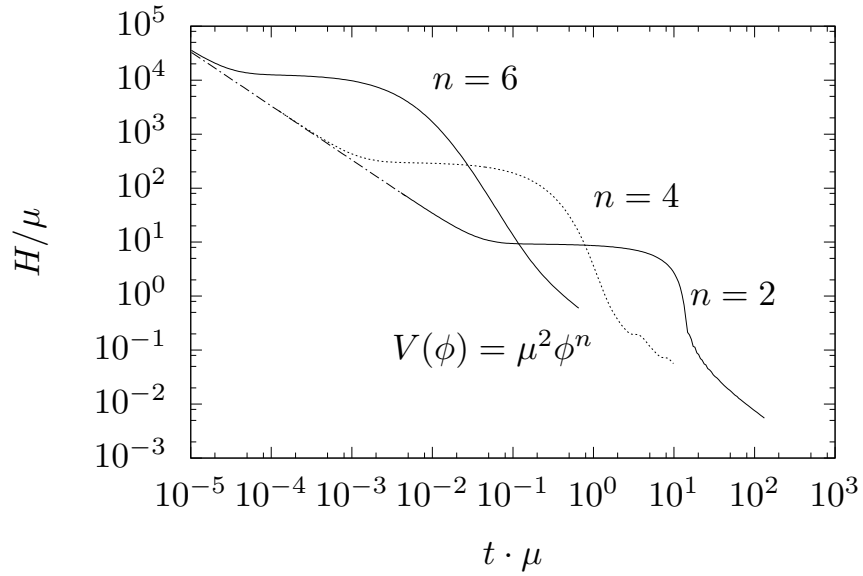


Figure 3.3: The evolution of the Hubble parameter for three polynomial potentials of the form in (3.114). The axes have been rescaled in terms of the parameter μ in the potential, so that this graph describes the evolution for any choice of μ. The initial conditions for inflation were set using the flat-universe kinetic conditions, and the parameter $\phi_{\rm p}$ was chosen so as to give 55 e-folds of inflation. All three universes emerge in a kinetically dominated phase with $H = 1/(3t)$, before entering a slow-roll inflationary phase with $H \sim$ constant. The universe then exits inflation, after which small 'wiggles' in H can be seen. These are due to the field ϕ executing oscillations about the base of the potential.

to be used as an investigative tool. $\mathcal{K}$ is the ratio of the kinetic energy to the total energy and has the properties that:

$$\mathcal{K}\begin{cases} \approx 1 & \Rightarrow \text{kinetic dominance} \\ > \frac{1}{3} & \Rightarrow \text{not inflating} \\ < \frac{1}{3} & \Rightarrow \text{fast-roll/power-law inflation} \\ \approx 0 & \Rightarrow \text{slow-roll inflation.} \end{cases} \tag{3.116}$$

This is used as a diagnostic tool in Figure 3.4 for the quadratic potential $V(\phi) = \mu^2\phi^2$. Examining Figure 3.4 one can see that our earlier expectations are verified. There are four stages of evolution:

A. the universe emerges from an initial singularity in a kinetically dominated phase,

B. it transitions through fast-roll inflation,

C. before entering a protracted slow-roll phase,

D. and thereafter the field ϕ quickly moves towards a minimum of the potential, about which it executes a decaying oscillation.

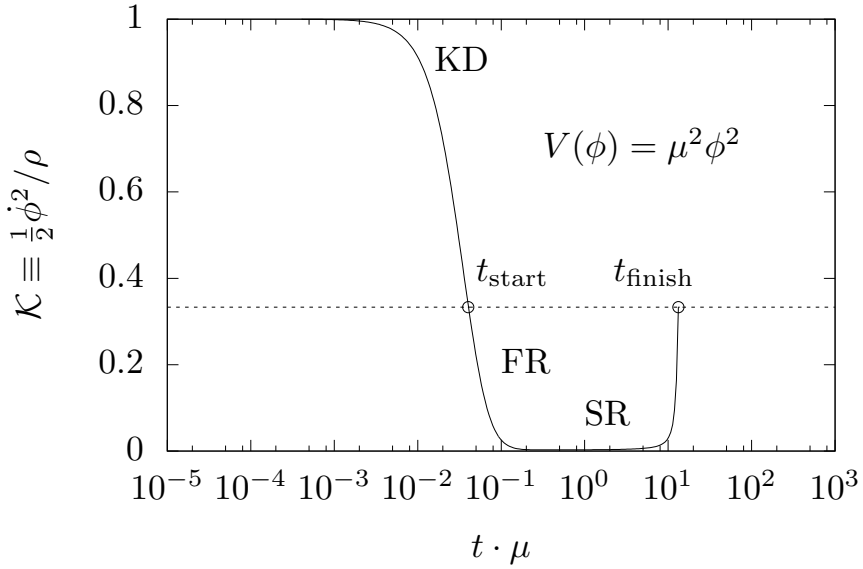


Figure 3.4: The evolution of $\mathcal{K} = \frac{1}{2}\dot{\phi}^2/\rho$ for the quadratic inflaton potential $V(\phi) = \mu^2\phi^2$. The universe can be seen to begin in a kinetically dominated state (KD). It enters inflation at $t_{\rm start}$. There is a brief fast-roll (FR) phase of inflation, before the universe enters a protracted slow-roll (SR) phase. The universe exits inflation at time $t_{\rm finish}$ after 55 e-folds. After the end of inflation the field ϕ executes oscillations about the base of the potential. This causes $\mathcal{K}$ to oscillate rapidly between 0 and 1. For clarity, the value of $\mathcal{K}(t)$ with $t > t_{\rm finish}$ has not been plotted. The above behaviour is common to all of the polynomial potentials shown in Figure 3.3.

The fast-roll transition in point (B) is potentially responsible for the damping of the CMB spectrum at low-ℓ observed in recent cosmological data. This will be discussed more fully in Section 3.5.5.

3.5.3 Exponential potentials

We now consider inflaton potentials of the form:

$$V_\epsilon^{\rm pow}(\phi) = 2V_0\left[\cosh\left(\frac{\sqrt{2\epsilon}}{m_{\rm p}}\phi\right) - 1\right], \tag{3.117}$$

which is a symmetrized form of the more common exponential potential; as $\phi \to \pm\infty$ the potential takes the asymptotic form:

$$V(\phi) = V_0 \exp\left(\frac{\sqrt{2\epsilon}}{m_{\rm p}}|\phi|\right). \tag{3.118}$$

Exponential potentials (3.118) have been well studied (Yokoyama and Maeda, 1988). For potentials of this form, the evolution equations have the analytical

power-law solutions:

$$a(t) \propto t^{1/\epsilon}, \tag{3.119}$$

$$\phi(t) = \pm m_{\rm p}\sqrt{\frac{2}{\epsilon}}\log\left(\sqrt{\frac{V_0}{(3-\epsilon)}}\frac{\epsilon}{m_{\rm p}}t\right), \tag{3.120}$$

$$H(t) = \frac{1}{\epsilon t}. \tag{3.121}$$

It is worth noting that for mathematical consistency one requires $\epsilon < 3$. For $\epsilon < 1$ these solutions are (continuously) inflating and are thus termed 'power-law inflation' (Lucchin and Matarrese, 1985). Moreover, these are attractor solutions as the universe evolves forwards in time. It is also straightforward to show that at all epochs $\dot{\phi}^2/V(\phi) = 2\epsilon/(3-\epsilon)$, so the ratio of the inflaton kinetic energy to its potential energy is constant. In particular, one notes that the solution is kinetically dominated only in the limit $\epsilon \to 3$. We may also interpret ϵ as the slow-roll parameter:

$$\epsilon \equiv \epsilon_H = -\frac{\dot{H}}{H^2}, \tag{3.122}$$

although one does not need to assume that it is small.

At first sight, solutions (3.119)–(3.121) appear to represent a counterexample to kinetic dominance as $t \to 0$, so it is worth exploring them further in the context of our proof of the generic nature of kinetic dominance in Section 3.3.

In terms of the master equation (3.31), in a flat universe, one finds $\phi = \sqrt{\frac{2}{3}}m_{\rm p}\psi + \text{const.}$ and thus for the exponential potential (3.118):

$$\frac{\mathrm{d}}{\mathrm{d}\psi}\log V_\epsilon^{\rm pow} = \frac{2\epsilon}{\sqrt{3}}. \tag{3.123}$$

Consequently, the master equation has the constant, finite solution:

$$f(\psi) = \log\left(1 - \frac{4}{3}\epsilon^2\right)^{-1/2}. \tag{3.124}$$

However, from the proof in Section 3.3, we know that this finite solution is unique. Any solution which is greater than this diverges as $|\phi| \to \infty$, and any solution less than this becomes negative. Indeed, this is already evident from the fact that the power-law solutions are attractors as the universe evolves forwards in time. By the same token, these solutions are unstable in the limit $t \to 0$, i.e. travelling backwards in time one diverges away from these solutions and generically arrives at kinetic dominance or a turn-around. We note that the proof that the power-law solutions are attractors is due to Halliwell (1987), and our work in Section 3.3 demonstrates that Halliwell's methodology is applicable more generally.

We show the evolution of the universe governed by the hyperbolic cosine inflaton potential (3.117) in Figures 3.5 & 3.6. The analysis is the same as that presented in Section 3.5.2. As in our previous example, the universe emerges from the initial singularity in kinetic dominance, which then transitions through a brief period of fast-roll inflation into a generically long-lasting power-law inflation state until the exit is reached as $\phi \to 0$, which corresponds to the minimum of the potential.

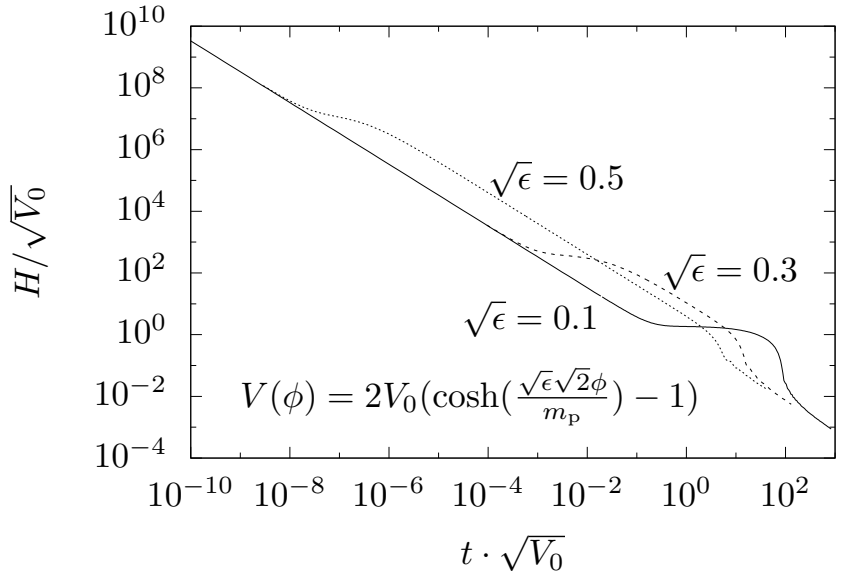


Figure 3.5: As in Figure 3.3, but for the hyperbolic cosine potential (3.117), which tends to an exponential potential for $\phi \to \pm\infty$. The scaling constant is now $\sqrt{V_0}$, and three values of $\sqrt{\epsilon}$ are considered. One can see clearly that the universe emerges in a kinetically dominated state with $H = 1/(3t)$. The universe then enters a protracted power-law phase where $H = 1/(\epsilon t)$. The field ϕ oscillates about the base of the potential after the exit of the inflationary phase, causing wiggles in the later sections of the above plot.

3.5.4 Another example of a finite solution $f(\psi)$

The $f(\psi)$ described above in equation (3.124) is one of the simplest examples of a finite solution. As shown earlier, there is at most one such finite solution for any given potential. For concreteness we demonstrate another less trivial example in this section.

By reverse-engineering the master equation (3.38), one can find a potential $V(\psi)$ for any specified $f(\psi)$. For example, if one chooses the oscillating solution (shown in Figure 3.7):

$$f(\psi) = \log\left(\frac{1}{\sqrt{1 - \left[a + b\cos(2k\psi)\right]^2}}\right), \tag{3.125}$$

then this $f(\psi)$ is the finite solution of the potential defined by:

$$V(\psi) = \left[1 - (a + b\cos 2k\psi)^2\right] e^{2a\psi + \frac{b}{k}\sin 2k\psi}, \tag{3.126}$$

whose shape is detailed in Figure 3.8.

To show explicitly that this is the only finite solution in this case, we consider a perturbed solution $y(\psi) = f(\psi) + \delta(\psi)$. From the uniqueness theorem (Appendix 3.A) if δ is initially positive (negative), then it is positive (negative)

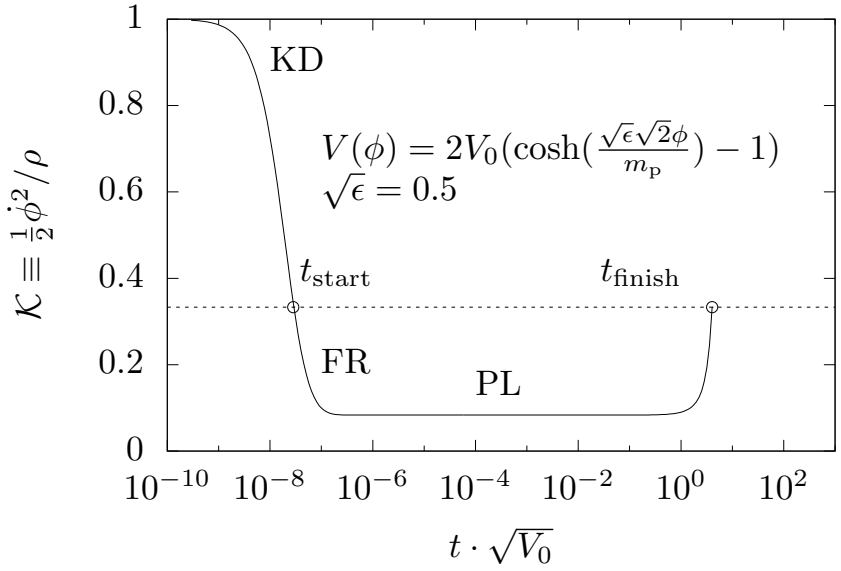


Figure 3.6: As in Figure 3.4, but for the hyperbolic cosine inflaton potential with $\sqrt{\epsilon} = 0.5$. The universe emerges in a kinetically dominated (KD) phase before going through transitory fast-roll (FR). Instead of a slow-roll inflation, the universe then settles into a power-law inflation (PL), characterized by $\mathcal{K} = \epsilon/3$. The universe exits inflation at t_finish, after which ϕ executes oscillations about the base of the potential. $\mathcal{K}$ therefore oscillates rapidly between 0 and 1, and the later $t > t_\mathrm{finish}$ section of this plot has been suppressed for clarity.

for all ψ. From the master equation (3.31) one may show that the perturbed solution satisfies:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\psi}\delta =& \sqrt{1 - e^{-2(f+\delta)}} - \sqrt{1 - e^{-2f}} \\ &> \sqrt{1 - e^{-2(f_\mathrm{max}+\delta)}} - \sqrt{1 - e^{-2f_\mathrm{max}}},\end{aligned} \tag{3.127}$$

where $f_\mathrm{max} = \log(1/\sqrt{1 - (a-b)^2})$. Substituting this in, one finds:

$$\frac{\mathrm{d}}{\mathrm{d}\psi}\delta > \sqrt{1 - e^{-\delta}(1 - (a-b)^2)} - \sqrt{1 - (1 - (a-b)^2)}. \tag{3.128}$$

One can see that if $\delta > 0$, then the right-hand side is strictly greater than some number greater than zero, hence δ grows without bound, and any solution greater than f initially diverges.

Working from equation (3.127), one also finds:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\psi}\delta <& \sqrt{1 - e^{-2(f_\mathrm{min}+\delta)}} - \sqrt{1 - e^{-2f_\mathrm{min}}} \\ &< \sqrt{1 - e^{-\delta}(1 - (a+b)^2)} - \sqrt{1 - (1 - (a+b)^2)}.\end{aligned} \tag{3.129}$$

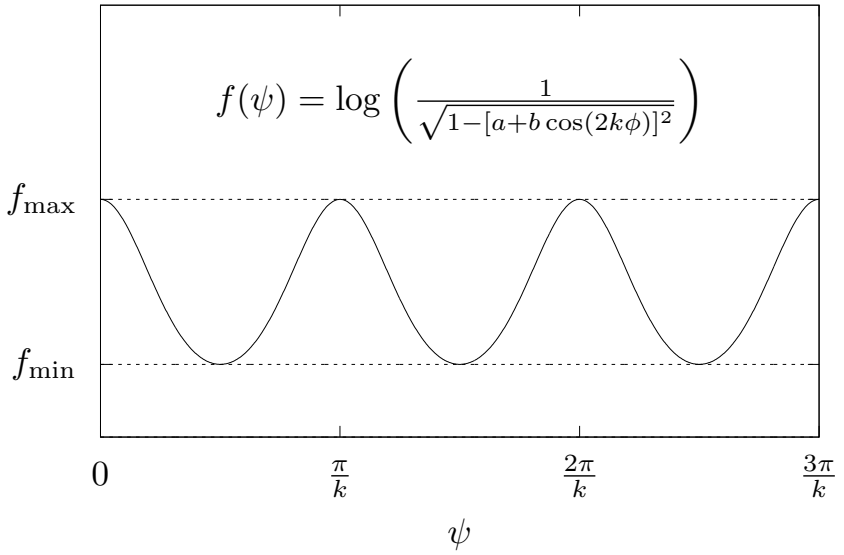


Figure 3.7: An oscillating finite solution to the master equation (3.31) with $V(\psi)$ defined as in equation (3.126) and demonstrated in Figure 3.8. If a and b are chosen such that $a > b > 0$, $0 < a + b < 1$, $0 < a - b < 1$, then this represents a finite, positive solution. From equation (3.125) it is easy to see that $f_{\min} = \log\left(\frac{1}{\sqrt{1-(a+b)^2}}\right)$, $f_{\max} = \log\left(\frac{1}{\sqrt{1-(a-b)^2}}\right)$.

One can see that if $\delta < 0$, then the right-hand side is strictly less than some number less than zero, hence δ falls without bound, and any solution less than f eventually becomes negative.

Thus one finds that $f(\psi)$ is the only finite positive solution. All other solutions either become negative or diverge.

3.5.5 Power spectrum of the curvature perturbation

The most interesting aspect of kinetic dominance is seen when the power spectrum of scalar curvature perturbations is examined. Recent observations of CMB power spectra (Hinshaw et al., 2013, Planck Collaboration et al., 2013a) show an unexpected suppression at low multipoles. While these deviations are not large enough to cause us to discard the standard Λ cold dark matter (ΛCDM) cosmology (Berera et al., 1998, Berera and Heavens, 2000, de Oliveira-Costa et al., 2004), there is still enough tension to be worthy of investigation. As we show below, kinetic dominance predicts a generic cutoff in the curvature power spectrum at large spatial scales. This is precisely what is needed to suppress low multipole moments of the CMB power spectrum while retaining the quality of the fit at higher ℓ values.

Figure 3.9 has been taken from Hlozek et al. (2012) and shows the current status of the observational constraints on the late-time matter power spectrum,

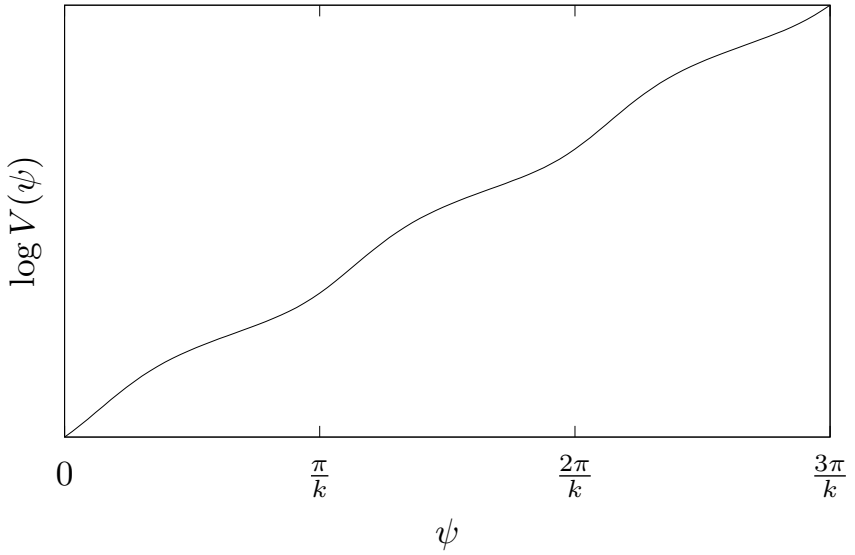


Figure 3.8: The master equation may be solved for the potential $V(\psi)$ if a form of $f(\psi)$ is chosen. For the choice of $f(\psi)$ detailed in equation (3.125), the potential is solvable in closed form (equation 3.126).

given by:

$$P(k, z = 0) = 2\pi^2 k \mathcal{P}_{\mathcal{R}}(k) G^2(z) T^2(k), \tag{3.130}$$

where $G(z)$ gives the growth of matter perturbations, $T(k)$ is the matter transfer function and $\mathcal{P}_{\mathcal{R}}(k)$ is the primordial curvature perturbation power spectrum. This mapping enables one to combine constraints on the power spectrum from CMB and other probes at $z \approx 0$.

As shown by Liddle and Lyth (2000), the primordial curvature perturbation power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ is given approximately by:

$$\mathcal{P}_{\mathcal{R}}(k) = \left(\frac{H^2}{2\pi\dot{\phi}} \right)^2_{k=aH}, \tag{3.131}$$

where, as denoted, the right-hand side is evaluated when a given scale crosses the horizon. If one has numerically calculated $a(t)$, $H(t)$ and $\dot{\phi}(t)$, then plotting $\left(H^2/2\pi\dot{\phi}\right)^2$ against aH will give the shape of the spectrum.

In order to perform predictive calculations, one must calibrate the aH axis to an observable scale today. This is easy to do if one defines a comoving pivot scale k_*, which leaves the horizon (at a time t_*) when N_* e-folds of inflation remain. In general, the relation between k_* and N_* depends on both the potential $V(\phi)$ and the details of cosmic reheating. For most reasonable models, $50 < N_* < 60$ for k_* with a value of $0.05\,\mathrm{Mpc}^{-1}$ today (Planck Collaboration et al., 2013b). For this work, we will take $N_* = 55$.

Once a value for N_* is chosen, one can determine numerically the time t_* at which N_* e-folds of inflation remain as well as $a_* \equiv a(t_*)$ and $H_* \equiv H(t_*)$.

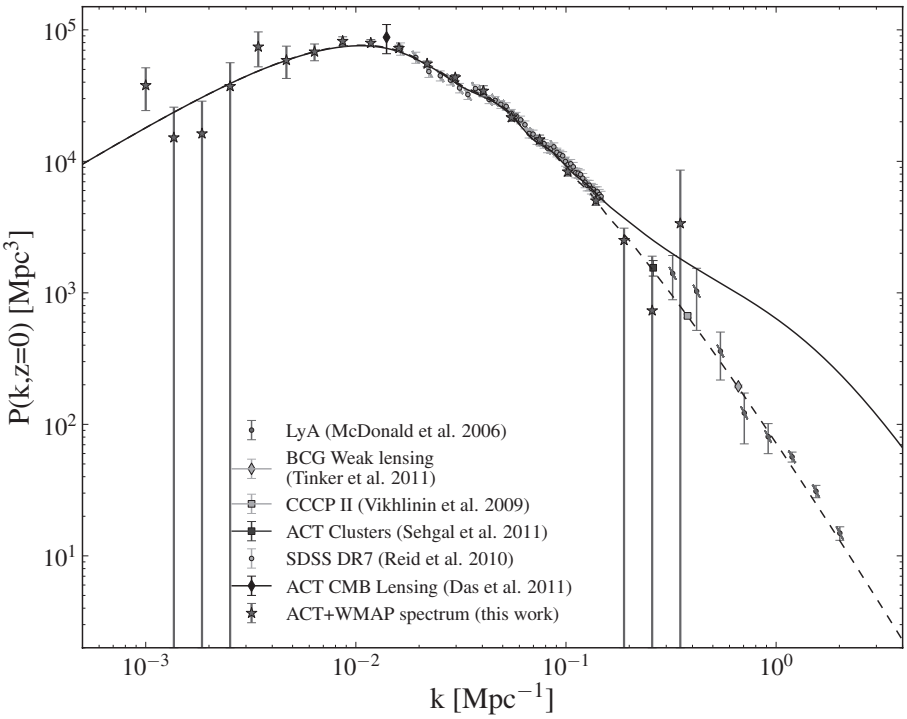


Figure 3.9: Figure taken from Hlozek et al. (2012) showing the current status of the observed late-time matter perturbation power spectrum. As one can see, the currently probed k range is $10^{-3} < k < 2\,\mathrm{Mpc}^{-1}$.

Since we know that the value of aH at t_* corresponds to a wave number today of $0.05\,\mathrm{Mpc}^{-1}$, we may calibrate the aH axis of the plot of the power spectrum using:

$$k_{\mathrm{today}} = \frac{aH}{a_* H_*} \times 0.05\,\mathrm{Mpc}^{-1}. \tag{3.132}$$

Calibrated plots of $\mathcal{P}_{\mathcal{R}}(k)$ are found in Figures 3.10 and 3.11.

The shape of the primordial power spectra obtained corresponds to that found in Lasenby and Doran (2005). We see that, in general, $\mathcal{P}_{\mathcal{R}}(k)$ has less power at low and high k values than would be expected from a canonical power-law primordial spectrum. The low-k cutoff is entirely generic and occurs as a result of the brief period of fast-roll prior to slow-roll or power-law inflation. This effect has been discussed previously by Boyanovsky et al. (2006a). The fast-roll regime behaves like an attractor potential in the wave equations for the mode functions of curvature and tensor perturbations. This potential leads then to the suppression of the primordial power spectra at low k. Hence, it might be able to account for the suppression of the quadrupole of the CMB in agreement with observational data, as discussed by Boyanovsky et al. (2006b). The exact position of the low-k cutoff is determined by the value of ϕ_{p}, as it controls the total number of e-folds of inflation. This effect has also been discussed in the context of "Open Inflation" (Yamauchi et al., 2011, Linde et al., 1999, Linde, 1999), and examined using WMAP data by Contaldi et al.

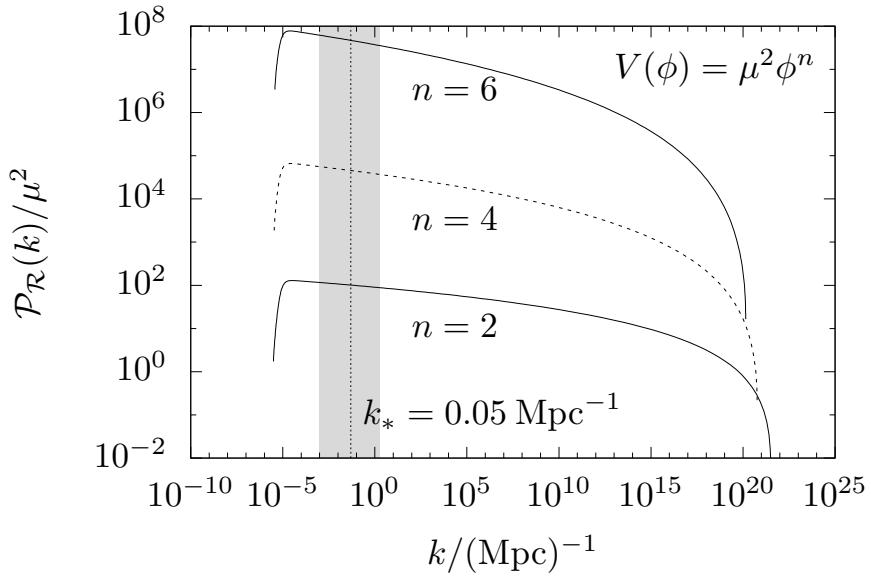


Figure 3.10: The approximate power spectrum of the primordial curvature perturbations for polynomial potentials, calculated using equation (3.131). The units of the k axis are determined by the requirement that there are $N_* = 55$ e-folds remaining when the pivot scale $k_* = 0.05\ \mathrm{Mpc}^{-1}$ exits the horizon $k = aH$. The magnitude of the power spectrum is determined by the scaling μ in the potential, and the low-k cutoff is determined by a choice of ϕ_p such that there are $N_\mathrm{tot} = 65$ e-folds of total inflation. The grey area indicates the angular scales that have been experimentally probed; see Figure 3.9.

(2003).

Further inspection shows $\mathcal{P}_\mathcal{R} \sim \log k$ after the low-k cutoff. This is identical to the result found by Lasenby and Doran (2005) and in contrast with the standard power-spectrum parametrization which assumes a near-flat power-law scaling: $\log \mathcal{P}_\mathcal{R} \sim \log k$.

It should be observed that we are using the approximation (3.131) outside the slow-roll regime for which it is valid; nonetheless we have performed full calculations that do not use the above approximation and which indicate that the resulting power spectrum is, in fact, a good representation of the true spectrum. These approximate spectra demonstrate the key generic aspects of the accurate calculation: both exhibit a low-k cut off and that $\mathcal{P}_\mathcal{R}(k) \sim \log(k)$. We shall follow this work with a second publication containing the full details and discussion of the accurate calculation, but a representative example is shown in Figure 3.12. It should be noted that such accurate calculations depend strongly on how one chooses initial conditions in the kinetically dominated phase for the comoving curvature perturbation. This is discussed further in Chapter 6. An alternative but related accurate calculation has been performed by Lello and Boyanovsky (2013), which uses kinetic initial conditions to show that the suppression at low-ℓ is entirely generic. It should also be noted that methods which reconstruct the primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ (Vázquez

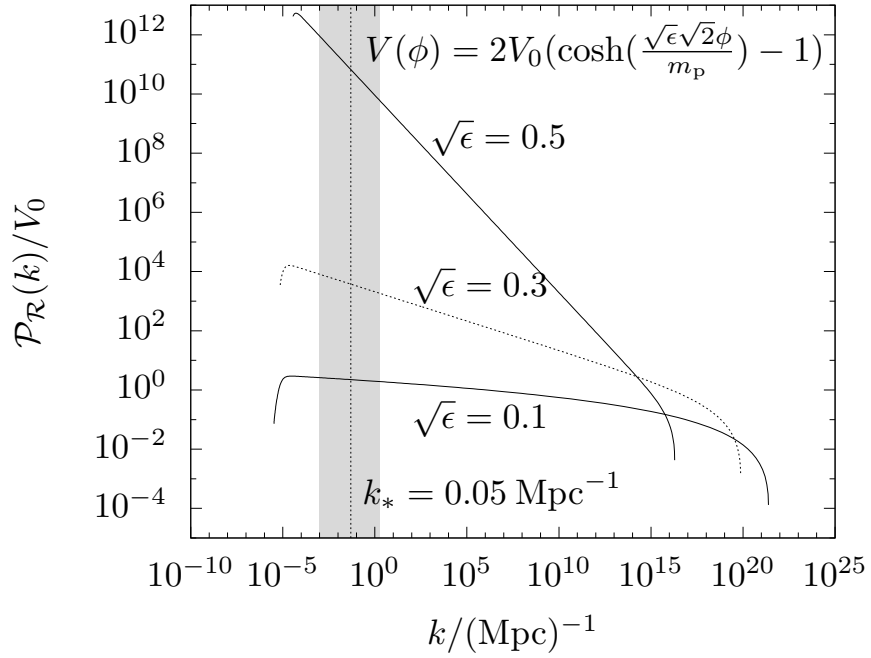


Figure 3.11: As in Figure 3.10, but for exponential potentials. The magnitude of the power spectrum now scales with V_0 rather than μ^2.

et al., 2012b, Hazra et al., 2013) using data also show a dip at low k values, updated results for which are reported in Chapter 5.

We demonstrate the suppression on large angular scales of the CMB and late-time matter power spectra qualitatively in Figure 3.13. In the standard six-parameter ΛCDM cosmology, the primordial power spectrum has a power-law form, parametrized by two variables A_s and $n_{\rm s}$, such that:

$$\mathcal{P}_\mathcal{R}(k) = A_s\left(\frac{k}{k_*}\right)^{n_{\rm s}-1}. \tag{3.133}$$

Using the best-fit parameters from *Planck*+WP+highL+BAO (Planck Collaboration et al., 2013a) yields the standard matter and CMB power spectra (dashed lines), whereas the alternative power spectra (solid line) were generated using $\mathcal{P}_\mathcal{R}(k)$ from Figure 3.10 (for $n = 2$), for which the axes were rescaled to agree with the values of A_s and $n_{\rm s}$ during the slow-roll phase. The resulting matter and CMB power spectra are seen to exhibit a suppression of power at low k and ℓ values, with the rest of the spectra perfectly intact, as required by cosmological observations. Further investigation is clearly required, but this analysis already demonstrates the utility of kinetic dominance.

The results of this section agree with the "just enough inflation" scenario introduced by Ramirez and Schwarz (2009, 2012), Ramirez (2012). In their work they assume that there is some physical mechanism which would limit the potential from above such that $V(\phi) < M^4_{\rm GUT}$, and thus find that $V(\phi) \ll \dot{\phi}^2$

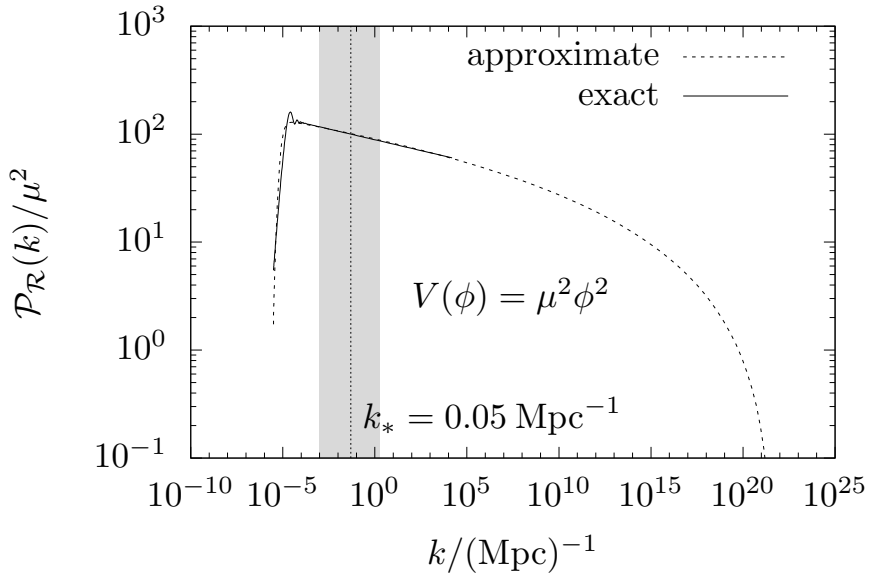


Figure 3.12: As in Figure 3.10, but now comparing the approximate and exact calculations of the primordial power spectrum for an $n = 2$ polynomial potential. The key feature of a low k suppression remains, with an additional ringing effect. The details of this ringing effect depend strongly on how one chooses initial conditions for the comoving curvature perturbation in the kinetically dominated phase. For this calculation, we chose "naïve" Bunch Davies initial conditions. For further detail, see Chapter 6.

at early times. Our results therefore place their observations in a more generic setting.

3.5.6 Comparison with equipartition initial conditions

An alternative method for setting the initial conditions for inflation models has been proposed by Boyanovsky et al. (2006a). Their work also shows that the low-multipole suppression of the scalar power spectrum is the result of a brief period of fast-roll inflation prior to the standard slow-roll regime. They arrange for such a fast-roll period by assuming "equipartition initial conditions". In this approach, the initial conditions are set at a time $t = t_{\rm eq}$, when there is approximate equipartition between the kinetic and potential energy of the inflaton:

$$\frac{1}{2}\dot{\phi}_{\rm eq}^2 \sim V(\phi_{\rm eq}). \tag{3.134}$$

Almost by definition, the subsequent evolution will generically exhibit a (brief) period of fast-roll inflation, before entering a slow-roll phase (which is an attractor solution). Boyanovsky et al. (2006a) verified this generic behaviour for a wide range of chaotic and new inflation potentials.

To address these issues further, we consider in more detail the main inflationary model used by Boyanovsky et al. (2006a) to illustrate their generic

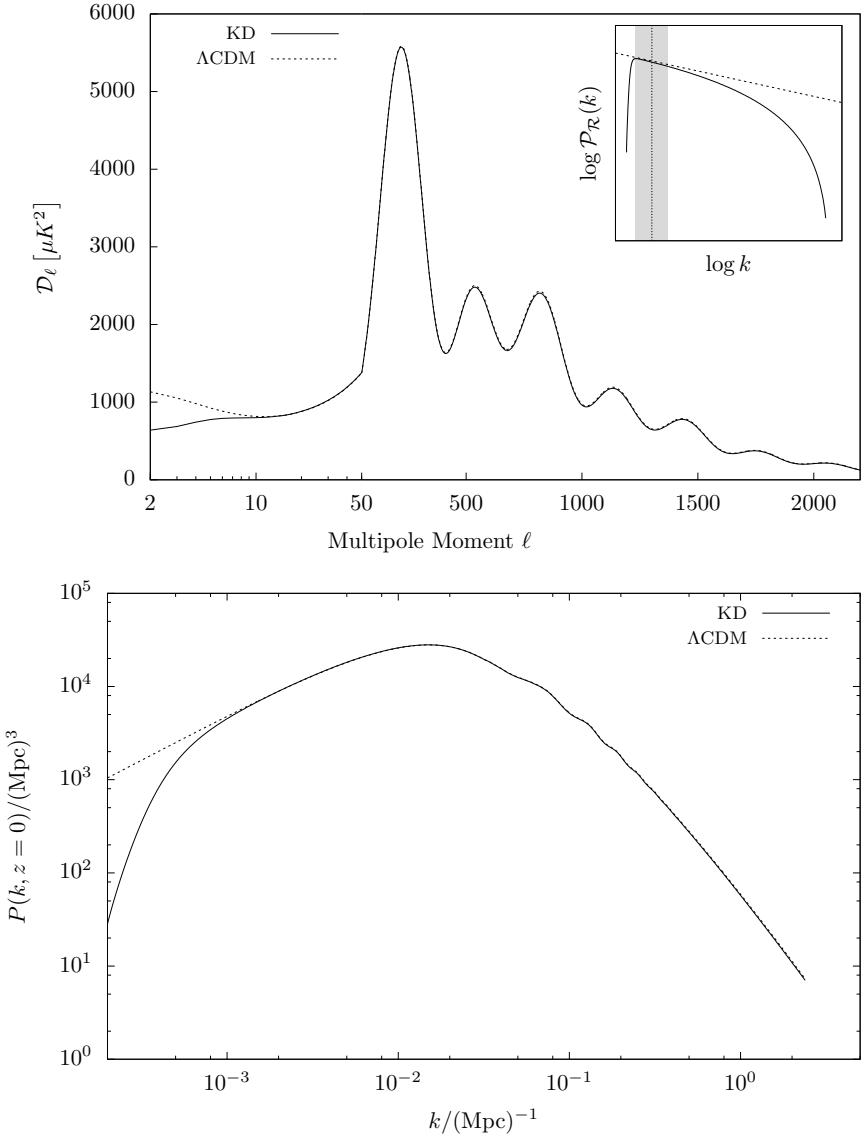


Figure 3.13: The CMB scalar power spectra (top) and the late-time matter power spectra (bottom), resulting from the curvature perturbation power spectra in the inset of the top figure. The solid line corresponds to a free inflaton with potential $V = \frac{1}{2}m^2\phi^2$ assuming kinetic initial conditions with $m = 0.81 \times 10^{-5} m_\mathrm{p}$, $\phi_\mathrm{p} = 21.8$, $N_* = 42.5$. The dashed line corresponds to the best-fit standard ΛCDM model. The presence of the cutoff in the curvature power spectrum for kinetic initial conditions causes a suppression of power on large angular scales in both the CMB and matter power spectra.

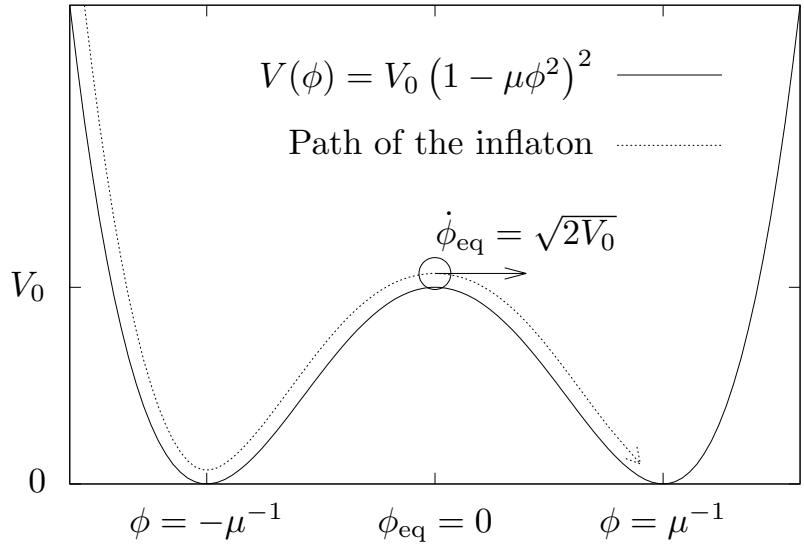


Figure 3.14: An illustration of the "equipartition" initial conditions imposed by Boyanovsky et al. (2006a), for the new inflation potential $V_0\big(1-\mu^2\phi^2\big)^2$. The initial conditions are set at some time $t_{\rm eq}$ with values denoted by a subscript "eq". The field is set with $\phi_{\rm eq} = 0$ and a velocity $\dot{\phi}_{\rm eq} = \sqrt{2V_0}$ so that the energy is partitioned equally between kinetic and potential energy: $V(\phi_{\rm eq}) = \frac{1}{2}\dot{\phi}_{\rm eq}$.

findings. The model is spatially flat, and uses a "new inflation" potential:

$$V(\phi) = V_0\big(1-\mu\phi^2\big)^2. \tag{3.135}$$

The initial conditions are set by choosing $\phi(t_{\rm eq}) = \phi_{\rm eq} = 0$ with an equipartition of kinetic and potential energies so that $\dot{\phi}(t_{\rm eq}) = \dot{\phi}_{\rm eq} = \sqrt{2V_0}$. Figure 3.14 illustrates this diagrammatically. The values of V_0 and μ are tuned so that the power spectrum has an appropriate index n_s and the correct number of e-folds are generated.

We now interpret this methodology using our formalism. Kinetic initial conditions will inevitably lead to a time $t_{\rm eq}$ when the equipartition condition (3.134) is satisfied. Thus, rather than considering (3.134) as a fundamental physical principle, it should be regarded as a natural consequence of kinetic initial conditions. Moreover, this is true independently of any tuning that we apply to μ or V_0 and indeed of the potential. Moreover, demanding that (3.134) is satisfied at $\phi = 0$ corresponds to choosing a specific value of $\phi_{\rm p}$ such that the field arrives at $\phi = 0$ with an equipartition of kinetic and potential energy (This is illustrated in Figure 3.14). If, however, $\phi_{\rm p}$ is required to lie within a certain range of values in order to produce an appropriate number of e-folds of inflation, then the choice $\phi_{\rm eq} = 0$ may be inconsistent with kinetic dominance. This can only be resolved if one is free to choose a general point $\phi_{\rm eq}$ for the position of equipartition.

3.6 When is kinetic dominance not the case?

Having looked in detail at the consequences of kinetic dominance, we enumerate the instances in which it does not hold. If kinetic dominance is not the case, then we can conclude that either

A. as one moves backward in time $\dot{\phi} \to 0$, and the inflaton tends to a constant value; or

B. there is no epoch before which we can say $\dot{\phi} \neq 0$, and the inflaton continues to oscillate as one moves backward in time.

We shall examine each of these cases in turn.

3.6.1 Resting inflaton: $\dot{\phi} \to 0$

No auxiliary fluids: eternal de Sitter

We shall begin by considering the case with no auxiliary fluids, for which the Friedmann (3.7) and Klein-Gordon (3.9) equations take the form:

$$H^2 = \frac{1}{3m_{\mathrm{p}}^2}\left(\tfrac{1}{2}\dot{\phi}^2 + V(\phi)\right), \tag{3.136}$$

$$0 = \ddot{\phi} + 3\dot{\phi}H + V'(\phi). \tag{3.137}$$

If $\dot{\phi} \to 0$ as one moves backward in time, then the inflaton ϕ and field $V(\phi)$ tend to constant values ϕ_0 and $V(\phi_0) \equiv V_0$. By examining the first of the above equations, one can see that H tends to a constant value:

$$H \to H_0 \equiv \sqrt{\frac{V_0}{3m_{\mathrm{p}}^2}}. \tag{3.138}$$

Thus, such a universe exhibits an eternal de Sitter phase (edS) as $t \to -\infty$.

By examining the Klein-Gordon equation (3.137), one can see that the only nonzero term remaining is $\frac{\mathrm{d}}{\mathrm{d}\phi}V(\phi)$. From this one can conclude that the inflaton must come to rest on an extremum of the potential. It is straightforward to show that the dynamical equations then have the following asymptotic solution as $t \to -\infty$:

$$\phi(t) = \phi_0 \pm A\exp(\alpha t), \tag{3.139}$$

$$H(t) = \sqrt{\frac{V(\phi_0)}{3m_{\mathrm{p}}^2}} \equiv H_0, \tag{3.140}$$

$$a(t) = Be^{H_0 t}, \tag{3.141}$$

where A and B are arbitrary constants, and α is a real, positive solution to the quadratic equation:

$$\alpha^2 + 3H_0\alpha + \left.\frac{d^2V}{d\phi^2}\right|_{\phi=\phi_0} = 0. \tag{3.142}$$

This solution is discussed in more detail by Destri et al. (2010). An example of this is "Hilltop inflation" (Linde, 1982, Albrecht and Steinhardt, 1982).

It should be noted that these solutions are not generic, as these solutions are rolling away from a position of unstable equilibrium: going backward in time, any small perturbation causes the inflaton to overshoot the extremum and move on to a kinetically dominated phase.

One can demonstrate the above statement formally by considering equations (3.137) and (3.136) in the Hamilton-Jacobi representation:

$$\left(\frac{\mathrm{d}H}{\mathrm{d}\phi}\right)^2 = \frac{3H^2}{2m_{\mathrm{p}}^2} - \frac{V(\phi)}{2m_{\mathrm{p}}^4}, \tag{3.143}$$

$$\frac{\mathrm{d}H}{\mathrm{d}\phi} = -\frac{\dot{\phi}}{2m_{\mathrm{p}}^2}. \tag{3.144}$$

The Hamilton-Jacobi representation is valid in the periods in which $\phi(t)$ is monotonic; i.e. $\dot{\phi}$ does not change sign. If one assumes that $\dot{\phi} > 0$, then the first of the above equations reads:

$$\frac{\mathrm{d}H}{\mathrm{d}\phi} = -\sqrt{\frac{3H^2}{2m_{\mathrm{p}}^2} - \frac{V(\phi)}{2m_{\mathrm{p}}^4}}. \tag{3.145}$$

Solutions to this equation are plotted in Figure 3.15, which demonstrates the following facts:

- There is a region in which solutions cannot exist, since by the Friedmann equation (3.136) we find that $H > \sqrt{V(\phi)/3m_{\mathrm{p}}^2}$.
- By equation (3.145) that solutions meet this region with zero gradient.
- Outside of this region, the right-hand side of equation (3.145) is Lipschitz continuous (see Appendix 3.A, or Agarwal et al., 1993), hence solutions do not cross over in the white region of the graph.

Consider the eternal de Sitter solution $H_{\mathrm{edS}}(\phi)$, plotted as the right-hand half of the dotted line in Figure 3.15. This solution has the property that as $\phi \to \phi_0$, $\dot{\phi} \propto \frac{\mathrm{d}H}{\mathrm{d}\phi} \to 0$. Consider also a solution H_{h} which at some value ϕ_1 is greater than the edS solution, $H_{\mathrm{h}}(\phi_1) \geq H_{\mathrm{edS}}(\phi_1)$. By uniqueness, this will remain greater than H_{edS} within the white region of the graph. Both H_{edS} and H_{h} satisfy equation (3.145). Taking the difference of these two equations and integrating from ϕ_1 to ϕ_0 shows:

$$\begin{aligned}
&H_{\mathrm{h}}(\phi_0) - H_{\mathrm{edS}}(\phi_0) = \\
&H_{\mathrm{h}}(\phi_1) - H_{\mathrm{edS}}(\phi_1) \\
&+ \int_{\phi_0}^{\phi_1} \sqrt{\frac{3H_{\mathrm{h}}^2}{2m_{\mathrm{p}}^2} - \frac{V(\phi)}{2m_{\mathrm{p}}^4}} - \sqrt{\frac{3H_{\mathrm{edS}}^2}{2m_{\mathrm{p}}^2} - \frac{V(\phi)}{2m_{\mathrm{p}}^4}}\, d\phi \\
&> H_{\mathrm{h}}(\phi_1) - H_{\mathrm{edS}}(\phi_1) > 0.
\end{aligned} \tag{3.146}$$

We thus find $\mathrm{d}H_{\mathrm{h}}/\mathrm{d}\phi \neq 0$ at $\phi = \phi_0$, and thus does not represent an eternal de Sitter phase. A similar argument holds for solutions $H_{\mathrm{l}}(\phi)$ which start out less than H_{edS}, only these must collide with the dark region. Upon this collision, $\dot{\phi}$ changes sign, causing the diagram to reflect about the $\phi = 0$ axis.

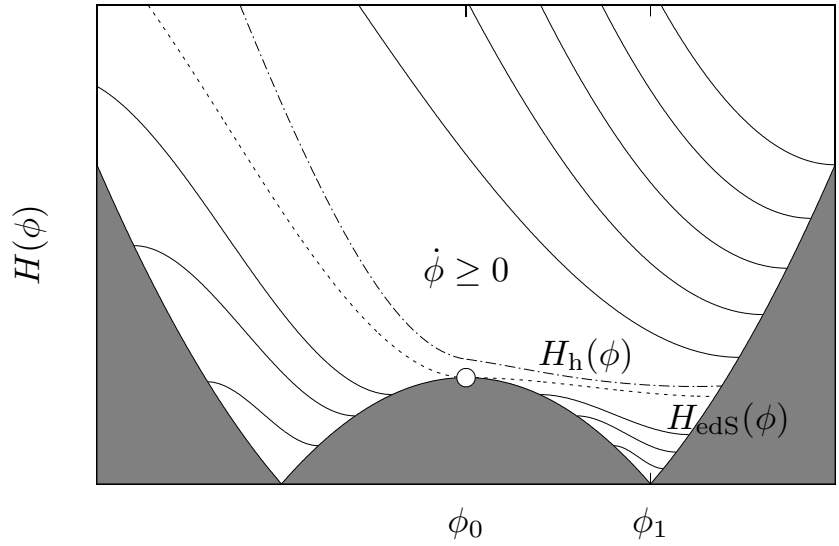


Figure 3.15: Schematic of the Hubble parameter against ϕ in the Hamilton-Jacobi representation. The potential $V(\phi)$ is the same potential as in Figure 3.14. The solid curves represent the portions of the solutions to (3.145) with $\dot{\phi} \geq 0$. The shaded region is defined by $H < \sqrt{V/3m_{\rm p}^2}$. Solutions meet the shaded region with zero gradient. Solutions within the white region are unique: they do not cross over. The right-hand side of the dashed curve represents a universe entering at $\phi = \phi_0$, ($t = -\infty$) in an eternal de Sitter phase. The left-hand side of the dashed curve indicates a universe entering in a kinetically dominated phase before settling into a de Sitter phase at $t = +\infty$.

Figure 3.15 reveals an interesting second eternal de Sitter solution. The right-hand half of the dashed line indicates the previously discussed phase: a universe which emerges at $t = -\infty$ in an inflating state with the inflaton slowly rolling off an extremum in the potential, exiting the inflationary phase when the inflaton oscillates about the bottom of the potential. The left-hand half of the dashed line indicates a universe which emerges at $t = 0$ in a kinetically dominated phase before the inflaton rests on top of the extremum, settling into an eternal de Sitter phase as $t \to \infty$.

Auxiliary fluids

The presence of an auxiliary fluid makes it slightly easier to engineer a solution in which $\dot{\phi} \to 0$, as one does not require that the inflaton comes to rest on an extremum of the potential.

In the limit that $a \to 0$, the ρ_i with the largest w_i dominates over all of the others. The relevant equations are therefore the Friedmann (3.7) and

Klein-Gordon (3.9) equations, with a single fluid ρ_w:

$$H^2 = \frac{1}{3m_{\rm p}^2}\left(\tfrac{1}{2}\dot{\phi}^2 + V(\phi) + \rho_w\right), \tag{3.147}$$

$$0 = \ddot{\phi} + 3\dot{\phi}H + V'(\phi). \tag{3.148}$$

In the first of these, if $\phi \to \phi_0$, then as $a \to 0$ on the right-hand side one only need worry about ρ_w. Since $H = \frac{\mathrm{d}}{\mathrm{d}t}\log a$, and $\rho_w \propto a^{-3(1+w)}$, one finds:

$$\int \frac{\mathrm{d}\log a}{\sqrt{\rho_w}} = \int \frac{t}{\sqrt{3}m_{\rm p}}, \tag{3.149}$$

$$\Rightarrow \qquad \rho_w = \frac{4m_{\rm p}^2}{3(1+w)^2}\frac{1}{t^2}. \tag{3.150}$$

We may then use the above equation along with (3.147) to show that:

$$H = \frac{2}{3(1+w)t} \quad \Rightarrow \quad a \propto t^{2/3(1+w)}. \tag{3.151}$$

We can now solve (3.148) for $\phi(t)$. As $\phi \to \phi_0$, we may assume the potential term tends to a constant value ${V_0}'$. Selecting the solution which tends to a constant, one finds:

$$\phi(t) = \phi_0 - \frac{{V_0}'(w+1)}{2(w+3)}t^2. \tag{3.152}$$

This constitutes a solution which is present for any potential with auxiliary fluids. In general one can find a specific solution with $\phi \to \phi_0$ for any given ϕ_0. However, for any small perturbation from this solution, one arrives back at kinetic dominance. The kinetically dominated solutions are generic.

3.6.2 Pathological oscillations: $\dot{\phi} = 0$

We now turn to the case where there is no epoch prior to which ϕ is monotonic. The inflaton continues to oscillate endlessly as $a \to 0$.

It is easy to engineer potentials that lead to this behaviour. In the case with no auxiliary fields, the limiting forms of $\dot{\phi}(t)$ and $\phi(t)$ are given by equations (3.68) and (3.69). We may solve these to find the relationship between $\dot{\phi}$ and ϕ:

$$\dot{\phi}^2 = \exp\left(\frac{\sqrt{6}}{m_{\rm p}}|\phi - \phi_{\rm p}|\right). \tag{3.153}$$

If one chooses a potential that grows faster than the right-hand side of this equation, then the universe cannot be kinetically dominated. We are thus forced to the conclusion that in such a universe either $\dot{\phi} \to 0$ or there is no epoch before which $\dot{\phi} \neq 0$. Typically, if one examines the numerical solutions of such equations, one sees that the inflaton oscillates at a faster and faster rate, with greater and greater amplitude until the numerical limit of the solver is reached.

These solutions are therefore somewhat pathological, though for the cases where they occur, kinetic dominance is not the generic solution.

3.7 Conclusions

We have shown that, if quantum gravitational effects are ignored, the coupled evolution equations for the inflaton field $\phi(t)$ and the Hubble parameter $H(t)$ in generic homogeneous and isotropic single-field inflation models imply that a universe beginning with a steadily moving inflaton ($|\dot{\phi}| > \xi > 0$ as $a \to 0$, for some positive constant ξ) generically emerges from an initial singularity in a non-inflating, kinetically dominated state ($\dot{\phi}^2 \gg V(\phi)$). In this kinetic-dominated regime, one obtains simple analytical solutions for $\phi(t)$ and $H(t)$, which are independent of the form of the inflaton potential $V(\phi)$ and of the presence of auxiliary fluids such as matter, radiation, dark energy or spatial curvature. These solutions provide a simple means of setting the initial conditions for such inflation models, from which numerical integration of the evolution equations may proceed.

For illustration, we applied this "kinetic" procedure for setting initial conditions to spatially flat polynomial and exponential inflation models. By making an appropriate choice of the time $t_{\rm i}$ at which the initial conditions are set, and the single free parameter $\phi_{\rm p}$ in the analytic kinetic-dominated solution, all models produce an amount of inflation compatible with observations. The background evolution in each case displays a generic behaviour. Following a non-inflating period of kinetic dominance, $H(t)$, $\phi(t)$ and their time derivatives continue to decrease until one obtains approximate equipartition $\dot{\phi}^2 \sim V(\phi)$ between the kinetic and potential energies of the inflaton. This marks the onset of a (typically brief) period of fast-roll inflation, which turns into a slow-roll [$\dot{\phi}^2 \ll V(\phi)$] or power-law inflation phase. At the end of the slow-roll phase, the inflation quickly moves towards a minimum of the potential, about which it executes a decaying oscillation.

We calculated the approximate spectrum of scalar perturbations for the polynomial and exponential models and find, in both cases, that it contains less power at low- and high-k values than would be expected from a power-law behaviour. The low-k effect is a generic consequence of the kinetic initial conditions, for any consistent inflaton potential or spatial curvature, resulting in particular from the brief period of fast-roll inflation that they imply. The damping of power on large scales may provide an explanation for the low-ℓ falloff in the matter and CMB power spectra seen in recent cosmological observations.

We also compared our kinetic initial conditions with an alternative proposal by Boyanovsky et al. (2006a) that inflationary initial conditions should be set by assuming approximate equipartition between the kinetic and potential energy of the inflaton. In the context of kinetic initial conditions, approximate equipartition is not a fundamental physical principle, but merely an inevitable consequence. Moreover, by considering a particular model used by Boyanovsky et al. (2006a), we demonstrate that assigning equipartition initial conditions with an arbitrary initial value for the inflaton field can lead to inconsistency with kinetic initial conditions.

Finally we enumerated the universes which do not have a steadily moving inflaton, and have shown that these are special cases, distinct from the generic kinetically dominated case.

Appendix 3.A Uniqueness theorem

We shall now prove that the solutions to the initial value problem of the master equation (3.31):

$$\frac{\mathrm{d}y}{\mathrm{d}\psi} = \sqrt{1-e^{-2y}} - \frac{\mathrm{d}}{\mathrm{d}\psi}\log\sqrt{V}, \tag{3.154}$$

$$y(\psi_0) = y_0 > 0, \tag{3.155}$$

are unique within any finite interval $\psi \in [\psi_0, \psi_1]$; i.e, if two positive solutions intersect at a point, then they intersect everywhere.

We begin by putting a lower bound on y in the interval $[\psi_0, \psi_1]$: From assumption (3.14), we know $\dot{\phi}^2 > \xi^2 > 0$. If we unpack the definition of y using equations (3.30), (3.21), (3.18) and (3.15), one finds:

$$y = \frac{1}{2}\log\left(\frac{\frac{1}{2}\dot{\phi}^2 + V(\phi)}{V(\phi)}\right) > \frac{\xi^2}{4V_{\max}}, \tag{3.156}$$

where $V_{\max}$ is the maximal value of $V(\psi)$ in the interval $[\psi_0, \psi_1]$. With this in hand we may prove the uniqueness of solutions of the initial value problem (3.154), (3.155) using standard techniques. For a good reference of such techniques the reader should consult the text by Agarwal et al. (1993). In this case, we shall prove it using *Peano iteration*.

If one assumes that $y(\psi)$ and $z(\psi)$ are two distinct solutions, then their difference satisfies:

$$\frac{\mathrm{d}}{\mathrm{d}\psi}(y-z) = \sqrt{1-e^{-2y}} - \sqrt{1-e^{-2z}}. \tag{3.157}$$

If in addition one assumes they meet at a common point ψ_0, so that $y(\psi_0) = z(\psi_0)$, then integrating away from this position yields:

$$\begin{aligned} |y(\psi) - z(\psi)| &= \left| \int_{\psi_0}^{\psi} \sqrt{1-e^{-2y}} - \sqrt{1-e^{-2z}}\ d\psi \right| \\ &\leq \int_{\psi_0}^{\psi} \left| \sqrt{1-e^{-2y}} - \sqrt{1-e^{-2z}} \right| d\psi. \end{aligned} \tag{3.158}$$

A generic property of the function $f(y) = \sqrt{1-e^{-2y}}$ is that in the interval $\left[\frac{\xi^2}{4V_{\max}}, \infty\right)$ it is *Lipschitz continuous*:

$$\left| \sqrt{1-e^{-2y}} - \sqrt{1-e^{-2z}} \right| \leq L\,|y-z|\,, \tag{3.159}$$

where L is the *Lipschitz constant*, taking the value:

$$L = \frac{\exp\left(-\frac{\xi^2}{2V_{\max}}\right)}{\sqrt{1-\exp\left(-\frac{\xi^2}{2V_{\max}}\right)}} > 0. \tag{3.160}$$

Applying Lipschitz continuity (3.159) to the inequality in (3.158) gives:

$$|y(\psi) - z(\psi)| \leq L \int_{\psi_0}^{\psi} |y(\psi) - z(\psi)|\,\mathrm{d}\psi. \tag{3.161}$$

Further, if the maximum value of the difference of $|y - z|$ between ψ_0 and ψ is Δ, then the above implies:

$$|y(\psi) - z(\psi)| \leq L\Delta \left| \int_{\psi_0}^{\psi} \mathrm{d}\psi \right| = L\Delta \left| \psi - \psi_0 \right|. \tag{3.162}$$

Applying this inequality back into (3.161) shows:

$$|y(\psi) - z(\psi)| \leq L^2\Delta \int_{\psi_0}^{\psi} |\psi - \psi_0| \, \mathrm{d}\psi = L^2\Delta \frac{|\psi - \psi_0|^2}{2!}. \tag{3.163}$$

Applying this back into (3.161) yields:

$$|y(\psi) - z(\psi)| \leq L^3\Delta \frac{|\psi - \psi_0|^3}{3!}, \tag{3.164}$$

and by induction on $n \in \mathbb{N}$ we find that:

$$|y(\psi) - z(\psi)| \leq L^n\Delta \frac{|\psi - \psi_0|^n}{n!}. \tag{3.165}$$

As $n \to \infty$ the term on the right-hand side drops to 0, and therefore $|y(\psi) - z(\psi)| = 0$. Thus, if y and z are equal at some point ψ_0, then they are equal at all points ψ within any finite interval $[\psi_0, \psi_1]$. Two separate solutions cannot "cross over", and if one positive solution f is initially less than a second solution h at ψ_0, $f(\psi_0) < h(\psi_0)$, then $f(\psi_1) < h(\psi_1)$ for any finite ψ_1.

Chapter 4

Further thoughts on kinetic dominance

This is a brief chapter whose content is based on discussions since the release of the material in Chapter 3 as the publication by Handley et al. (2014). To recap, the classical equations of motion for single field inflation are governed by:

$$\dot{H} + H^2 = -\frac{1}{3m_{\rm p}^2}\left(\dot{\phi}^2 - V(\phi)\right), \tag{4.1}$$

$$0 = \ddot{\phi} + 3H\dot{\phi} + V'(\phi). \tag{4.2}$$

We have proved that if these equations are integrated backward in time they almost always begin on a solution in which $\dot{\phi}^2 \gg V(\phi)$ (independent of the potential). This is demonstrated in Figure 4.1. In this kinetically dominated regime, the solutions take the analytical form:

$$H = \frac{1}{3t}, \qquad \dot{\phi} = \sqrt{\frac{2}{3}}\frac{m_{\rm p}}{t}, \qquad \phi = \phi_{\rm p} \pm \sqrt{\frac{2}{3}} m_{\rm p} \log\left(\frac{t}{t_{\rm p}}\right). \tag{4.3}$$

The key question is how physically relevant these observations are, and how they relate to more traditional inflationary set-ups and assumptions.

The classical equations have two opportunities to break down. First, it may be that this kinetically dominated state occurs before the Planck time, when we know that we would need a quantum theory of gravity to describe the universe. Second, it may be that the homogeneity assumption breaks down at early times.

4.1 The Planck time

The Planck epoch occurs at the earliest times and is defined as the regime in which we are certain we would need a quantum theory of gravity. Effectively this occurs when:

$$\rho = \frac{1}{2}\dot{\phi}^2 + V(\phi) \sim m_{\rm p}^4. \tag{4.4}$$

If the universe exits the kinetically dominated solutions before the Planck time, then kinetic dominance cannot be physically relevant. We may use this to put bounds on $\phi_{\rm p}$, or alternatively the total number of e-folds of inflation.

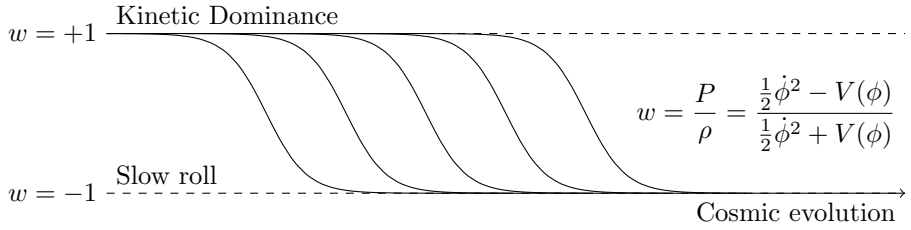


Figure 4.1: Schematic of the equation of state parameter w against cosmic evolution. The generic scenario entails beginning in a kinetically dominated phase with $w = 1$ (i.e. $\dot{\phi}^2 \gg V(\phi)$) before moving to the typical slow roll attractor solution with $w = -1$ (i.e. $\dot{\phi}^2 \ll V(\phi)$). We term the intermediate stage "Fast roll".

If we are in the kinetically dominated phase with $\dot{\phi}^2 \gg V(\phi)$ then:

$$\rho \approx \frac{1}{2}\dot{\phi}^2 = \frac{1}{3t^2} m_{\rm p}^2, \tag{4.5}$$

so $\rho \sim m_{\rm p}^4$ implies that $t \sim t_{\rm p} = m_{\rm p}^{-1}$. Thus, in order for kinetic dominance to be true at the Planck time we require that:

$$V(\phi_{\rm p}) \ll m_{\rm p}^4. \tag{4.6}$$

For $V(\phi) = \frac{1}{2}m^2\phi^2$ inflation, this amounts to requiring that $m^2\phi_{\rm p}^2 \ll m_{\rm p}^4$. In order to agree with CMB observations, we require that $m \sim 10^{-5} m_{\rm p}$, which amounts to requiring that:

$$\phi_{\rm p}^2 \ll 10^{10} m_{\rm p}^2. \tag{4.7}$$

We may phrase this in terms of the total number of e-folds:

$$N_{\rm tot} = \int_{t_{\rm begin}}^{t_{\rm end}} H dt = \int_{\phi_{\rm begin}}^{\phi_{\rm end}} \frac{H}{\dot{\phi}} d\phi, \qquad (H = \dot{N}) \tag{4.8}$$

$$= \frac{1}{m_{\rm p}^2} \int_{\phi_{\rm begin}}^{\phi_{\rm end}} \frac{V(\phi)}{V'(\phi)} d\phi. \qquad \text{(Slow Roll)} \tag{4.9}$$

For $V \sim \phi^n$, this gives:

$$N_{\rm tot} = \frac{\Delta(\phi^2)}{n\, m_{\rm p}^2}. \tag{4.10}$$

If we constrain $\phi_{\rm begin}$ by (4.7), and assume that the end of inflation is somewhere near the base of the potential $\phi_{\rm end} \sim 0$, one finds:

$$N_{\rm tot} \ll 10^{10}, \tag{4.11}$$

in order for there to be a pre-inflationary phase after the Planck time. There is therefore ample room in ϕ^n inflation for a kinetically dominated phase in between the required number of e-folds of $N_* \sim 50$–60 and the upper limit somewhere much less than 10^{10}. This upper limit agrees with the work of Remmen

and Carroll (2014), who derive the expected number of e-folds from a phase-space distribution approach.

It is difficult to imagine how one could possibly constrain the total number of e-folds observationally, unless the number was sufficiently small so as to be observable directly in the CMB power spectrum, for example in "just-enough" inflation by Ramirez and Schwarz (2009, 2012). For a given inflationary potential, one may place theoretical limits or make predictions on the total number of e-folds (Remmen and Carroll, 2014), but these estimates typically tend to be much, much larger than 50–60.

4.2 Eternal inflation

Theorists typically assume an effectively eternal inflationary phase all the way back to the Planck time. The typical argument by Linde (2008) goes something like this:

A. We should set initial conditions at the Planck time, when $\rho \sim m_{\rm p}^4$.

B. There will be some partitioning between kinetic and potential energy at this moment.

C. Given that we know nothing of the physics we should put uniform priors the amount of energy in the potential at this moment: $V(\phi_{\rm p}) \in [0, m_{\rm p}^4]$.

D. Therefore, on average we expect equipartition between kinetic and potential energy at the Planck time.

From Linde's arguments, the kinetically dominated regime is only present after the Planck time for $V(\phi_{\rm p})$ at the very bottom end of the prior range.

We can see this graphically in Figure 4.2. Solutions are plotted so that they are equally spaced in kinetic energy at the Planck time. Whilst graphs similar to the upper half of the figure are typically shown in lectures and textbooks, this is for an artificially high value of m. As m is decreased, the phase space look more like the lower half of Figure 4.2, and the "Planck circle" becomes an elongated ellipse when plotted in $(\phi, \dot{\phi})$ phase space.

A more modern and concrete encapsulation of Linde's arguments can be seen in Remmen and Carroll (2014), where they arrive from phase-space arguments at a prior on the Planck circle of:

$$\mathrm{P}\left(\theta\right) = \frac{3}{4}\left|\cos^2\theta \sin\theta\right|, \tag{4.12}$$

with $\tan\theta = \dot{\phi}/(m\phi)$, which puts even less weight on the kinetically dominated solutions.

If m is set to the more physically reasonable $m \sim 10^{-5} m_{\rm p}$, then a prior that is uniform in potential energy puts a lot of weight on extremely large values of ϕ. It might be equally reasonable therefore to suggest some kind of logarithmic prior on the value of ϕ at the Planck time, as shown in Figure 4.3. In this case, almost by definition, most solutions begin in a kinetically dominated phase.

The problem is that all of these arguments tend to the more philosophical end of physics: "What is the natural prior to assume on initial conditions?". It

is far better to try and answer these questions observationally. If the kinetically dominated phase is observable, then all of this is moot.

In regards to theory, what we can say with some certainty is that *if* there is a pre-inflationary phase (significantly after the Planck time) where the physics is still encapsulated by equations (4.1) & (4.2), then that phase must be kinetically dominated. For example, in just-enough inflation (Ramirez and Schwarz, 2009, 2012), to give consideration to initial conditions one must work in a kinetically dominated phase.

4.3 Breakdown of homogeneity

We have shown that all homogeneous universes classically begin in a kinetically dominated phase. It is profitable to examine the form of the perturbed equations (2.43), to take into account the evolution of any spatial variation in the solutions.

To first order, the i-j, i-0 and 0-0 Einstein equations read:

$$\Phi = \Psi, \tag{4.13}$$

$$0 = -\dot{\phi}\,\delta\phi + 2H\Phi + 2\dot{\Phi}, \tag{4.14}$$

$$0 = \left(6H^2 - \dot{\phi}^2 + 2\frac{k^2}{a^2}\right)\Phi + \left(-3H\dot{\phi} - \ddot{\phi}\right)\delta\phi + \dot{\phi}\dot{\delta\phi} + 6H\dot{\Phi}, \tag{4.15}$$

where we have absorbed the explicit potential dependence into $\ddot{\phi}$ terms. One may rearrange these to gain second-order equations in Φ or $\mathcal{R}$:

$$0 = \ddot{\Phi} + \left(H - 2\frac{\ddot{\phi}}{\dot{\phi}}\right)\dot{\Phi} + \left(\frac{k^2}{a^2} - \frac{1}{m_{\mathrm{p}}^2}\dot{\phi}^2 - 2\frac{\ddot{\phi}}{\dot{\phi}}H\right)\Phi, \tag{4.16}$$

$$0 = \ddot{\mathcal{R}} + \left(\frac{\dot{\phi}^2}{m_{\mathrm{p}}^2 H} + 3H + 2\frac{\ddot{\phi}}{\dot{\phi}}\right)\dot{\mathcal{R}} + \frac{k^2}{a^2}\mathcal{R}, \qquad \mathcal{R} = \Psi - \frac{H}{\dot{\phi}}\delta\phi. \tag{4.17}$$

There is of course also a second-order equation purely in $\delta\phi$, but this is particularly unappetising, and given that we have an explicit solution for it in (4.14), we shall not state it here. Technically, these equations are derived in the Newtonian gauge ($E = B = 0$), but since everything here is manifestly gauge invariant one may interpret all of the above equations as the relations in the equivalent gauge invariant variables.

If we apply the kinetically dominated solutions to the background variables, we arrive at:

$$0 = \ddot{\Phi} + \frac{7}{3}\frac{1}{t}\dot{\Phi} + \frac{k^2}{a^2}\Phi, \tag{4.18}$$

$$0 = \ddot{\mathcal{R}} + \frac{1}{t}\dot{\mathcal{R}} + \frac{k^2}{a^2}\mathcal{R}, \tag{4.19}$$

$$a \propto t^{1/3}. \tag{4.20}$$

Both equations are solvable with Bessel functions as:

$$0 = \ddot{x} + (1+p)\frac{1}{t}\dot{x} + \frac{k^2}{a^2}x, \tag{4.21}$$

$$\Rightarrow \qquad x = t^{-\frac{p}{2}}\left[AJ_{\frac{3}{4}p}\left(\frac{3k}{2a}t\right) + BY_{\frac{3}{4}p}\left(\frac{3k}{2a}t\right)\right], \tag{4.22}$$

$$\sim t^{-\frac{p}{2}-\frac{1}{3}}\left[C\cos\left(\frac{3k}{2a}t - \frac{3p\pi}{8} - \frac{\pi}{4}\right) + D\sin\left(\frac{3k}{2a}t - \frac{3p\pi}{8} - \frac{\pi}{4}\right)\right], \qquad t \gg 1, \tag{4.23}$$

where A, B, C and D are integration constants. The solutions are generically oscillatory with a polynomial decay with power $-\frac{p}{2} - \frac{1}{3}$.

Physically that means that as we integrate further backwards in time, although the homogeneous universe is drawn toward a kinetically dominated state, spatial inhomogeneities begin to increase.

Thus, if some portion of the universe at an early stage has $\dot{\phi}^2 \gg V(\phi)$, and is approximately homogeneous, then evolving forwards in time its homogeneity will tend to increase. Kinetic dominance thus causes patches of the universe to homogenise. This somewhat strengthens the assumption of homogeneity used in the proof of Chapter 3.

Equally, this means that there will be a breakdown in the homogeneity assumption at early times. It remains to be determined at what moment this breakdown occurs for physically observable scales in the early universe. More analytical investigation is required, and is a subject of my current research.

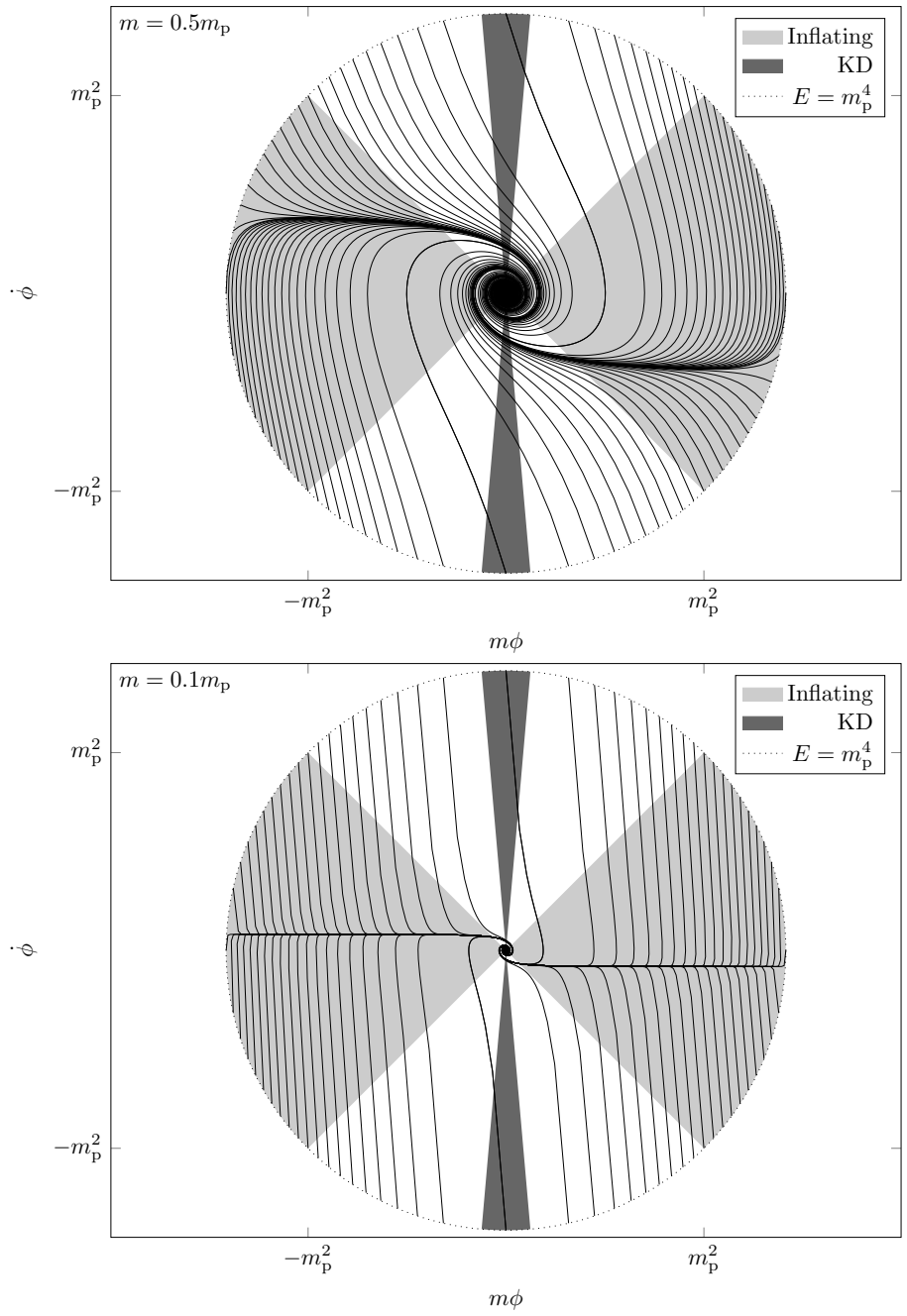


Figure 4.2: Phase-space plot of the evolution of $(\phi, \dot{\phi})$ for a chaotic inflationary potential $V(\phi) = \frac{1}{2}m^2\phi^2$. The upper plot has $m = 0.5m_\mathrm{p}$. The outer circle is the Planck time t_p when $\rho = m_\mathrm{p}^4$. One can see that each path through phase space is generically drawn to the slow roll attractor solutions in the centre of the plot. The lines are plotted so that they have a uniform spacing of kinetic energy at the Planck time. The upper plot is somewhat misleading, since it has an artificially high value of m. When the value of m is lowered to $0.1m_\mathrm{p}$, as in the lower plot, then the slow roll phase is a much stronger attractor solution, and generates a sustained slow-roll accelerated expansion.

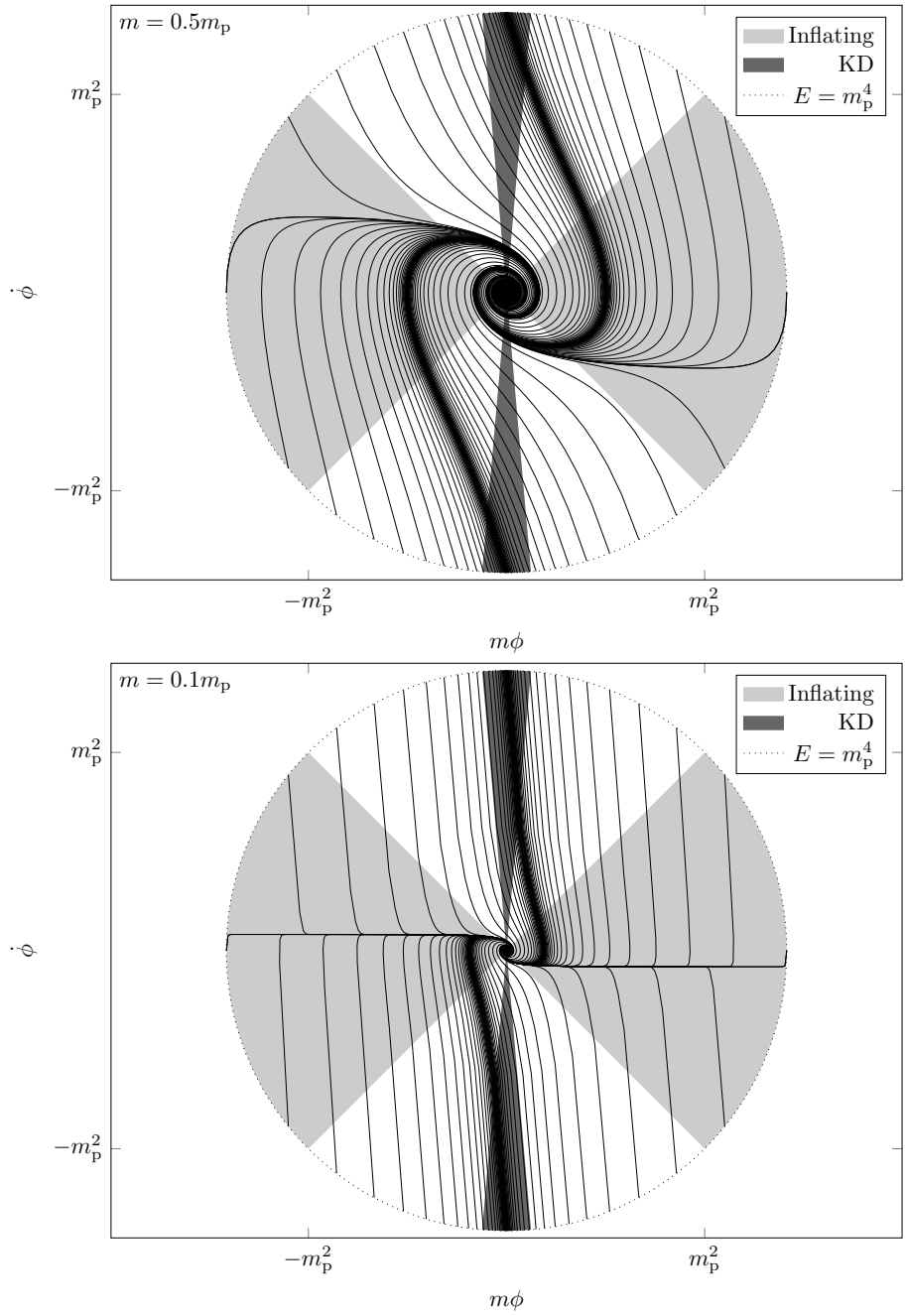


Figure 4.3: Phase space plot as in Figure 4.2, but with initial conditions that are log-uniformly spaced in ϕ at the Planck time. Here, most solutions begin in a kinetically dominated state, whilst still maintaining a sustained inflationary phase.

Chapter 5

Reconstructing the Primordial Power Spectrum

In Chapter 3 it was shown that all pre-inflationary universes take the same generic form. In addition, I demonstrated that the pre-inflationary phase could have produced a signature in the primordial power spectrum (PPS) of curvature perturbations (Sections 2.9–2.10). Given this, it is instructive to ask what our current state of knowledge is of this spectrum, and whether observations are consistent with this prediction.

In this chapter, I reconstruct the primordial power spectrum of curvature perturbations from a model-independent standpoint using Bayesian techniques applied to Planck 2015 data (Planck Collaboration et al., 2016a). A more detailed discussion of Bayesian analysis and techniques is given in Chapters 7–9. It is worth remarking that the analysis detailed in this chapter only became possible after the creation of PolyChord, which is detailed in Chapter 9.

I performed this analysis as a member of the Planck team, and appears in the inflation paper (Planck Collaboration et al., 2016b). The analysis is entirely my own work.

5.1 Strategy

In this chapter we model the primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ using a nested family of models where $\mathcal{P}_{\mathcal{R}}(k)$ is piecewise linear in the $\log(\mathcal{P})$-$\log(k)$ plane between a number of knots, $N_{\rm knots}$, that is allowed to vary. The question arises as to how many knots one should use, and we address this question using Bayesian model comparison. A family of priors is chosen where both the horizontal and vertical positions of the knots are allowed to vary. We examine the "Bayes factor" or "Bayesian evidence" as a function of $N_{\rm knots}$ to decide how many knots are statistically justified. A similar analysis has been performed by Vázquez et al. (2012a) and Aslanyan et al. (2014). In addition, we marginalize over all possible numbers of knots to obtain an averaged reconstruction weighted according to the Bayesian evidence.

The generic prescription is illustrated in Fig. 5.1. $N_{\rm knots}$ knots $\{(k_i, \mathcal{P}_i) : i = 1, \ldots, N_{\rm knots}\}$ are placed in the $(k, \mathcal{P}_{\mathcal{R}})$ plane and the function $\mathcal{P}_{\mathcal{R}}(k)$ is constructed by logarithmic interpolation (a linear interpolation in log-log space)

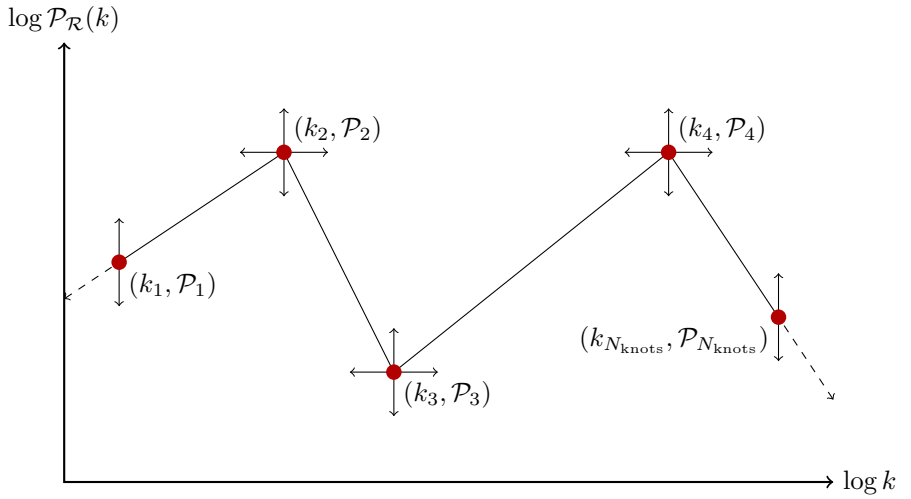


Figure 5.1: Linear spline reconstruction. The primordial power spectrum is reconstructed using $N_{\rm knots}$ interpolation points $\{(k_i, \mathcal{P}_i) : i = 1, 2, \ldots N_{\rm knots}\}$. The end knots are fixed in k but allowed to vary in $\mathcal{P}$, whereas the internal knots can vary subject to the constraint that $k_1 < k_2 < \cdots < k_{N_{\rm knots}}$. The function $\mathcal{P}_{\mathcal{R}}(k)$ is constructed within the range $[k_1, k_{N_{\rm knots}}]$ by interpolating logarithmically between adjacent knots (i.e., linearly in log-log space). Outside this range the function is extrapolated smoothly as indicated by the dashed lines in the diagram. The function $\mathcal{P}_{\mathcal{R}}(k; \{k_i, \mathcal{P}_i\})$ thus has $2N_{\rm knots} - 2$ parameters.

between adjacent points. Outside the horizontal range $[k_1, k_N]$ the function is extrapolated using the outermost interval.

Within this framework, base ΛCDM arises when $N_{\rm knots} = 2$—in other words, when there are two boundary knots and no internal knots, and the parameters $\mathcal{P}_1$ and $\mathcal{P}_2$ (in place of $A_{\rm s}$ and $n_{\rm s}$) parameterise the simple power-law PPS. There are also, of course, the four standard cosmological parameters ($\Omega_{\rm b}h^2$, $\Omega_{\rm c}h^2$, $100\theta_{\rm MC}$, and τ), as well the numerous foreground parameters associated with the *Planck* high-ℓ likelihood, all of which are unrelated to the PPS. This simplest model can be extended iteratively by successively inserting an additional internal knot, thus requiring with each iteration two more variables to parameterise the new knot position.

We run models for a variety of numbers of internal knots, $N_{\rm int} = N_{\rm knots} - 2$, evaluating the evidence for $N_{\rm int}$. Under the assumption that the prior is justified, the most likely, or preferred, model is the one with the highest evidence. Evidences are evaluated using the PolyChord sampler (Chapter 9 and Handley et al., 2015a,b) in CAMB and CosmoMC. The use of PolyChord is essential, as the posteriors in this parameterisation are often multimodal. Also, the ordered log-uniform priors on the k_i are easy to implement within the PolyChord framework. All runs were performed with 1000 live points, oversampling the semi-slow and fast parameters by a factor of 5 and 100, respectively.

Priors for the reconstruction and cosmological parameters are detailed in Table 5.1. We report evidence ratios with respect to the base ΛCDM case.

Parameter range	Prior type
$10^{-4}\,\mathrm{Mpc}^{-1} = k_1 < k_2 < \cdots < k_{N_{\mathrm{knots}}} = 0.3\,\mathrm{Mpc}^{-1}$	log uniform (sorted)
$2 < \log\left(10^{10}\mathcal{P}_1\right), \ldots, \log\left(10^{10}\mathcal{P}_{N_{\mathrm{knots}}}\right) < 4$	log uniform
$2 \leq N_{\mathrm{knots}} \leq 10$	integer uniform
$0.019 < \Omega_{\mathrm{b}}h^2 < 0.025$	uniform
$0.095 < \Omega_{\mathrm{c}}h^2 < 0.145$	uniform
$1.03 < 100\theta_{\mathrm{MC}} < 1.05$	uniform
$0.01 < \tau < 0.4$	uniform

Table 5.1: Prior for moveable knot positions. The $\mathcal{P}_{\mathcal{R}}$ positions are distributed in a log-uniform manner across a wide range. The k positions are also log-uniformly distributed across the entire range needed by CosmoMC and are sorted so that $k_1 < \cdots < k_{N_{\mathrm{knots}}}$. When we marginalize over the number of knots, N_{knots}, we assume a uniform prior between 2 and 10.

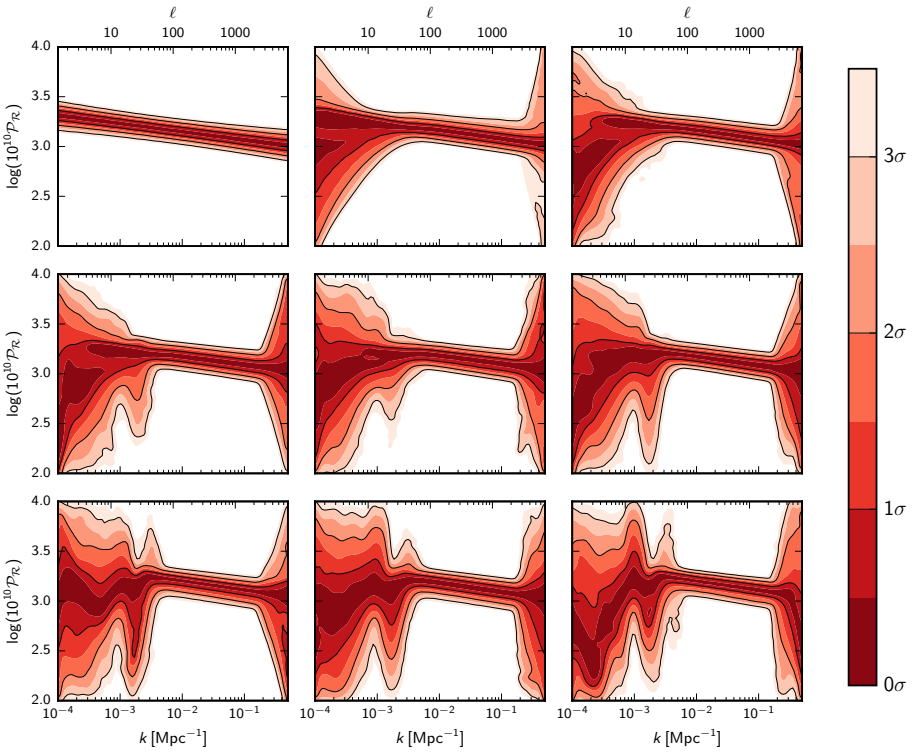


Figure 5.2: Bayesian movable knot reconstructions of the primordial power spectrum $\mathcal{P}_{\mathcal{R}}(k)$ using *Planck* TT data. The plots indicate our knowledge of the PPS $\mathrm{P}\left(\mathcal{P}_{\mathcal{R}}(k)|k, N\right)$ for a given number of knots. The number of internal knots N_{int} increases (left to right and top to bottom) from 0 to 8. For each k-slice, equal colours have equal probabilities. The colour scale is chosen so that darker regions correspond to lower-σ confidence intervals. $1\,\sigma$, $2\,\sigma$ and $3\,\sigma$ confidence intervals are also sketched (black curves). The upper horizontal axes give the approximate corresponding multipoles via $\ell \approx kD_{\mathrm{rec}}$, where D_{rec} is the comoving distance to recombination.

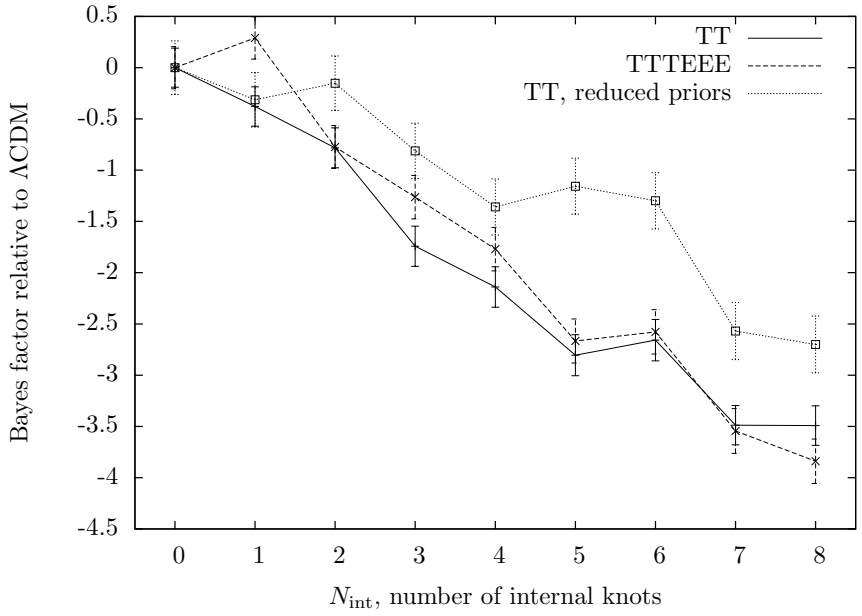


Figure 5.3: Bayes factor (relative to the base ΛCDM model) as a function of the number of knots for three separate runs. Solid line: *Planck* TT. Dashed line: *Planck* TT,TE,EE. Dotted line: *Planck* TT, with priors on the $\mathcal{P}$ parameters reduced in width by a factor of 2 ($2.5 < \log(10^{10}\mathcal{P}) < 3.5$).

The cosmological priors remain the same for all models, and this part of the prior has almost no impact on the evidence ratios. The choice of prior on the reconstruction parameters $\{\mathcal{P}_i\}$ does affect the Bayes factor. CosmoMC, however, puts an implicit prior on all models by excluding parameter choices that render the internal computational approximations in CAMB invalid. The baseline prior for the vertical position of the knots includes all of the range allowed by CosmoMC, so slightly increasing this prior range will not affect the evidence ratios. If one were to reduce the prior widths significantly, the evidence ratios would be increased. The allowed horizontal range includes all k-scales accessible to *Planck*. Thus, altering this width would be unphysical.

After completion of an evidence calculation, PolyChord generates a representative set of samples of the posterior for each model $\mathrm{P}(\Theta) \equiv \mathrm{P}(\Theta|\mathrm{data}, N_{\rm int})$. We may use this to calculate a marginalized probability distribution for the PPS:

$$\mathrm{P}(\log \mathcal{P}_\mathcal{R}|k, N_{\rm int}) = \int \delta(\log \mathcal{P}_\mathcal{R} - \log \mathcal{P}_\mathcal{R}(k;\Theta))\, \mathrm{P}(\Theta)\ \mathrm{d}\Theta. \qquad (5.1)$$

This expression encapsulates our knowledge of $\mathcal{P}_\mathcal{R}$ at each value of k for a given number of knots. Plots of this PPS posterior are shown in Fig. 5.2 using *Planck* TT data.

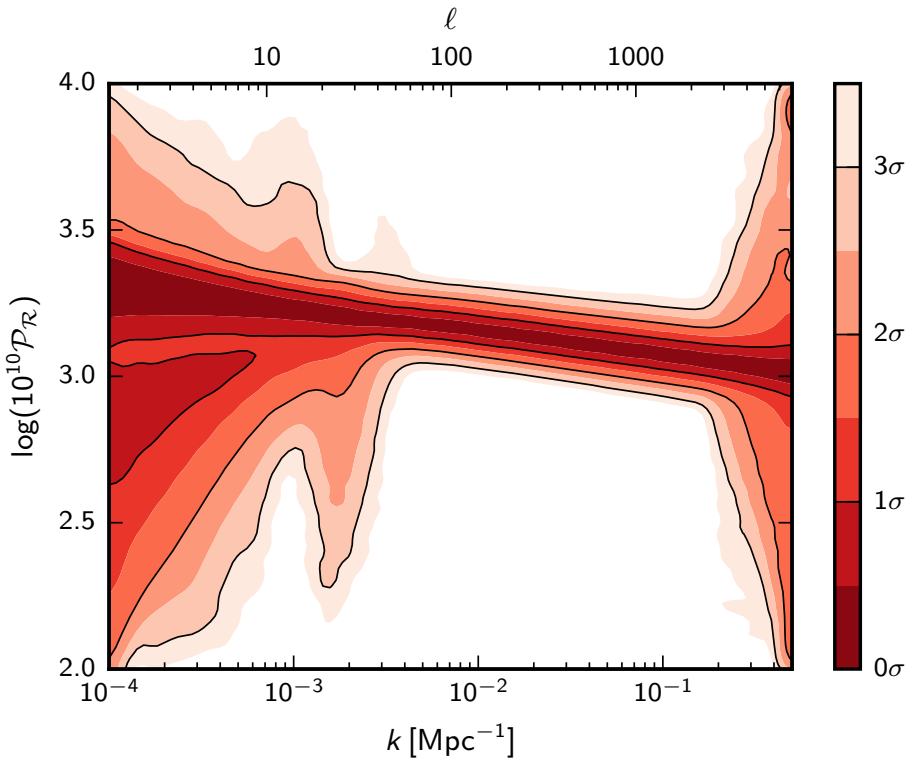


Figure 5.4: Bayesian reconstruction of the primordial power spectrum averaged over different values of $N_{\rm int}$ (as shown in Fig. 5.2), weighted according to the Bayesian evidence. The region $30 < \ell < 2300$ is highly constrained, but the resolution is lacking to say anything precise about higher ℓ. At lower ℓ, cosmic variance reduces our knowledge of $\mathcal{P}_{\mathcal{R}}(k)$. The weights assigned to the lower $N_{\rm int}$ models outweigh those of the higher models, so no oscillatory features are visible here.

5.2 Results

If one considers the Bayesian evidences of each model, Fig. 5.3 shows that although no model is preferred over base ΛCDM, the case $N_{\rm int} = 1$ is competitive. This model is analogous to the broken-power-law spectrum detailed in Section 4.4 of Planck Collaboration et al. (2016a), although the models differ significantly in terms of the priors used. In this case, the additional freedom of one knot allows a reconstruction of the suppression of power at low ℓ. Adding polarization data does not alter the evidences significantly, although $N_{\rm int} = 1$ is strengthened. We also plot a *Planck* TT run, but with the reduced vertical priors $2.5 < \log\left(10^{10}\mathcal{P}\right) < 3.5$. As expected, this increases the evidence ratios, but does not alter the above conclusion.

For increasing numbers of internal knots, the Bayesian evidence monotonically decreases. Occam's razor dictates, therefore, that these models should not be preferred, due to their higher complexity. However, there is an intriguing

stable oscillatory feature, at $20 \lesssim \ell \lesssim 50$, that appears once there are enough knots to reconstruct it. This is a qualitative feature predicted by several inflationary models (discussed in Section 9 of Planck Collaboration et al., 2016a), and a possible hint of new physics, although its statistical significance is not compelling.

A full Bayesian analysis marginalizes over all models weighted according to the normalized evidence $Z_{N_{\rm int}}$, so that:

$$\mathrm{P}\left(\log \mathcal{P}_{\mathcal{R}}|k\right) = \sum_{N_{\rm int}} \mathrm{P}\left(\log \mathcal{P}_{\mathcal{R}}|k, N_{\rm int}\right) Z_{N_{\rm int}}, \tag{5.2}$$

as indicated in Fig. 5.4. This reconstruction is sensitive to how model complexity is penalized in the prior distribution.

5.3 Conclusion

Whilst the fully marginalised power spectrum in Figure 5.4 certainly does not provide any evidence for a kinetically dominated phase, neither does it rule it out. There are hints of a feature at $\ell \sim 30$, and a possibility of a downturn in power at low k. Unfortunately, it appears that if there is in fact a suppression of power at low k, it is well into a regime that may well be cosmic variance limited. It would be instructive to examine via simulations what the reconstruction limits are for a fully cosmic-variance limited dataset, and whether one could ever detect a downturn at this scale. An analysis of this nature forms part of my current research.

Chapter 6

Defining the Quantum Vacuum

6.1 Introduction

Traditionally, quantum initial conditions for inflation are set using the Bunch-Davies vacuum. This approach is valid in de Sitter space and other asymptotically static spacetimes. Rapidly evolving spacetimes, however, do not admit such an easy quantisation.

In Chapter 3, we showed that the classical equations of motion suggest that the universe in fact emerged from the initial singularity in a rapidly evolving state, with the kinetic energy of the inflaton dominating the potential in a pre-inflationary phase. This can be used to set initial conditions on the background variables such as the inflaton value and Hubble parameter. In order to make contact with real observations, the effect that this phase has on the primordial power spectrum requires a semi-classical quantum mechanical treatment of the comoving curvature perturbation.

Hamiltonian diagonalisation is the simplest approach for setting quantum initial conditions in a general spacetime, and derives the vacuum from the minimisation of the Hamiltonian density. This approach has been criticised in the past as it does not admit a consistent interpretation in terms of particles (Fulling, 1989, 1979). Other approaches such as the adiabatic vacuum go some way to rescuing the particle concept, but have additional theoretical issues.

The issue of the particle interpretation stems from an attempt to apply a Minkowski spacetime concept outside the region of its validity. We postulate that the minimisation of an energy density is still an appropriate way to define a vacuum. In order to avoid the issues raised against Hamiltonian diagonalisation, we motivate our initial conditions from the minimisation of the *renormalised* stress-energy density. Indeed, if one takes care to minimise the correct quantity (using the theory of quantum fields in curved spacetime), then novel initial conditions can be derived which differ from the traditional Hamiltonian diagonalisation conditions.

After the relevant background material is reviewed, we develop a generic mechanism for setting initial conditions. These reduce to the Bunch-Davies case in asymptotically static spacetimes (such as de Sitter space), but yield different results otherwise. The aim is that these should be more theoretically robust. Additionally, these conditions are potentially distinguishable using observational data.

We then apply this procedure to the kinetically dominated universe, but defer the observational analysis to a later work.

6.2 Background

We denote a general action via:

$$S_I = \int \mathrm{d}^4x\sqrt{|g|}\mathcal{L}_I, \tag{6.1}$$

where $\mathcal{L}_I$ is the *Lagrangian density*. We work in natural units $\hbar = c = 1$ and set the reduced Planck mass $m_\mathrm{p} = (8\pi G)^{-1/2} = 1$. Dots denote differentiation with respect to cosmic time $\dot{f} \equiv \frac{\mathrm{d}}{\mathrm{d}t}f$, and primes denote differentiation with respect to conformal time $f' \equiv \frac{\mathrm{d}}{\mathrm{d}\eta}f$.

We begin by briefly summarising the classical theory of cosmological perturbations for a general scalar field, before discussing the quantisation of such a theory

6.2.1 The classical action

Consider (Baumann, 2009) a canonical scalar field ϕ minimally coupled to gravity $S = S_G + S_\phi$ with:

$$\mathcal{L}_G = \frac{1}{2}R, \qquad \mathcal{L}_\phi = \frac{1}{2}g^{\mu\nu}\nabla_\mu\phi\nabla_\nu\phi - V(\phi). \tag{6.2}$$

Extremising this action with respect to the fields ϕ and $g_{\mu\nu}$ recovers the Klein-Gordon and Einstein equations respectively:

$$\left(g^{\mu\nu}\nabla_\mu\nabla_\nu + \frac{\mathrm{d}V}{\mathrm{d}\phi}\right)\phi = 0, \tag{6.3}$$

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = T_{\mu\nu}, \tag{6.4}$$

where the stress-energy tensor is:

$$T_{\mu\nu} = \nabla_\mu\phi\nabla_\nu\phi - \frac{1}{2}g_{\mu\nu}\nabla_\alpha\phi\nabla^\alpha\phi + g_{\mu\nu}V(\phi). \tag{6.5}$$

In cosmology, we assume that at zeroth order both the metric $g_{\mu\nu}$ and scalar field ϕ are homogeneous and isotropic. Applying these assumptions to equations (6.3) & (6.4), we find:

$$\dot{H} + H^2 = -\frac{1}{3}\left(\dot{\phi}^2 - V(\phi)\right), \tag{6.6}$$

$$0 = \ddot{\phi} + 3H\dot{\phi} + \frac{\mathrm{d}V}{\mathrm{d}\phi}, \tag{6.7}$$

where the Hubble parameter $H = \dot{a}/a$.

6.2.2 Inflationary perturbations

One then considers scalar perturbations about these background solutions:

$$\begin{aligned} \phi =& \phi(t) + \delta\phi(t,x), && (6.8)\\ \mathrm{d}s^2 =& (1+2\Phi)dt^2 - 2a\partial_i B\,\mathrm{d}x_i\,\mathrm{d}t \\ & - a^2\left((1-2\Psi)\delta_{ij} + 2\partial_i\partial_j E\right)\mathrm{d}x_i\,\mathrm{d}x_j. && (6.9)\end{aligned}$$

We are interested in the gauge-invariant co-moving curvature perturbation:

$$\mathcal{R} \equiv \Psi - \frac{H}{\dot{\phi}}\delta\phi, \tag{6.10}$$

since it is this quantity which defines the primordial power spectrum for seeding cosmological perturbations. Working in the co-moving gauge $\delta\phi = 0$, and expanding the action S to second-order in $\mathcal{R}$, gives:

$$S^{(2)} = \int \mathrm{d}^4x a^3 \frac{\dot{\phi}^2}{H^2}\left[\dot{\mathcal{R}}^2 - a^{-2}(\partial_i\mathcal{R})^2\right]. \tag{6.11}$$

Note that the dependence on $V(\phi)$ is implicit in the variables $H, \dot{\phi}, a$ and $\mathcal{R}$. Defining the Mukhanov variable,

$$v = z\mathcal{R}, \qquad z = \frac{a\dot{\phi}}{H}, \tag{6.12}$$

and transforming t into conformal time $\eta = \int^t \mathrm{d}\tau/a(\tau)$ yields:

$$S^{(2)} = \int \mathrm{d}\eta\,\mathrm{d}^3x\left[(v')^2 - (\partial_i v)^2 + \frac{z''}{z}v^2\right]. \tag{6.13}$$

This is the canonically normalised action for a scalar field with time-dependent "effective" mass $m^2_{\text{eff}} = -z''/z$.

6.3 Quantisation via Hamiltonian diagonalisation

We now consider the traditional quantisation of the action (6.13) in spatially flat spacetimes via Hamiltonian diagonalisation. This is a standard method in the inflationary literature, but has several theoretical issues which will be discussed. To begin, one writes;

$$v = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\left[a_{\mathbf{k}}\chi_{\mathbf{k}}(\eta)e^{i\mathbf{k}\cdot\mathbf{x}} + a^\dagger_{\mathbf{k}}\chi_{\mathbf{k}}{}^*(\eta)e^{-i\mathbf{k}\cdot\mathbf{x}}\right], \tag{6.14}$$

which expresses the operator v as a superposition of creation and annihilation operators $\{a_{\mathbf{k}}, a^\dagger_{\mathbf{k}}\}$ (Mukhanov, 2007), with the mode functions written in separated form $u_{\mathbf{k}} = \chi_{\mathbf{k}}(\eta)e^{i\mathbf{k}\cdot\mathbf{x}}$. If one requires that the scalar field satisfies the equations of motion, along with the canonical commutator relation:

$$[a_{\mathbf{k}}, a^\dagger_{\mathbf{k}'}] = (2\pi)^3\delta^{(3)}(\mathbf{k}-\mathbf{k}'), \tag{6.15}$$

holds true, then the temporal part $\chi_{\mathbf{k}}(\eta)$ of the mode functions $u_{\mathbf{k}}$ must satisfy:

$$\chi_{\mathbf{k}}{}'' + \left(k^2 - \frac{z''}{z}\right)\chi_{\mathbf{k}} = 0, \tag{6.16}$$

$$\chi_{\mathbf{k}}{}'\chi_{\mathbf{k}}{}^* - \chi_{\mathbf{k}}{}^{*\prime}\chi_{\mathbf{k}} = -i. \tag{6.17}$$

The first of these is the classical equation of motion of the action (6.13), whilst the second is a normalisation constraint.

6.3.1 Choosing a vacuum

The complex mode functions $\chi_{\mathbf{k}}$ are not fully determined by condition (6.17). Although the overall phase of the mode $\chi_{\mathbf{k}}$ is unimportant, there is an additional degree of freedom for each $\mathbf{k}$ to be determined. The choice of this is equivalent to choosing a vacuum state $|0\rangle$, defined by $a_{\mathbf{k}}|0\rangle = 0$.

The traditional approach is to consider the Hamiltonian of the Mukhanov variable, which after normal ordering takes the form:

$$H = \frac{1}{2}\int \frac{\mathrm{d}^3k}{(2\pi)^3}\Big[a_{\mathbf{k}}a_{-\mathbf{k}}F_{\mathbf{k}}(\eta) + a^\dagger_{\mathbf{k}}a^\dagger_{-\mathbf{k}}F_{\mathbf{k}}{}^*(\eta) + \left(2a^\dagger_{\mathbf{k}}a_{\mathbf{k}} + \delta^{(3)}(0)\right)E_{\mathbf{k}}(\eta)\Big], \tag{6.18}$$

where

$$E_{\mathbf{k}}(\eta) = |\chi'_{\mathbf{k}}|^2 + \omega_k^2|\chi_{\mathbf{k}}|^2, \quad F_{\mathbf{k}}(\eta) = \chi'_{\mathbf{k}}{}^2 + \omega_k^2\chi_{\mathbf{k}}{}^2, \tag{6.19}$$

$$\omega_k^2(\eta) = k^2 - \frac{z''}{z}. \tag{6.20}$$

It is therefore attractive to choose either (i) the vacuum as an eigenstate of the Hamiltonian:

$$H|0\rangle \propto |0\rangle \Rightarrow F_{\mathbf{k}} = 0, \tag{6.21}$$

or (ii) that the vacuum minimises the expected energy:

$$\langle 0|H|0\rangle \propto \int \frac{\mathrm{d}^3k}{(2\pi)^3}E_{\mathbf{k}}. \tag{6.22}$$

As can be shown with standard linear algebra, these two conditions are equivalent, and result in the requirement that:

$$|\chi_{\mathbf{k}}|^2 = \frac{1}{2\omega_k}, \qquad \chi_{\mathbf{k}}{}' = -i\omega_k\chi_{\mathbf{k}}, \tag{6.23}$$

which provides enough information to set unambiguous initial conditions for the mode equation (6.16). When the condition (6.23) is satisfied, the Hamiltonian is diagonalised such that:

$$H(\eta_0) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\left(a^\dagger_{\mathbf{k}}a_{\mathbf{k}} + \frac{1}{2}\delta^{(3)}(0)\right)\omega_k(\eta_0). \tag{6.24}$$

We henceforth refer to (6.23) as the Hamiltonian diagonalising (HD) vacuum choice.

From (6.24), one may easily show that $a^\dagger_{\mathbf{k}}|0\rangle$ is a state with energy $\omega_k(\eta_0)$ and momentum $\mathbf{k}$. One therefore traditionally interprets the action of $a^\dagger_{\mathbf{k}}$ at time η_0 as creating a "particle" from the vacuum. This is a well established interpretation in flat Minkowski space.

Note that in general, the conditions (6.23) are only satisfied at a specific time η_0, and the vacuum is thus a time-dependent notion. Indeed, it is straightforward to show that at some different time η_1, the expectation $\langle 0|H(\eta_1)|0\rangle$ of the Hamiltonian in the ground state $|0\rangle$ is in general larger than the minimum possible value at η_1. This is interpreted as the vacuum state $|0\rangle$ at η_0 containing particles at other times η_1.

Thus, if one sets these conditions at a time η_0, one is effectively setting the universe to be in the vacuum state at that time. The expansion of the universe then excites the vacuum, creating "particles" at later times η_1. The question as to what is the "correct" η_0 at which to set these conditions is currently an unresolved theoretical (or indeed observational) issue.

6.3.2 Criticism of Hamiltonian diagonalisation

Astute readers will have spotted that the expected energy (6.22) is divergent. Whilst the implicit $\delta^{(3)}(0)$ in the proportionality constant of (6.22) is harmless, and merely accounts for the contribution from the infinite volume of space, there is a second divergence which requires closer attention. For large k, $E_{\mathbf{k}} \sim \omega_k \sim k$, and hence the integral (6.22), which represents the energy density, is ultraviolet divergent as k^4.

In traditional quantum field theory, this divergence is subtracted as one only measures energy differences. This is also applicable to spacetimes that are asymptotically static (such as de Sitter space). However, in changing spacetimes, where the vacuum is time dependent, this subtraction can only be performed at a single instant. If one then advances in time by some finite amount, the space-time generates an infinite particle density (Fulling, 1979, 1989).

This is clearly unphysical, causing some authors (Fulling, 1979) to discard Hamiltonian diagonalisation as an inappropriate methodology for choosing a vacuum state.

6.4 Alternative quantisations

The particle concept can be somewhat rescued by considering the adiabatic vacuum. This is well defined when the spacetime is changing slowly, as one can then perform an adiabatic expansion. The n^{th} order adiabatic vacuum at time η_0 is defined by matching the general solution of the mode equation (6.16) onto the n^{th} order adiabatic expansion at time η_0. This has the satisfying property of more closely corresponding to what a freely falling particle detector would measure, and is generally agreed to be superior to Hamiltonian diagonalisation.

However, there are still some issues with this vacuum. First, it is only usable in slowly changing spacetimes, so only goes halfway to solving the general problem. Second, it introduces a further ambiguity in vacuum choice, namely that of which value of n to choose. Since the adiabatic expansion is asymptotic, it does not in general converge for large n. One must pick a specific term of the

series to truncate at, and there is little theoretical guidance as to what value n to choose.

We believe that the adiabatic vacuum is in fact trying to rescue the particle concept unnecessarily. A particle interpretation is doomed to failure in general curved spacetime because of the global nature of their definition. Particles are defined in terms of field modes over a large patch of the manifold. Whilst for higher $\mathbf{k}$ modes the environment looks effectively Minkowski, low $\mathbf{k}$ modes are sensitive to the large scale structure of spacetime.

It would be more sensible to base the notion of a vacuum not in terms of a "particle-less" state, but in terms of the minimisation of a *local energy density*, such as the 0-0 component of the stress-energy tensor. Unfortunately being quadratic in the field ϕ, like the Hamiltonian, $\langle 0|T_{00}|0\rangle$ is also divergent.

In order to ameliorate this difficulty, we must adopt a more sophisticated approach.

6.5 Quantum fields in curved spacetime

This is the semi-rigorous theory of fields in which gravity is strong enough to generate curvature, but the quantum mechanics only affects spacetime to low order. It can therefore be thought of as a one-loop approximation to quantum gravity.

Traditionally (Birrell, 1984, Parker, 2009), one considers a scalar field Lagrangian with mass m, with action:

$$S = \int \mathrm{d}^4x\sqrt{|g|}\left(\frac{1}{2}g^{\mu\nu}\nabla_\mu\phi\nabla_\nu\phi - \frac{1}{2}m^2\phi^2\right), \tag{6.25}$$

where for simplicity we are considering the case of minimal coupling $\xi = 0$. In the context of FRW spacetime, the modes are quantised as:

$$\phi(x) = \int \frac{\mathrm{d}^3k}{(2\pi)^3 a(\eta)}\left[a_{\mathbf{k}}\chi_{\mathbf{k}}(\eta)e^{i\mathbf{k}\cdot\mathbf{x}} + a^\dagger_{\mathbf{k}}{\chi_{\mathbf{k}}}^*(\eta)e^{-i\mathbf{k}\cdot\mathbf{x}}\right], \tag{6.26}$$

where the mode functions are written in separated form $u_{\mathbf{k}} = a(\eta)^{-1}\chi_{\mathbf{k}}(\eta)e^{i\mathbf{k}\cdot\mathbf{x}}$. The additional conformal factor of $a(\eta)^{-1}$ generates mode equations without first order derivatives in η. Requiring that the scalar field satisfies the equations of motion, and that the commutation relation (6.15) remains true, one finds that the mode functions $\chi_{\mathbf{k}}$ must satisfy:

$${\chi''}_{\mathbf{k}} + \left[k^2 + a^2m^2 - \frac{a''}{a}\right]\chi_{\mathbf{k}} = 0, \tag{6.27}$$

$${\chi_{\mathbf{k}}}'{\chi_{\mathbf{k}}}^* - {\chi_{\mathbf{k}}}^{*\prime}\chi_{\mathbf{k}} = -i. \tag{6.28}$$

6.5.1 Application to inflation

The similarity between equations (6.16) and (6.27) is striking. It suggests solving for the quantum curvature perturbation is equivalent to solving a massless scalar field in an alternative spacetime with scale factor satisfying:

$$\frac{a''}{a} = \frac{z''}{z}. \tag{6.29}$$

This may be explicitly solved for $a(\eta)$ as:

$$a(\eta) = A\, z(\eta) + B\, z(\eta) \int^{\eta} \frac{\mathrm{d}x}{z(x)^2}, \tag{6.30}$$

where A and B are constants of integration.

Considering the special case of the inflating universe; during inflation H and $\dot{\phi}$ are approximately constant, so $z \propto a$, as in the exact de Sitter case. Thus, quantising the Mukhanov variable during inflation is equivalent to quantising a massless, minimally coupled scalar field on the same background spacetime. Note however, that in a more general scenario, the two spacetimes will not be the same.

6.6 Minimising the renormalised stress-energy tensor

Within the theory of quantum fields in curved spacetime, one is able to compute a *renormalised* stress-energy tensor $\langle 0|T_{\mu\nu}|0\rangle_{\mathrm{ren}}$. There are a variety of methods of doing this, but if carried out carefully they yield the same result.

6.6.1 Hadamard point splitting

We briefly recap the procedure for evaluating a renormalised stress-energy tensor via a Hadamard point splitting procedure. The Hadamard Green function is defined by:

$$G^{(1)}(x,x') = \frac{1}{2}\,\langle 0|\{\phi(x),\phi(x')\}|0\rangle\,. \tag{6.31}$$

The coincidence limit $x' \to x$ formally would yield the expectation $\langle 0|\phi^2|0\rangle$, but this is unfortunately divergent. The strategy therefore is to subtract off de-Witt Schwinger geometrical terms $G^{(1)}_{\mathrm{DS}}(x,x')$ which may be absorbed into a renormalisation of the "bare" constants G_B and Λ_B. One then takes the coincidence limit to yield a non-divergent quantity.

To form the stress-energy tensor from the Green function, one operates with a bi-scalar derivative function $D_{\mu\nu}(x,x')$:

$$\begin{aligned}\langle 0|T_{\mu\nu}(x)|0\rangle_{\mathrm{ren}} &= \lim_{x'\to x} \mathcal{D}_{\mu\nu}(x,x')\left[G^{(1)}(x,x') - G^{(1)}_{\mathrm{DS}}(x,x')\right], \\ \mathcal{D}_{\mu\nu}(x,x') &= \frac{1}{2}\left(\nabla_\mu\nabla_{\nu'} + \nabla_{\mu'}\nabla_\nu\right) - \frac{1}{2}g_{\mu\nu}\nabla_\alpha\nabla^{\alpha'} \\ &\quad + g_{\mu\nu}\frac{1}{2}m^2,\end{aligned} \tag{6.32}$$

where, as we are taking the coincidence limit, the metric $g_{\mu\nu}$ may be evaluated at either x or x'. The Hadamard Green function (6.31) using the mode expansion (6.26) becomes:

$$G^{(1)}(x,x') = \int \frac{\mathrm{d}^3k}{(2\pi)^3 a(\eta)a(\eta')}\Big(\chi_{\mathbf{k}}(\eta)\chi_{\mathbf{k}}{}^*(\eta')e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')} + \chi_{\mathbf{k}}{}^*(\eta)\chi_{\mathbf{k}}(\eta')e^{-i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')}\Big).$$

Inserting this expression into (6.32) will yield an expression which depends on the specific choice of mode function $\chi_{\mathbf{k}}$. We now regard this expression as a *functional* of the independent variables:

$$\mathcal{X} = \{\chi_{\mathbf{k}}, {\chi_{\mathbf{k}}}^*, {\chi_{\mathbf{k}}}', {\chi_{\mathbf{k}}}^{*\prime}\}, \tag{6.33}$$

and aim to minimise this with respect to the functions. Since $G^{(1)}_{\rm DS}$ does not depend on these variables, this term can be ignored for the purposes of extremisation. Further, the functional derivatives such as $\frac{\delta}{\delta\chi_{\mathbf{k}}}$ commute with the limit expression, so in fact minimising the renormalised tensor with respect to the mode functions is equivalent to naïvely minimising the traditional stress-energy tensor (6.5). Inserting the mode function (6.26) into (6.32) and taking the coincidence limit, one finds:

$$\begin{aligned}\langle 0|T_{00}(x)|0\rangle_{\rm ren} =& \frac{1}{2}\int \frac{{\rm d}^3k}{(2\pi)^3a^2}\Big[(\chi_{\mathbf{k}}' - \frac{a'}{a}\chi_{\mathbf{k}})({\chi_{\mathbf{k}}}^{*\prime} - \frac{a'}{a}{\chi_{\mathbf{k}}}^*) \\ &+ \left(k^2 + m^2a^2\right)\chi_{\mathbf{k}}{\chi_{\mathbf{k}}}^* + \tilde{T}\Big],\end{aligned} \tag{6.34}$$

where $\tilde{T}$ signifies the plethora of additional terms arising from the renormalisation process that have no dependence on the variables $\mathcal{X}$. Minimising this with respect to $\mathcal{X}$ subject to the constraint (6.28) yields the relations:

$$|\chi_{\mathbf{k}}|^2 = \frac{1}{2\sqrt{k^2+m^2a^2}}, \tag{6.35}$$

$$\chi_{\mathbf{k}}' = \left(-i\sqrt{k^2+m^2a^2} + \frac{a'}{a}\right)\chi_{\mathbf{k}}. \tag{6.36}$$

6.6.2 Application to the Mukhanov variable.

Recalling Section 6.5.1, to apply this formalism to the Mukhanov variable, one should set $m = 0$ and replace a with z:

$$|\chi_{\mathbf{k}}|^2 = \frac{1}{2k}, \qquad \chi_{\mathbf{k}}' = \left(-ik + \frac{z'}{z}\right)\chi_{\mathbf{k}}. \tag{6.37}$$

This should now be compared with the more usual HD conditions (6.23). Deep inside the horizon ($k \gg -z'/z$) these two initial conditions are equivalent, but yield very different answers for infra-red modes (small k). The second of these equations may be re-written in a more illuminating form:

$$\left(\frac{\chi_{\mathbf{k}}}{z}\right)' = -ik\left(\frac{\chi_{\mathbf{k}}}{z}\right), \tag{6.38}$$

which suggests that the co-moving curvature $\mathcal{R} = v/z$ is set with a "positive frequency mode" independent from any spacetime variation.

It is important to recognise setting these conditions at η_0 is equivalent to forcing the universe into a vacuum state at that moment, but there is minimal theoretical guidance as to when this should be[1]. Indeed, there is little reason

[1] Although Lasenby (2009) has suggested using successive adiabatic approximations to pick the vacuum epoch as the moment when the field is most "particle-like".

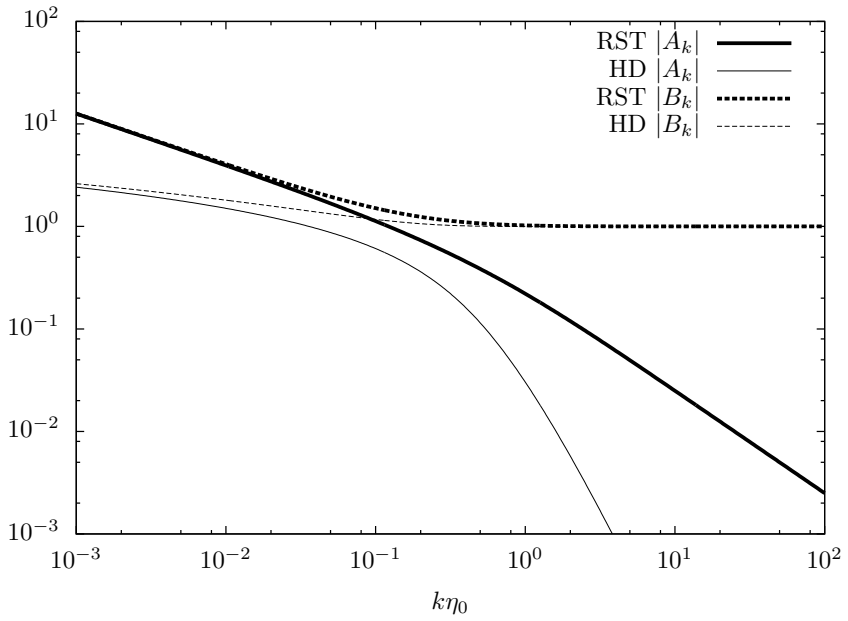


Figure 6.1: The modulus of the A_k and B_k coefficients in a kinetically dominated universe for the Hamiltonian diagonalising vacuum (HD) and the vacuum from the renormalised stress tensor (RST). Under these conditions, the universe will be in a vacuum state at conformal time η_0. Note that at large k, the mode functions tend to $A_k = 0$, $B_k = 1$.

to imagine that the universe should be in a vacuum state at any given moment. However, these conditions could also be used to build a formalism of excited states.

It is also important to realise that this vacuum does not claim to be interpretable in terms of particles. It is merely the mode function that minimises the renormalised stress tensor. In the language of Hamiltonian diagonalisation, or adiabatic vacuums, it would be a superposition of "particle states".

Armendáriz-Picón and Lim (2003) provides a review (particularly in the appendix) of various choices of initial conditions analogous to (6.23) and (6.37). It is interesting to note that the Danielsson (2002) vacuum bears a striking similarity to the renormalisation initial conditions (6.37) we have derived. The Danielsson (2002) vacuum (also discussed by Easther et al., 2002) is instead derived from phenomenological grounds by imposing initial conditions around a high energy cutoff.

6.7 Renormalising the KD universe

We now consider these observations in the context of the kinetically dominated universe. As is detailed in Chapter 3 (and in Handley et al., 2014), the classical solutions to the evolution equations (6.3) & (6.6) emerge almost always from a kinetically dominated phase with $\dot{\phi}^2 \gg V(\phi)$. In this regime, there is a significant period of cosmic time in which the theory of Section 6.5 is valid.

In this semi-classical pre-inflationary context, one finds that $\dot{\phi} \propto H$ and hence $z \propto a$. In the same manner as a de Sitter universe, quantising the co-moving curvature perturbation is equivalent to quantising a massless scalar field on the same background spacetime. In this case though, the scale factor $a \propto \eta^{1/2}$, so the mode equations have the general solution:

$$\chi_{\mathbf{k}}(\eta) = \frac{1}{2}\sqrt{\pi\eta}\left(A_k H_0^{(1)}(k\eta) + B_k H_0^{(2)}(k\eta)\right),$$
$$1 = |B_k|^2 - |A_k|^2, \tag{6.39}$$

where without loss of generality we assume A_k is real. Applying HD conditions (6.23), or our new renormalised stress tensor conditions (6.37) yields different values for A_k and B_k, as indicated in Figure 6.1. This difference is potentially observationally distinguishable, and will be analysed in a following paper.

6.8 Conclusions

We have presented a novel procedure for setting the initial conditions on the Mukhanov-Sazaki equation. We define the vacuum state via the instantaneous minimisation of the renormalised stress-energy tensor. This procedure is valid for any background cosmology, independent of the thorny issue of a particle-type concept. It reduces to the Bunch-Davies vacuum in an asymptotically static region. Further, it makes theoretical predictions that may be observationally testable.

Conclusion: Cosmology

This part began by showing in Chapter 3 that almost all classical inflationary solutions begin in a generic kinetically dominated phase. The generality of this statement was discussed in Chapter 4. Whether or not this phase occurs in reality can only be established by observation. If inflation was sufficiently short, then this pre-inflationary epoch may be observable as a suppression in power at low-ℓ in the C_ℓ spectrum, or at low-k in the $\mathcal{P}_\mathcal{R}(k)$ spectrum.

In Chapter 5, I showed that if one reconstructs the primordial power spectrum $\mathcal{P}_\mathcal{R}(k)$ from a Bayesian perspective there is some weak evidence for a suppression of power at low-k, as well as an anomaly at $\ell \sim 30$. Whilst this by no means provides evidence for an observable kinetically dominated epoch, it does suggest the possibility that with better data there could be.

In order to gain more theoretical guidance on the precise predictions which the kinetically dominated universe makes about the primordial power spectrum, we have to gain a greater understanding of the quantum mechanics of this epoch. The details of this are non-trivial, since as soon as one migrates away from a de Sitter limit, the theory as to how to set initial conditions becomes far more murky. Chapter 6 has enumerated some of these issues, and provided an alternative possibility for quantising a kinetically dominated universe.

Future work

Theoretically investigating kinetically dominance

The results of Chapter 4 require further investigation. In particular, since linear spatial perturbations grow backwards in time, the universe is more inhomogeneous closer to $t = 0$. This will lead to a breakdown in the assumptions at some point, and it would be helpful to quantify this fully. In particular, it would useful to know if the breakdown is earlier than the Planck scale for k-scales of observational interest.

Further, since the kinetically dominated universe stabilises forwards in time, it would be interesting to quantify how generic our initial conditions are, and whether a homogeneous kinetically dominated phase naturally arises for any universe beginning with $\dot{\phi}^2 \gg V(\phi)$. This would involve similar machinery to that used in "eternal inflation" scenarios.

Constraining the kinetically dominated universe

The next task would be to ask if the data can provide any further insight into the quantum vacuum of the kinetically dominated universe. It would be particularly interesting to find out if the data themselves were capable of distinguishing between vacua. This could be done for the current set of cosmological data, or one could ask about the feasibility of future data sets in providing constraints on this portion of the universe. In addition, one could also attempt to observationally constrain the epoch η_0 at which one should set the vacuum, by treating it as a free parameter in the model.

A full analysis would involve numerically integrating the quantum mechanical equations through the pre-inflationary phase all the way to horizon exit. Since these equations are highly oscillatory with time-varying coefficients, this would require a numerical method capable of tackling these. In fact, I have begun work on such a technique, which is detailed in Chapter 10.

If one of these vacua is the "correct" one, it still remains to be determined when, if at all, it was in its vacuum state. It may be that we can provide observational constraints on the precise value of this moment.

Further constraints on inflation

In addition, as a member of Planck core team II, I intend to continue more traditional observational reconstructions of inflationary and cosmological functions.

In general, inflationary analysis begins with the definition of the potential $V(\phi)$. This then predicts a primordial power spectrum $P_{\mathcal{R}}(k)$ via a numerical integration of the Mukhanov-Sazaki (MS) equations. The primordial power spectrum is then converted via cosmological transfer functions (Δ) into a set of multipole moments:

$$V(\phi) \xrightarrow{\text{MS}} P_{\mathcal{R}}(k) \overset{\Delta}{\rightsquigarrow} C_\ell$$

In Chapter 5 we applied a Bayesian reconstruction procedure to the middle stage, and reconstructed the primordial power spectrum in a model independent manner. We intend to apply our Bayesian reconstruction procedure separately to all three of these stages of the analysis, with the new updated polarisation data.

Part II

Methods

Chapter 7

Bayesian Inference

> We anticipate the sun will rise tomorrow, not just because it has always done so far, but because this is predicted by *models*, which accord with *data*. Any perceived failure of the sun to rise would more likely be a hallucination.
>
> David MacKay (2002)

Despite popular conception, science is not the search for truth. Scientists instead concern themselves with the construction of descriptive and predictive models that have their relative merit determined using data. The natural language to encapsulate these ideas is that of Bayesian probability. This chapter reviews the concepts and notation of this "meta-science".

7.1 Probability

Probability is the mathematical language of uncertainty. A numerical weighting is assigned to all events, which roughly corresponds to the "chance" that such an event would occur. An event $E \subset \Omega$ is defined as a subset of all possible outcomes Ω. Probability is defined as being additive over the set of all possible events, such that:

$$A \cap B = \phi \Leftrightarrow A,\, B \text{ disjoint} \Leftrightarrow \mathrm{P}\left(A\right) + \mathrm{P}\left(B\right) = \mathrm{P}\left(A \cup B\right). \tag{7.1}$$

John Skilling has derived these properties of probability from measurement-theoretical grounds (Hobson et al., 2009, chap. 1), which provides a more intuitive backing to the standard Kolmogorov axioms.

7.1.1 Bayes' theorem

The conditional probability of event B given event A is defined via:

$$\mathrm{P}\left(B|A\right) = \frac{\mathrm{P}\left(A \cap B\right)}{\mathrm{P}\left(A\right)}. \tag{7.2}$$

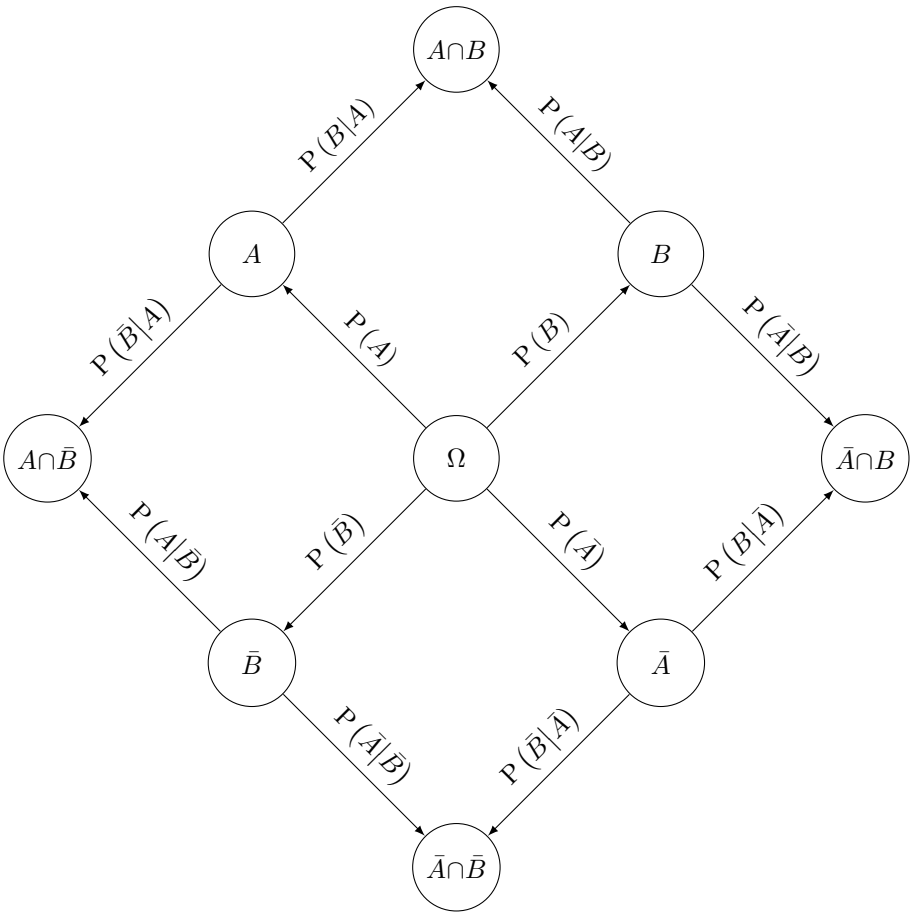


Figure 7.1: Diagrammatic Bayes' theorem. Nodes indicate the occurrence of an event. Edges indicate the probability of an event occurring, and are multiplicative in the direction of the arrows. Bayes' theorem (7.4) may be derived by equating the path $A \to A \cap B$ with $A \to \Omega \to B \to A \cap B$.

Multiplying both sides by $\mathrm{P}(A)$, and noting the symmetrical alternative:

$$\mathrm{P}(B|A)\,\mathrm{P}(A) = \mathrm{P}(A \cap B) = \mathrm{P}(A|B)\,\mathrm{P}(B)\,, \tag{7.3}$$

one may then derive Bayes' theorem:

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\,\mathrm{P}(B)}{\mathrm{P}(A)}. \tag{7.4}$$

Bayes' theorem indicates how to reverse the conditionality in a probability distribution $\mathrm{P}(A|B)$: one first deconditions by multiplying by $\mathrm{P}(B)$ and then reconditions via the normalising constant $\mathrm{P}(A)$. Figure 7.1 extends this to all possibilities for two events. With the algebra of conditionals, probability is the only measurable quantity with both a closed multiplication and addition.

7.2 Bayesian vs. Frequentist

There are two fundamental types of probability: *Aleatoric* and *Epistemological*. Aleatoric systems are genuinely random, for example: the flip of a coin, or the click on a Geiger counter. Epistemological probability governs systems where the randomness is associated with a subjective lack of knowledge.

Along with this, there are two interpretations of probability and the meaning of chance: *Bayesian* and *Frequentist*.

The Frequentist school of thought defines probability as:

Frequentist Probability: "The limiting relative frequency of an event."

If a coin has $\mathrm{P}(\text{Head}) = \frac{1}{2}$, then if you were to toss it an arbitrarily large number of times, the fraction of events that are heads would get closer and closer to $\frac{1}{2}$. This is the version of probability that most people encounter early in their mathematical education.

This definition applies relatively well to aleatoric systems (since an experiment can often be run a large number of times). However, for epistemological systems, it is all but useless. The frequentist solution is therefore to disregard the latter kind of probability.

For example, most people's experiences with probability will likely involve betting scenarios. For example, when given $7:1$ odds[1] on a horse's victory in the Grand National, it should be clear that these numbers do not refer to an event that can be repeated an arbitrarily large number of times.

As another example, before Prince George was born, bookmakers took bets on whether William and Kate's baby was a boy or a girl. Here it is obvious that the betting odds should be approximately $50:50$.[2,3] This is a classic example of epistemological probability. At the time of betting, the unborn infant is almost certainly either a boy or a girl. Any "chance" is down to the betting population's lack of knowledge. Interestingly, for a given couple, the aleatoric likelihood of male or female conception is not $50:50$. Factors that affect the sex at conception include maternal age, environmental factors, menstrual phase and individual genetics. As such, a sufficiently well informed bookmaker could make significant profit by tailoring bets to individuals, despite the fact that the population odds are $50:50$.

The above two examples suggest that a definition of probability that confines itself to aleatoric circumstances is too narrow. The alternative definition of probability is *Bayesian*:

Bayesian Probability: "A degree of belief that an event will occur."

Note that Bayesian probability is *subjective*, it matters whose degree of belief one is considering. Probabilities are assigned only with a given state of knowledge. Effectively, Bayesians place aleatoric and epistemological probability under the same umbrella.

---

[1] Betting odds $a:b$ indicates a probability of success of $\frac{b}{a+b}$.

[2] In fact the probability of a boy being conceived across the population is roughly 51%, in order to biologically account for the higher rate of male infant mortality.

[3] A good bookmaker will obviously take this into account, and give you slightly worse odds in order to ensure their profit.

Pure mathematics deals in relative truth or falsity, i.e. given initial assumptions, all statements are assigned to the set $\{F, T\} \equiv \{0, 1\}$. Bayesian probability can be thought of as a blurring of this process, namely from initial assumptions, various conclusions are assigned a number from the continuum between $[0, 1]$.

7.3 An example: biased coins

As an concrete introduction to the terminology, we will consider the case of determining whether a coin is biased. For example, assume that you toss a coin $N = 20$ times, and observe a dataset of $\mathcal{D} = 16$ heads. Is this enough to cause us to doubt the fairness of the coin?

The standard model $\mathcal{M}_0$ is that a coin toss consists of a binomial trial with probability $\frac{1}{2}$. Elementary probability tells us that the chance of getting this data is:

$$\mathrm{P}\left(\mathcal{D}|\mathcal{M}_0\right) = {}^{N}C_{\mathcal{D}}\left(\frac{1}{2}\right)^{N} = 4.6 \times 10^{-3}. \tag{7.5}$$

Note that as should be expected, the chance of obtaining this exact dataset is small.

If we allow for the possibility that the coin could yield a head with any probability p, we can encapsulate this in a second model $\mathcal{M}_1$.

$$\mathrm{P}\left(\mathcal{D}|p, \mathcal{M}_1\right) = {}^{N}C_{\mathcal{D}}\, p^{\mathcal{D}}(1-p)^{N-\mathcal{D}}. \tag{7.6}$$

This second model is not ideal, since we have not specified the parameter p. If we choose $p = \frac{1}{2}$ we recover $\mathcal{M}_0$. It is tempting to choose p such that the probability of getting the data is maximised. This would be an example of a *maximum likelihood* approach. In this case, the method indicates we should choose $p = \frac{\mathcal{D}}{N}$.

Maximum likelihood has its drawbacks, particularly in the case of a low volume of data. After $N = 1$ toss, it seems a little premature to choose either $p = 1$ or $p = 0$ depending on whether we see a head or a tail. Instead picking a specific value of p, one could instead "spread your bets" and consider several different values of p. The full generalisation of this is to consider a continuum of models, and define an initial probability distribution on p, $\mathrm{P}\left(p|\mathcal{M}_1\right)$. This can be interpreted as our initial assumptions on the value of p, or our initial degree of belief in its value. A non-partisan assumption is to try and be as unbiased as possible and assume that p is equally likely to take any value in between 0 and 1. We therefore have:

$$\mathrm{P}\left(p|\mathcal{M}_1\right) = \begin{cases} 1 & : 0 \leq p \leq 1 \\ 0 & : \text{otherwise.} \end{cases} \tag{7.7}$$

With this distribution, we can work out what the overall probability of obtaining the dataset $\mathcal{D}$ is, by marginalising over all values of p:

$$\mathrm{P}\left(\mathcal{D}|\mathcal{M}_1\right) = \int \mathrm{P}\left(\mathcal{D}|p, \mathcal{M}_1\right) \mathrm{P}\left(p|\mathcal{M}_1\right)\, \mathrm{d}p, \tag{7.8}$$

$$= \int_0^1 {}^{N}C_{\mathcal{D}}\, p^{\mathcal{D}}(1-p)^{N-\mathcal{D}}\, \mathrm{d}p, \tag{7.9}$$

$$= \frac{1}{N+1} = 4.8 \times 10^{-2}. \tag{7.10}$$

This is quite telling, because our initial choice (7.7) for $\mathrm{P}(p|\mathcal{M}_1)$ indicates that we expect all data sets with equal probability (independent of $\mathcal{D}$). This is therefore a "minimally suspicious" choice in spread.

We now have two equations (7.5) & (7.10) which detail the probability of getting the dataset $\mathcal{D}$ given the choice of model. However, what we are really after is the probability of the model, given the dataset $\mathrm{P}(\mathcal{M}|\mathcal{D})$. To compute this, we use Bayes' theorem:

$$\mathrm{P}(\mathcal{M}|\mathcal{D}) = \frac{\mathrm{P}(\mathcal{D}|\mathcal{M})\,\mathrm{P}(\mathcal{M})}{\mathrm{P}(\mathcal{D})}. \tag{7.11}$$

In order to complete the calculation, we must assign a probability to each model. A natural choice is to consider $\mathrm{P}(\mathcal{M}_0) = \mathrm{P}(\mathcal{M}_1) = \frac{1}{2}$. $\mathrm{P}(\mathcal{D})$ is just a normalising constant, computed as:

$$\mathrm{P}(\mathcal{D}) = \sum_i \mathrm{P}(\mathcal{D}|\mathcal{M}_i)\,\mathrm{P}(\mathcal{M}_i)\,, \tag{7.12}$$

so we may finally compute:

$$\mathrm{P}(\mathcal{M}_0|\mathcal{D}) = 0.088, \qquad \mathrm{P}(\mathcal{M}_1|\mathcal{D}) = 0.922. \tag{7.13}$$

In other words, a betting man would give the odds of the coin being unbiased $\mathcal{M}_0$ as approximately $10:1$. Another way of thinking of this is that $\mathcal{M}_1$ is ten times better at describing the data than $\mathcal{M}_0$.

$\mathcal{M}_1$ can go further however. We may also use Bayes' theorem to compute:

$$\mathrm{P}(p|\mathcal{D},\mathcal{M}_1) = \frac{\mathrm{P}(\mathcal{D}|p,\mathcal{M}_1)\,\mathrm{P}(p|\mathcal{M}_1)}{\mathrm{P}(\mathcal{D}|\mathcal{M}_1)}. \tag{7.14}$$

This here gives us the distribution on p *given the data*. Namely, how the data should update our "spread bet". We already have the ingredients for the above construction from equations (7.6), (7.7), & (7.10), and given the uniform prior, we find our updated bet on p is proportional to $p^{\mathcal{D}}(1-p)^{N-\mathcal{D}}$, a function of p. This is a beta distribution, and is indicated in Figure 7.2.

7.4 Parameter estimation & model comparison

We shall now take the concepts of the previous section, and put them in a general setting.

The typical problem of science is the construction of a model $\mathcal{M}$ in order to explain some dataset $\mathcal{D}$. In general, scientific models have a set of continuous parameters $\Theta_{\mathcal{M}}$, where $\Theta_{\mathcal{M}}$ is normally multi-dimensional, and may contain a variety of parameter types such as integers, vectors, tensors and more exotic components.

Elementary probability theory then enables us to calculate the probability of the data, given the choice of model along with a specific parameter choice.

$$\mathcal{L} \equiv \mathrm{P}(\mathcal{D}|\Theta_{\mathcal{M}},\mathcal{M})\,. \tag{7.15}$$

This distribution is called the *likelihood*, which is denoted with a calligraphic $\mathcal{L}$ to differentiate between rapidly proliferating conditional probabilities.[4] It

[4]This should not be confused with a Lagrangian, also denoted with $\mathcal{L}$.

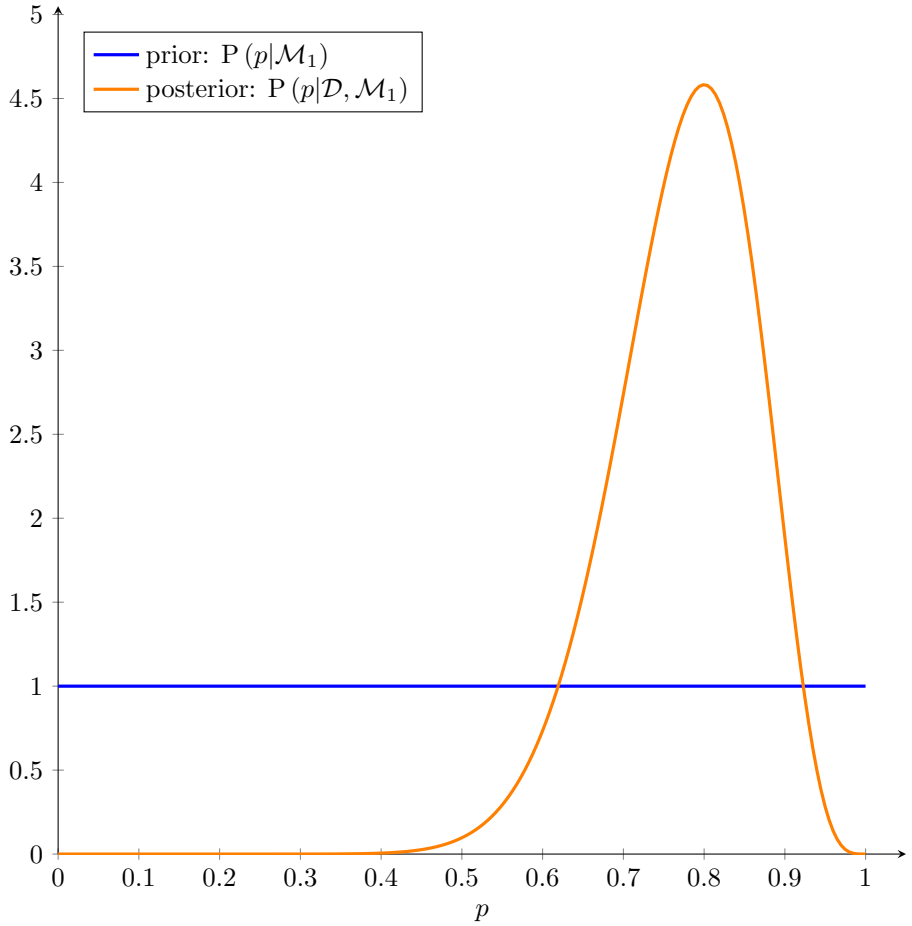


Figure 7.2: Our beta-function posterior degree of belief in the bias of the coin p, after observing a dataset of $(\mathcal{D}, N) = (16, 20)$, and assuming a uniform prior on p.

is also clearer in many situations to suppress explicit data, model and/or parameter-dependence of the likelihood, and instead write:

$$\mathrm{P}\left(\mathcal{D}|\Theta_{\mathcal{M}}, \mathcal{M}\right) \equiv \mathcal{L}_{\mathcal{M}}(\Theta_{\mathcal{M}}) \equiv \mathcal{L}(\Theta) \equiv \mathcal{L}_{\mathcal{M}} \equiv \mathcal{L}. \tag{7.16}$$

This is a generic overloading technique which I utilise throughout this thesis.

In order to perform Bayesian inference, another requirement of the model $\mathcal{M}$ is that it must specify an initial degree of knowledge of the parameters:

$$\pi \equiv \mathrm{P}\left(\Theta_{\mathcal{M}}|\mathcal{M}\right). \tag{7.17}$$

Since no model occurs in isolation, it is generally not difficult to theoretically produce upper and lower bounds on parameter values. The normal strategy is to choose fairly conservative uniform or Gaussian priors on parameter values. In general, the prior should encapsulate the scale and spread of our current expectation of the parameter value.

Once a prior has been specified, the model $\mathcal{M}$ is complete. The science of the problem is finished, and the rest of the analysis is statistics. Statistical analysis may be neatly partitioned into two problems: *model comparison* and *parameter estimation*.

7.4.1 Model comparison

It is usually the case in science that there is more than one model available to explain the data. Typically one will have a set of models $\{\mathcal{M}_1, \mathcal{M}_2, \ldots\}$, whose relative merit must be scientifically determined using data $\mathcal{D}$.

We may use the prior to marginalise out any of the parameter dependence within each model:

$$\mathcal{Z} \equiv \mathrm{P}\left(\mathcal{D}|\mathcal{M}\right) = \int \mathrm{P}\left(\mathcal{D}|\Theta_{\mathcal{M}}, \mathcal{M}\right) \mathrm{P}\left(\Theta_{\mathcal{M}}|\mathcal{M}\right) \, \mathrm{d}\Theta_{\mathcal{M}}. \tag{7.18}$$

This quantity is termed the *evidence* $\mathcal{Z}$, or *marginalised likelihood*, and gives the probability of observing the data $\mathcal{D}$, conditioned on the model $\mathcal{M}$. The quantity we seek however is the probability of each model $\mathcal{M}$ given the data $\mathcal{D}$, which may be obtained using Bayes' theorem:

$$\mathrm{P}\left(\mathcal{M}_i|\mathcal{D}\right) = \frac{\mathrm{P}\left(\mathcal{D}|\mathcal{M}_i\right) \mathrm{P}\left(\mathcal{M}_i\right)}{\mathrm{P}\left(\mathcal{D}\right)}. \tag{7.19}$$

In order to utilise this, we must specify our prior degree of belief in each model:

$$\mathrm{P}\left(\mathcal{M}_i\right) = \phi_i. \tag{7.20}$$

These may have been obtained from previous analyses, but a simple choice would be to choose the models to be equally weighted. The denominator of (7.19) is then a normalising constant, and the posterior degree of belief in each model may be obtained via:

$$\mathrm{P}\left(\mathcal{M}_i|\mathcal{D}\right) \equiv \mathcal{W}_i = \frac{\mathcal{Z}_i \phi_i}{\sum_j \mathcal{Z}_j \phi_j}. \tag{7.21}$$

These model weights $\mathcal{W}$ may then be used to determine the "most probable model". In some cases there is a clear winner, and the other models may be safely discarded, but the more usual scenario is that there are several competing alternatives. It is for this reason that we prefer the term "model comparison" to the more oft-quoted *model selection*. Additional datasets may determine a clear winner, but in the mean-time the weights $\mathcal{W}$ can be used to perform proper inference.

7.4.2 Parameter estimation

Of equal interest to scientists is to ask what the data tells us about the various parameters. Bayes' theorem allows us to invert the conditioning in equation (7.15) and find the *posterior* $\mathcal{P}$ by combining the likelihood (7.15), prior (7.17) and evidence (7.18):

$$\mathcal{P} \equiv \mathrm{P}\left(\Theta_{\mathcal{M}}|\mathcal{D}, \mathcal{M}\right) = \frac{\mathrm{P}\left(\mathcal{D}|\Theta_{\mathcal{M}}, \mathcal{M}\right) \mathrm{P}\left(\Theta_{\mathcal{M}}|\mathcal{M}\right)}{\mathrm{P}\left(\mathcal{D}|\mathcal{M}\right)}, \tag{7.22}$$

which is schematically written as:

$$\mathcal{P} = \frac{\mathcal{L} \times \pi}{\mathcal{Z}}, \tag{7.23}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}. \tag{7.24}$$

This describes how our initial knowledge π of the parameters updates to $\mathcal{P}$ in light of the data $\mathcal{D}$. Note that the evidence $\mathcal{Z}$ features in both model selection and parameter estimation, and its computation is therefore of great significance.

Evidences also enter into parameter estimation when performing a combined inference on several models. For example, models often predict a distribution for a common derived parameter $\mathrm{P}\left(y|\mathcal{M}, \mathcal{D}\right)$.[5] If the data are not strong enough to distinguish a given model, the correct inference on y is to use a posterior which marginalises over all models:

$$\mathrm{P}\left(y|\mathcal{D}\right) = \sum_i \mathrm{P}\left(y|\mathcal{D}, \mathcal{M}_i\right) \mathrm{P}\left(\mathcal{M}_i|\mathcal{D}\right) = \sum_i \mathrm{P}\left(y|\mathcal{D}, \mathcal{M}_i\right) \mathcal{W}_i, \tag{7.25}$$

where the $\mathcal{W}_i$ are defined in equation (7.21), and depend on the evidence of each model. This fully Bayesian approach has been historically under-utilised due to the difficulties in numerically computing the evidence.

7.5 Numerical statistics: sampling

Having discussed the theory of Bayesian statistics, we now turn to the more challenging aspect of actually computing these various inferences. The likelihood (7.15) $\mathcal{L}(\Theta)$ is a routine, if often challenging, quantity to compute. It is the job of observational scientists to provide this function, and for the purposes of inference we may consider it a "black box". In general, $\mathcal{L}$ will be not be analytical, but instead is a numerical and computationally expensive quantity. Any calculation we perform must aim to minimise the number of times we attempt to evaluate $\mathcal{L}$. The prior π is typically much less expensive to compute and is normally expressed using analytic functions such as a uniform or Gaussian distribution. In particular, it is usually straightforward to obtain samples distributed according to such priors.

Typically, in inference calculation we wish to compute quantities that are marginalised over by the posterior. For example means and variances will typically take the form:

$$\langle f \rangle = \int f(\Theta)\mathcal{P}(\Theta)\ \mathrm{d}\Theta, \tag{7.26}$$

$$= \int f(\Theta)\frac{\mathcal{L}(\Theta)\pi(\Theta)}{\mathcal{Z}}\ \mathrm{d}\Theta. \tag{7.27}$$

Given a likelihood $\mathcal{L}$ and a prior π, a naïve approach would be to first compute the evidence:

$$\mathcal{Z} = \int \mathcal{L}(\Theta)\pi(\Theta)\ \mathrm{d}\Theta, \tag{7.28}$$

[5] E.g. in cosmology, different models of the universe will predict an age distribution.

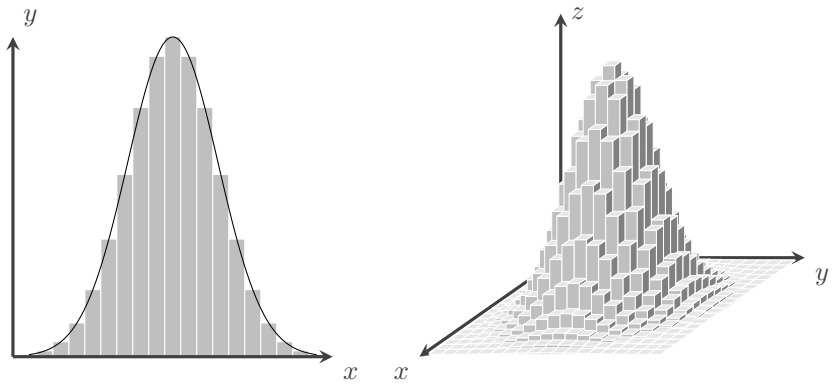


Figure 7.3: Approximating an integral using a quadrature rule. In moving a one-dimensional integral (left) to a two-dimensional integral (right), the number of quadrature points becomes exponentially larger. This is an example of the curse of dimensionality.

and then perform the integral (7.26) using a traditional numerical quadrature procedure. In most cases, this method fails at both steps, due to the fact that the dimensionality of the integration is too high for numerical quadrature to succeed.

To see this issue, consider the quadrature integration of a function g along $[0, 1]$

$$\int_0^1 g(x)\,\mathrm{d}x \approx \sum_{i=0}^{n} g(x_i) w_i, \tag{7.29}$$

where $x_i \in [0, 1]$ are the quadrature points $w_i \in \mathbb{R}$ are the quadrature weights. For example, if $x_i = \frac{i}{n}$ and $w_i = \frac{1}{n+1}$ then one obtains the left rectangle rule (Figure 7.3). This calculation therefore requires $\sim \mathcal{O}(n)$ calculations of the function. In the D-dimensional case the generalisation is:

$$\int_0^1 \cdots \int_0^1 g(x_1, \ldots, x_D)\,\mathrm{d}x_1 \cdots \mathrm{d}x_D \approx \sum_{i_1,\ldots,i_D=0}^{n} g(x_{i_1}, \ldots, x_{i_D}) w_{i_1,\ldots,i_D}. \tag{7.30}$$

This calculation however requires $\sim \mathcal{O}\left(n^D\right)$ function evaluations. This is an exponential scaling with D, and is an example of the *curse of dimensionality*. Even for modest number of parameters [$D \sim \mathcal{O}(6)$] these kind of integrations require unfeasible amounts of computational time. Worse still, for most likelihoods, the region about which $\mathcal{L}$ is significantly non-zero is much smaller than the prior range[6]. The number of grid points in each dimension n must therefore be taken as reasonably large in order to ensure that enough function evaluations occur at the peak.

Fortunately there is a better way. If one has a set of n_s samples:

$$S = \{\Theta_i \sim \mathcal{P} : i = 1, \ldots, n_s\}, \tag{7.31}$$

[6]This should be expected, since the good data will update the prior information by some significant amount in each dimension.

distributed according to the posterior $\mathcal{P}$, then we may use the samples to compute the sample average of f:

$$\hat{E}_f = \frac{1}{n_s} \sum_{i=1}^{n_s} f(\Theta_i). \tag{7.32}$$

The central limit theorem states that $\hat{E}_f$ is normally distributed with mean $\langle f \rangle$ and standard deviation $\sigma_f / \sqrt{n_s}$, where σ_f is the true standard deviation $\left(\langle f^2 \rangle - \langle f \rangle^2 \right)^{1/2}$. If $n_s \gg 1$ is large, then $\hat{E}_f$ will closely approximate the true mean.

This approach has several advantages. If one already has the set S of samples (7.31), the computation (7.32) is extremely cheap, in contrast to a full integral (7.30). Further, it has excellent scaling with dimensionality.

All that remains in order to compute (7.32) is a methodology for cheaply generating samples according to a given distribution $\mathcal{P}$. In many cases, one does not even need the normalisation constant $\mathcal{Z}$ in order to sample from $\mathcal{P}$, merely an unnormalised posterior $\mathcal{P}^* = \mathcal{L} \times \pi$. Suffice to say that this is a well established problem, and for readers unfamiliar with standard sampling methods I refer them to Appendix 7.A.

However, even if one can generate reliably and cheaply a full set S of samples, this is not a full solution to the entire Bayesian inference problem problem. Computing $\langle f \rangle$ with sampling is only a reasonable methodology when σ_f is small (or even defined at all). A very important example of a case where this is not true is the computation of the evidence integral (7.28). For these kind of problems we must turn to a more sophisticated methodology

7.6 Nested sampling

Nested sampling is a methodology devised by Skilling (2006) to efficiently compute the evidence:

$$\mathcal{Z} = \int \mathcal{L}(\Theta) \pi(\Theta) \, \mathrm{d}\Theta. \tag{7.28 revisited}$$

This integral is typically over a high-dimensional parameter space, only a small fraction of which contributes to $\mathcal{Z}$. We may quantify the size of this region by considering the *relative entropy*[7] of the prior from the posterior distribution:

$$H = -\int \log \left(\frac{\mathcal{P}(\Theta)}{\pi(\Theta)} \right) \mathcal{P}(\Theta) \, \mathrm{d}\Theta. \tag{7.33}$$

This quantifies the number of nats[8] of information that have been gained on account of the data. Equivalently, the fraction of the prior which contributes non-negligibly to (7.28) is approximately $X \sim e^{-H}$.

The size and position of the region surrounding the peak (or peaks) will not be known *a priori*, and in high dimensions is challenging to find.

[7] Also known as: Kullback-Leibler divergence, information divergence, information gain, KLIC, KL divergence.

[8] If "bits" are a measure of information in base 2, then "nats" is information in base e.

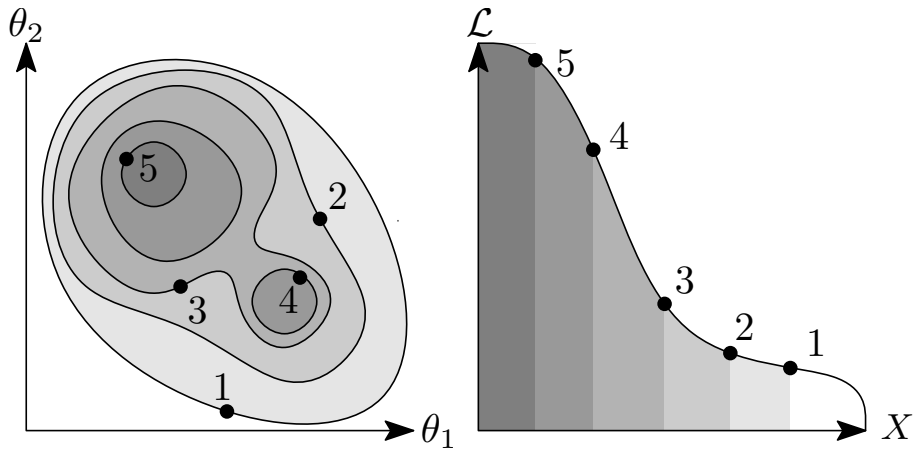


Figure 7.4: The nested sampling volume transformation. Left: five iso-likelihood contours of a two-dimensional multi-modal likelihood function $\mathcal{L}(\Theta)$. Each contour encloses some fraction of the prior X, indicated by colour. Right: Likelihood $\mathcal{L}$ as a function of the volume X enclosed by the contour. The evidence is the area under this curve.

7.6.1 The prior volume transformation $\mathcal{L}(X)$

The first trick is to transform (7.28) into an effectively one dimensional problem. We define an iso-likelihood contour as:

$$C(\mathcal{L}_*) = \{\Theta : \mathcal{L}(\Theta) = \mathcal{L}_*\}. \tag{7.34}$$

Each contour defined by $\mathcal{L}_*$ will also enclose a Θ-region containing some fraction of the prior:

$$X(\mathcal{L}_*) = \int_{\mathcal{L}(\Theta)>\mathcal{L}_*} \pi(\Theta)\,\mathrm{d}\Theta, \tag{7.35}$$

which is termed the *prior volume* by physicists and *prior mass* by mathematicians. For each contour C, there corresponds a likelihood and a prior volume. We may use this to re-write the likelihood as a function $\mathcal{L}(X)$,[9] and thus the integral (7.28) as:

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X)\ \mathrm{d}X. \tag{7.36}$$

This process is graphically depicted in Figure 7.4 for a two dimensional likelihood. Note the limits are from 0 to 1 since X denotes the *fraction* of the prior enclosed by the likelihood contour[10].

An inverse derivation is to examine (7.28), and define a new integration variable $\mathrm{d}X = \pi(\Theta)\,\mathrm{d}\Theta$. From this definition, the notion of prior volume (7.35) follows, which may be inverted to give $\mathcal{L}(X)$. Mathematically inclined readers may like to consider how this procedure is related to Lebesgue integration.

The beauty of this process is that it completely removes all complications such as geometry, topology and dimensionality, reducing (7.36) to a one dimensional integral. We may now evaluate the evidence using standard quadrature

[9] More rigorously, one may invert equation (7.35) since $X(\mathcal{L})$ is monotonic.
[10] Alternatively because X is a form of cumulative distribution function.

techniques:

$$\mathcal{Z} \approx \sum_i w_i \mathcal{L}(X_i), \tag{7.37}$$

where $\{X_i\}$ are the quadrature points and $\{w_i\}$ are the quadrature weights.

7.6.2 Computing the prior volumes

This is all well and good, but so far all we have done is transfer the complications (geometry, topology and dimensionality) from the evidence integral (7.28) to the integral which calculates the prior volume (7.35). John Skilling's true genius is to come up with a scheme that allows one to *estimate* the prior volume X in a quantifiable way.

The method relies on having a sampling procedure capable of producing a sample $\tilde{\Theta}$ distributed according to the prior π, subject to the constraint that it lies within the likelihood contour $C(\mathcal{L}_*)$. This is termed "sampling within a hard-edged likelihood constraint". In theory it should not be any more challenging than traditional sampling techniques such as Metropolis-Hastings (Skilling, 2006).

Importantly, a sample within $C(\mathcal{L}_*)$ distributed according to the prior π will have its prior volume X drawn *uniformly* within $[0, X_*]$, $X_* = X(\mathcal{L}_*)$. This demonstrates another reason why the prior volume is a powerful variable to work with.

7.6.3 Single-point nested sampling

If one draws an initial sample from $C(0)$ (i.e. the entire space) according to the prior π, one will have generated a point with prior volume X_1 and likelihood $\mathcal{L}_1$. On average, a single sample will cut the prior in two, and the mean $\langle X_1 \rangle = \frac{1}{2}$. At this value of X_1, it is unlikely that the likelihood is large enough to contribute to the evidence integral ($X_1 \gg e^{-H}$) so we should go deeper.

If one then generates X_2 by sampling from the contour defined by the previous point $C(\mathcal{L}_1)$, then $X_2 = t \times X_1$, where t is a uniform variable drawn from $[0, 1]$. We thus find that on average the second sample will have a mean $\langle X_2 \rangle = \frac{1}{4}$.

Repeating this process by using the previous contour as the hard likelihood constraint will mean that we sample within regions closer and closer to the peak, with $\langle X_i \rangle = 2^{-i}$. When $X_i \sim e^{-H}$, then the samples with these quadrature points start contributing to the evidence integral. Eventually, the samples compress too far ($X_i \ll e^{-H}$) and no longer contribute to the evidence, and the algorithm should stop.

More precisely, one finds that:

$$X_i = \prod_{j=1}^{i} t_j, \tag{7.38}$$

where each t_j is an independent random variable drawn uniformly from $[0, 1]$.

As i increases, X_i becomes distributed *log-normally* with:

$$\log X_i \approx -i \pm \sqrt{i}, \tag{7.39}$$

$$\mathrm{P}\left(X_i\right) = \frac{1}{\sqrt{2\pi i} X} \exp\left[-\frac{\left(\log X_i - i\right)^2}{2i}\right]. \tag{7.40}$$

This follows by taking logarithms of (7.38) and applying the central limit theorem. This is particularly impressive, since one finds that the samples compress the prior space *exponentially*. In general, the logarithmic compression from the prior to the posterior H tends to be linear in dimensionality d, so single point nested sampling scales linearly with dimensionality.

As this scheme progresses, one is left with a set of likelihoods $\mathcal{L}_i$, each paired with a volume X_i. These volumes are not deterministically known, but defined by the probabilistic definition (7.38). The evidence may be approximated as:

$$Z = \sum_{i=1}^{n} \mathcal{L}_i \times (X_{i-1} - X_i), \tag{7.41}$$

where for simplicity the quadrature weighting is taken as $w_i = X_{i-1} - X_i$. This means that the evidence is also implicitly probabilistic. Instead of computing an exact evidence, one may infer the distribution of evidence values that it could take.

In general, single point nested sampling will not produce a particularly accurate evidence estimation, but we can go one better.

7.6.4 Multi-point nested sampling

Instead of working with a single sample, one could initially generate a set of n_{live} "live points" distributed across the prior. The lowest likelihood point will have likelihood $\mathcal{L}_1$ and a volume $X_1 = t$ where t is distributed as the largest of n_{live} uniform samples in $[0, 1]$. Elementary probability shows that:

$$\mathrm{P}\left(t\right) = n_{\text{live}} t^{n_{\text{live}}-1}. \tag{7.42}$$

If this point is deleted, and replaced with a new point drawn from the contour defined by $\mathcal{L}_1$, then one will still have n_{live} points, now uniformly distributed uniformly in $[0, X_1]$. Proceeding onward, one finds that the volume becomes log-normally distributed:

$$\log X_i \approx -\frac{i}{n_{\text{live}}} \pm \sqrt{\frac{i}{n_{\text{live}}}}. \tag{7.43}$$

The prior space compression is slower than single point nested sampling, but with a correspondingly lower error. This therefore produces a more accurate inference on the evidence.

The deleted point is then added to a list of "dead points" which can then be used to make inferences on the evidence using equation (7.41).

7.6.5 Posterior inference

Nested sampling is designed to calculate the evidence, but it also produces posterior samples as a by-product of the calculation. Samples can be obtained

by sampling randomly under the posterior curve. Since we have partitioned the area into regions of size $\mathcal{L}_i w_i$, we can use this as an importance weighting.

Thus, dead points may be used as posterior samples with importance weighting:

$$p_i = \frac{\mathcal{L}_i w_i}{\mathcal{Z}}, \tag{7.44}$$

where the additional factor of $\mathcal{Z}$ merely ensures that they sum to 1 for consistency.

7.6.6 Terminating the algorithm

One does not wish to continue compressing the space once $X_i \ll e^{-H}$ for two reasons. The first reason is that this will not give any further accuracy on the evidence calculation. The second is more important. The region beneath $X_i \ll e^{-H}$ is not statistically relevant. Scientists used to maximum likelihood methods find this difficult to swallow, and often desire algorithms that will maximise the function to find the highest peak. In high dimensions, the absolute peak occupies such a vanishingly small percentage of the prior mass that it is totally irrelevant for the purposes of inference. In the case of a Gaussian likelihood, the peak value is a useful *summary statistic* but nothing more than that. Otherwise, one should not concern oneself with maximum likelihood, only in an accurate description of the peak provided by a set of samples. To paraphrase with a thermodynamic analogy: Frequentists like temperature, but Bayesians prefer heat.

As nested sampling proceeds, the likelihoods $\mathcal{L}_i$ monotonically increase, but the weights w_i monotonically decrease. This results in a peak in importance weights (7.44) that can be seen in Figure 7.5. We terminate the algorithm once the remaining posterior mass (white region) left in the live points is some small fraction of the currently calculated evidence (dark region). The posterior mass left in the live points at iteration i can be estimated by:

$$\mathcal{Z}_{\text{live}} \approx \langle \mathcal{L} \rangle_{\text{live}} X_i, \tag{7.45}$$

where the average is taken over the live points. Since this is typically an underestimate at early times, this will not cause premature termination.

7.6.7 Nested sampling: summary

- The algorithm initialises by sampling n_{live} points from the prior distribution $\pi(\Theta)$.
- At iteration i, the point with the lowest likelihood $\mathcal{L}_i$ is deleted, and then replaced by a new point, which is drawn from the prior subject to the constraint that its likelihood is greater than $\mathcal{L}_i$.
- The evidence may be inferred from the dead points via:

$$\mathcal{Z} = \sum_{i=1}^{n} w_i \mathcal{L}_i, \qquad X_i = \prod_{j=1}^{i} t_j,$$
$$w_i = X_{i-1} - X_i, \qquad t_j \sim U[0,1].$$

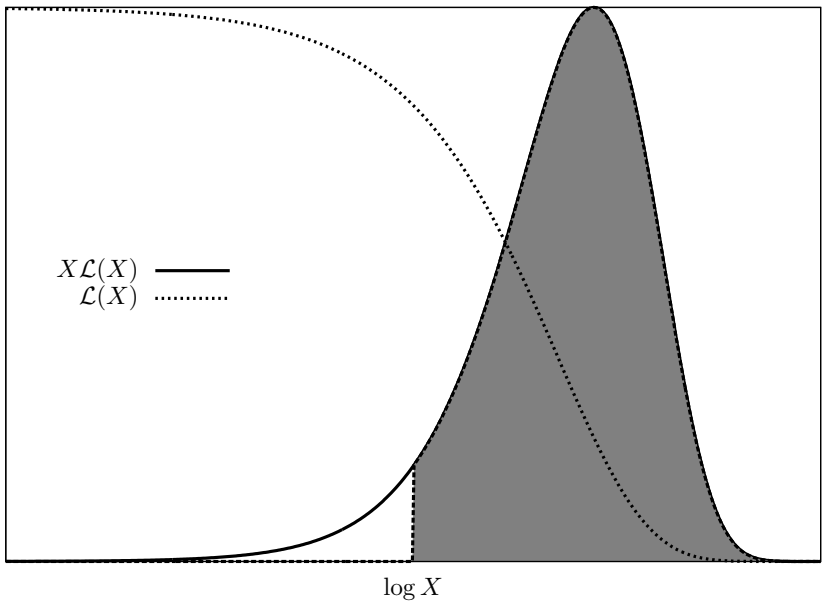


Figure 7.5: Plot of a generic likelihood as a function of the prior volume $\mathcal{L}(X)$. In high dimensions, the likelihood is only visible if plotted against $\log X$ (dashed curve). However, the evidence is better visualised by plotting $X\mathcal{L}(X)$ (solid curve). The area under the solid curve corresponds to the evidence. The magnitude of the solid curve is proportional to the importance weighting. Nested sampling proceeds from high to low volumes. After some time, the live points no longer contribute significantly to the evidence, and the algorithm terminates at this point.

- The algorithm terminates when:

$$X_i \times \langle \mathcal{L} \rangle_{\text{live}} \ll \mathcal{Z}.$$

- A set of posterior samples may be produced using the dead points with importance weighting:

$$p_i = \frac{w_i \mathcal{L}_i}{\mathcal{Z}}.$$

7.7 Conclusion

The following two chapters are concerned with describing my contributions to nested sampling theory and implementation in the algorithm PolyChord. After that I describe a novel numerical approach to solving differential equations which is relevant for my work in the field of quantum cosmology.

Appendix 7.A Traditional sampling methods

7.A.1 Inverse transform sampling

In the one-dimensional case, this amounts to converting a uniform random variable (which are easy to generate) into a variable sampled from a general distribution $f(\theta)$. One first finds its cumulative distribution function (CDF):

$$F(\theta) = \int_{-\infty}^{\theta} f(\theta')\,\mathrm{d}\theta', \tag{7.46}$$

computes the inverse of the CDF, and then applies this function to a uniform random variable $x \sim U(0,1)$ to generate a variable $\theta = F^{-1}(x)$, which is distributed according to $f(\theta)$.

In the general D-dimensional case $f = f(\theta) = f(\theta_1, \ldots, \theta_D)$ one calculates D conditional distributions $\{f_i : i = 1, \ldots, D\}$, by marginalising over parameters with indices greater than i and conditioning on parameters with indices less than i:

$$f_i(\theta_i|\theta_{i-1}, \ldots, \theta_1) = \frac{\int f_i(\theta) d\theta_{i+1} \ldots \mathrm{d}\theta_D}{\int f_i(\theta) d\theta_i \ldots \mathrm{d}\theta_D}. \tag{7.47}$$

Integrating these yields D conditional CDFs:

$$x_i = F_i(\theta_i|\theta_{i-1}, \ldots, \theta_1) = \int_{-\infty}^{\theta_i} f_i(\theta_i'|\theta_{i-1}, \ldots, \theta_1)\,\mathrm{d}\theta_i'. \tag{7.48}$$

Inverting this gives $\theta_i = F_i^{-1}(x_i|\theta_{i-1}, \ldots, \theta_1)$, which constitutes a set of relations sequentially transforming D uniform random variables $\{x_i\}$ into $\{\theta_i\}$ distributed according to $f(\theta)$.

7.A.2 Prior transformations

Here we give some concrete examples of inverse transform sampling. In the language of nested sampling, the D uniform random variables x_i are termed elements of the "unit hypercube", θ_i are elements of the "physical space", and the function f is the prior π on the physical parameters θ.

In many cases, the prior $\pi(\theta)$ is separable, and the above equations are easily calculated. For sections of the parameters which are not separable, the calculation can become more involved. We demonstrate it for a separable case, and a more complicated dependent case

To recap, we aim to compute the inverse of the functions F_i:

$$F_i(\theta_i|\theta_{i-1}, \ldots, \theta_0) = \int_{0}^{\theta_i} \pi_i(\theta_i'|\theta_{i-1}, \ldots, \theta_1)\,\mathrm{d}\theta_i', \tag{7.49}$$

where:

$$\pi_i(\theta_i|\theta_{i-1}, \ldots, \theta_0) = \frac{\int \pi_i(\theta) d\theta_{i+1} \ldots d\theta_N}{\int \pi_i(\theta) d\theta_i \ldots d\theta_N}, \tag{7.50}$$

and $\mathbf{F}$ maps from θ in the physical space onto the unit hypercube injectively.

Separable priors

A separable prior satisfies:

$$\pi(\theta) = \prod_i \pi_i(\theta_i). \tag{7.51}$$

This has the fortunate side effect that the functions F_i only depend on θ_i:

$$F_i(\theta_i|\theta_{i-1}, \ldots, \theta_0) = F_i(\theta_i). \tag{7.52}$$

Solving a separable prior thus amounts to solving a one-dimensional inverse-transform sampling problem. We demonstrate this procedure for two cases, a rectangular uniform prior, and a Gaussian prior.

Uniform prior

A rectangular uniform prior is defined by two parameters, $\theta_{\min}, \theta_{\max}$:

$$\pi(\theta) = \begin{cases} (\theta_{\max} - \theta_{\min})^{-1} & \text{for } \theta_{\max} < \theta_i < \theta_{\min} \\ 0 & \text{otherwise.} \end{cases} \tag{7.53}$$

Computing $F(\theta)$ we find:

$$F(\theta) = \int_{-\infty}^{\theta} \pi(\theta')\,\mathrm{d}\theta', \tag{7.54}$$

$$= \frac{\theta - \theta_{\min}}{\theta_{\max} - \theta_{\min}}, \tag{7.55}$$

with $F = 0$ or 1 either side of $\theta_{\min}$ and $\theta_{\max}$ respectively. Inverting the equation $F(\theta) = x$ we find:

$$\theta = \theta_{\min} + (\theta_{\max} - \theta_{\min})x, \tag{7.56}$$

is the transformation from x in the unit hypercube to θ in the physical space.

Gaussian prior

Defining a Gaussian prior with mean μ and standard deviation σ:

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{7.57}$$

we find that the procedure above yields:

$$\theta = \mu + \sqrt{2}\sigma \mathrm{erfinv}(2x - 1), \tag{7.58}$$

where erfinv is the conventional inverse error function.

Forced identifiability priors

As an example of a prior that is not separable in the parameters, we consider a forced identifiability prior. Here, n parameters are distributed uniformly between $\theta_{\min}$ and $\theta_{\max}$, but subject to the constraint that they are ordered numerically. This is a particularly useful prior in the reconstruction of functions

using a spline with movable knots (Vázquez et al., 2012a, Aslanyan et al., 2014, Abazajian et al., 2014, Planck Collaboration et al., 2016b). In this case, the horizontal locations of the knots must be ordered.

The required prior is uniform in the hyper-triangle defined by $\theta_{\min} < \theta_1 < \cdots < \theta_n < \theta_{\max}$, and zero everywhere else:

$$\pi(\theta) = \begin{cases} \frac{1}{n!(\theta_{\max}-\theta_{\min})^n} & \text{for } \theta_{\min} < \theta_1 < \cdots < \theta_n < \theta_{\max} \\ 0 & \text{otherwise.} \end{cases} \tag{7.59}$$

To calculate equations (7.49) & (7.50) we integrate over the constant distribution, taking care with the limits. We find:

$$\pi_i(\theta_i|\theta_{i-1},\ldots,\theta_0) = \frac{(n-i+1)(\theta_i-\theta_{i-1})^{n-i}}{(\theta_{\max}-\theta_{\min})^{n-i+1}}, \tag{7.60}$$

$$F_i(\theta_i|\theta_{i-1},\ldots,\theta_0) = \left(\frac{\theta_i-\theta_{i-1}}{\theta_{\max}-\theta_{i-1}}\right)^{n-i+1}, \tag{7.61}$$

where for consistency we define $\theta_0 = \theta_{\min}$. Hence solving $x_i = F(\theta_i|\theta_{i-1},\ldots,\theta_0)$ for θ_i we find:

$$\theta_i = \theta_{i-1} + (\theta_{\max}-\theta_{i-1})x_i^{1/(n-i+1)}. \tag{7.62}$$

This enables $\{\theta_i\}$ to be calculated sequentially from $\{x_i\}$. We may interpret this transformation as θ_i being distributed as the smallest of $n-i+1$ uniformly distributed variables in the range $[\theta_{i-1}, \theta_{\max}]$.

7.A.3 Rejection sampling

The principle behind rejection sampling is simple, and best demonstrated graphically as in Figure 7.6. If one requires some samples from a complicated distribution $f(x)$, but has the knowledge that it satisfies $f(x) < g(x)$ for some simpler distribution $g(x)$, then one may sample from g, and accept only those samples with probability less than f.

Obviously this will be very inefficient unless $g(x)$ happens to be rather close to $f(x)$. One finds that in general this inefficiency is exaggerated by the curse of dimensionality.

7.A.4 Metropolis—Hastings

Metropolis—Hastings approaches are an extremely widely used methodology for generating a sequence of samples (or points) from some posterior $\mathcal{P}(x)$.

A sequence of T points is termed a *chain* $\mathcal{C}_T$:

$$\mathcal{C}_T = \{x^{(t)} : t = 1\cdots T\}. \tag{7.63}$$

A new point $x^{(t+1)}$ is generated from a proposal density $\mathcal{Q}$ which depends on the current state $x^{(t)}$. Typically $\mathcal{Q}(x|x^{(t)})$ might be a Gaussian distribution centred on the current value of $x^{(t)}$:

$$\log \mathcal{Q}(x|x^{(t)}) = \text{const.} - \frac{\left[x-x^{(t)}\right]^2}{2\varepsilon^2}, \tag{7.64}$$

but in general the proposal density can be any fixed density from which we may draw samples easily.

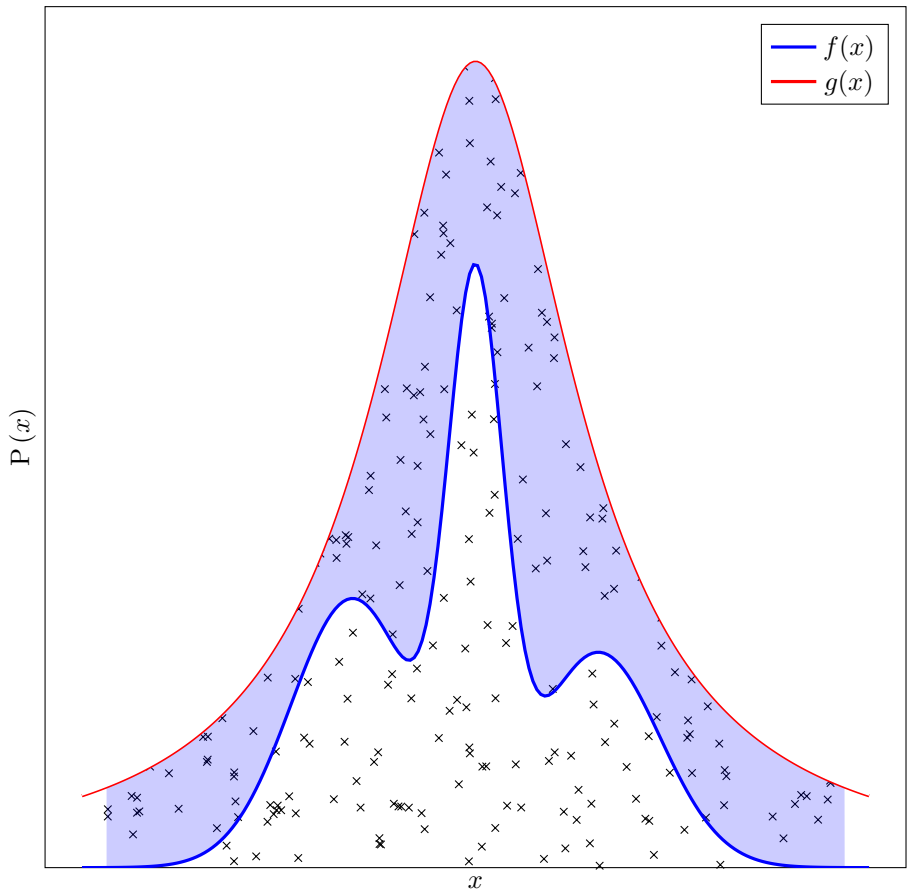


Figure 7.6: Rejection sampling. One may produce samples from some complicated distribution $f(x)$ using a simpler distribution $g(x)$ provided that $g(x) > f(x)$ within the domain of interest. One samples from g, and accepts only those samples that lie under the f curve.

A new state x is proposed from the proposal density $\mathcal{Q}(x|x^{(t)})$, and it is accepted with probability:

$$P_{\text{accept}}(x|x^{(t)}) = \frac{\mathcal{P}(x)}{\mathcal{P}(x^{(t)})}\frac{\mathcal{Q}(x^{(t)}|x)}{\mathcal{Q}(x|x^{(t)})}. \tag{7.65}$$

If the step is accepted, then $x^{(t+1)} = x$, otherwise $x^{(t+1)} = x^{(t)}$. This procedure can be seen schematically in Figure 7.7.

This has the distinct advantage that it does not require one to take into account the overall normalisation of the posterior, circumnavigating any issues associated with computing the evidence $\mathcal{Z}$, since:

$$\frac{\mathcal{P}(x)}{\mathcal{P}(x^{(t)})} \equiv \frac{\mathcal{L}(x)\pi(x)}{\mathcal{L}(x^{(t)})\pi(x^{(t)})}. \tag{7.66}$$

It is important to note that the chain C_T will comprise T samples which are *correlated*. Consider Figure 7.7 once more, where the shortest and longest

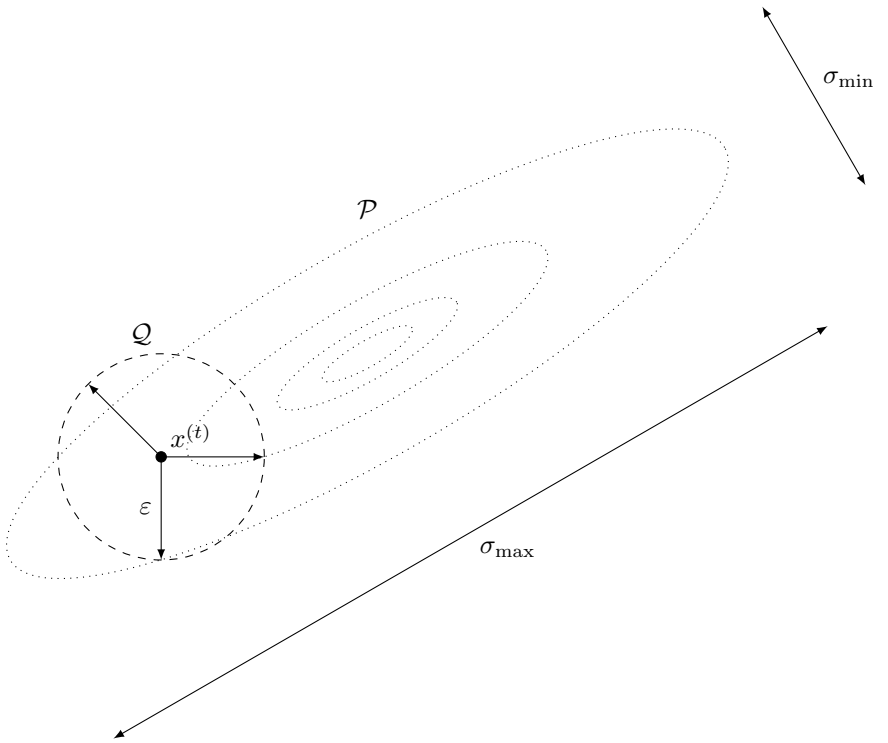


Figure 7.7: Metropolis-Hastings. Here we have some degenerate two-dimensional posterior $\mathcal{P}$, and a circular proposal distribution $\mathcal{Q}$.

lengthscales of the probability region are $\sigma_{\min}$ and $\sigma_{\max}$ respectively. If the proposal distribution amounts to a random walk with step size of order the shortest lengthscale $\varepsilon \sim \sigma_{\min}$, then the diffusive Brownian nature of a random walk leads one to expect the timescale to fully traverse the region will be:

$$\tau \sim \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^2 . \tag{7.67}$$

It is this timescale which determines the degree of correlation between samples within a chain C_T.

In theory, given that the relation (7.67) has no dimensionality attached to it, Metropolis Hastings can be extremely successful in high dimensions. However, there are several issues.

First, one does not initially a-priori have a good starting point. There is therefore a period of "burn-in" attached to MH methods. Knowing when this period is truly over, and one is starting to generate random samples can be challenging in the general case.

In many cases, $\tau \gg 1$ resulting in unacceptable run-times. This can be ameliorated by choosing a proposal distribution $\mathcal{Q}$ that better agrees with the posterior $\mathcal{P}$. However, this amounts to adding in many effective tuning parameters. Many algorithms work by having a "learning" phase, whereby the algorithm deduces for itself a good correlated Gaussian proposal distribution, but given the fact that this is not strictly Markovian care must be taken.

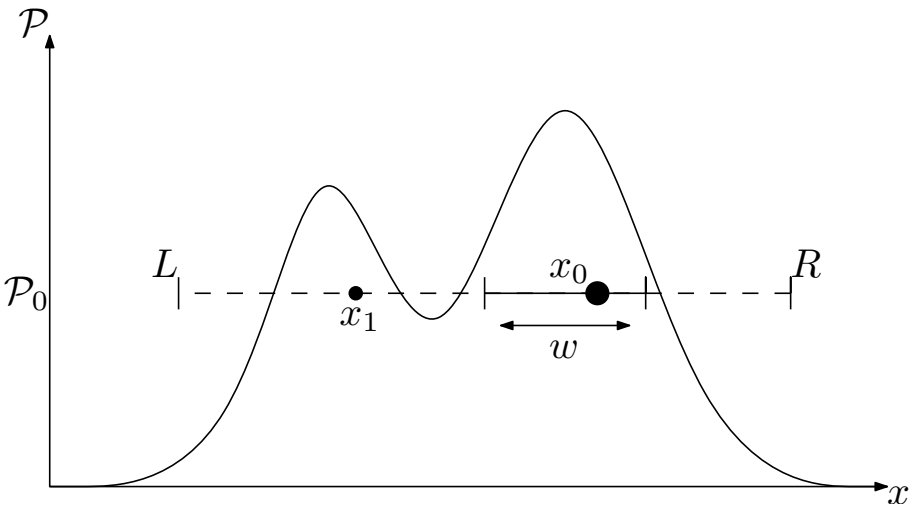


Figure 7.8: Slice sampling in one dimension. Given a probability level (or slice) $\mathcal{P}_0$, slice sampling samples within the horizontal region defined by $\mathcal{P} > \mathcal{P}_0$. From an initial point x_0 within the slice ($\mathcal{P}(x_0) > \mathcal{P}_0$), a new point x_1 is generated within the slice with a distribution $\mathrm{P}(x_1|x_0)$. External bounds are first set on the slice $\hat{L} < x_0 < \hat{R}$ by uniformly expanding a random initial bound of width w until they lie outside the slice (Neal terms this the *stepping out* procedure). x_1 is then sampled uniformly within these bounds. If x_1 is not in the slice, then $\hat{L}$ or $\hat{R}$ is replaced with x_1, ensuring that x_0 is still within the slice. This procedure is guaranteed to generate a new point x_1, and satisfies detailed balance $\mathrm{P}(x_0|x_1) = \mathrm{P}(x_1|x_0)$. Thus, if x_0 is drawn from a uniform distribution within the slice, so is x_1.

7.A.5 Slice sampling

Radford Neal initially proposed slice sampling as an effective methodology for generating samples numerically from a given posterior $\mathcal{P}(\Theta)$. One first chooses a 'slice' (or probability level) $\mathcal{P}_0$ uniformly within $[0, \mathcal{P}_{\max}]$. One then samples uniformly within the Θ-region defined by $\mathcal{P}(\Theta) > \mathcal{P}_0$. The similarity with the iso-likelihood contour sampling required by nested sampling should be clear. In the one-dimensional case, he suggests the sampling procedure detailed in Figure 7.8.

In higher dimensions, Neal (2000) suggests a variety of MCMC-like methods. The simplest of these is implemented by sampling each of the parameter directions in turn. Since each one-dimensional slice requires $\sim \mathcal{O}$(a few) likelihood calculations, the number of likelihood calculations required scales linearly with dimensionality, providing the region is efficiently navigated.

Multi-dimensional slice sampling has many of the benefits of a traditional MH approach, and uses a proposal distribution which is much more efficient at sampling a hard likelihood constraint. This makes is a very suitable candidate for use within a nested sampling algorithm (Skilling, 2006).

Problems may occur with slice sampling if the peaks in Figure 7.8 are separated by a large distance. In this case it is still theoretically true that procedure satisfies detailed balance, and thus the algorithm will eventually

jump one from one peak to the other. Practically however, this could take many Hubble times, and to all intents and purposes the distribution is not explored correctly. It is worth noting that PolyChord (Chapter 9) circumvents this difficulty by explicitly evolving posterior modes separately.

7.A.6 Thermodynamic integration

The traditional method of computing the evidence from MCMC procedures is via *thermodynamic integration*. The aim is to compute:

$$\mathcal{Z} = \int \mathcal{L}(\Theta)\pi(\Theta)\,\mathrm{d}\Theta. \tag{7.68}$$

Inspired by thermodynamics, we define the posterior $\mathcal{P}_\beta \propto \mathcal{L}^\beta\pi$ at temperature β as:

$$\mathcal{P}_\beta(\Theta) = \frac{\mathcal{L}^\beta(\Theta)\pi(\Theta)}{\mathcal{Z}_\beta}, \tag{7.69}$$

$$\mathcal{Z}_\beta = \int \mathcal{L}^\beta(\Theta)\pi(\Theta)\,\mathrm{d}\Theta. \tag{7.70}$$

If the likelihood is interpreted as a Boltzmann-like probability distribution, then:

$$\mathcal{L}(\Theta)^\beta = \exp[-\beta E(\Theta)], \tag{7.71}$$

is the Boltzmann probability of finding a system with inverse temperature $\beta = T^{-1}$ in state Θ with energy E, and the partition function is $\mathcal{Z}_\beta$.

Inspired by thermodynamics, one can see that:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\log\mathcal{Z}_\beta = \frac{\int \mathcal{L}^\beta\pi\log\mathcal{L}\,\mathrm{d}\Theta}{\int \mathcal{L}^\beta\pi\,\mathrm{d}\Theta} = \int \mathcal{P}_\beta\log\mathcal{L}\,\mathrm{d}\Theta = \langle\log\mathcal{L}\rangle_\beta. \tag{7.72}$$

Hence, integrating the above equation yields:

$$\log\mathcal{Z}_1 - \log\mathcal{Z}_0 = \log\mathcal{Z} = \int_0^1 \langle\log\mathcal{L}\rangle_\beta\,\mathrm{d}\beta. \tag{7.73}$$

The key point is that we may compute the mean $\langle\log\mathcal{L}\rangle_\beta$ from a set of samples from the posterior $\mathcal{P}_\beta$. If we therefore generate a set of samples for a range of temperatures β, the one-dimensional integral above may be computed easily. Thermodynamic calculation of the evidence therefore is equivalent to running multiple MCMC runs at several temperatures. This however suffers from the disadvantage that it is difficult to produce an estimate of the error in computing (7.73).

Chapter 8

Extending Nested Sampling

In this chapter, I describe some of my analytical contributions to the theory of nested sampling (detailed in Chapter 7). In particular, the PolyChord algorithm (Chapter 9) requires the ability to handle a variable number of live points, and the capacity to evolve separate posterior modes semi-independently. These analytics detail how one can keep track of inferences on evolving prior volumes and evidences. Some of this work was published in an appendix in Handley et al. (2015b).

8.1 Evidence estimates and errors

Skilling (2006) initially advocated using Monte-Carlo methods to estimate the evidence error, although this requires the storage of the entire chain of dead points, rather than just the subset usually stored for posterior inferences. For high-dimensional problems, the number of dead points is prohibitively large, and cannot be stored. Feroz et al. (2009) use an alternative method based on the relative entropy (also suggested by Skilling, 2006). Keeton (2011) suggests a more intuitive methodology of estimating the error, and it is this which we use, although it must be heavily adapted for the case of variable numbers of live points and clustering.

8.1.1 Basic theory

We wish to compute the sum:

$$\mathcal{Z} = \sum_i (X_{i-1} - X_i)\mathcal{L}_i. \tag{8.1}$$

However, we do not know the volumes X_i exactly, so we can only make inferences about $\mathcal{Z}$, in terms of a probability distribution $\mathrm{P}(\mathcal{Z})$. In practice, all we need to compute is the mean and variance of this distribution:

$$\mathrm{mean}(\mathcal{Z}) \equiv \overline{\mathcal{Z}}, \tag{8.2}$$

$$\mathrm{var}(\mathcal{Z}) \equiv \overline{\mathcal{Z}^2} - \overline{\mathcal{Z}}^2. \tag{8.3}$$

At iteration i, the n_{live} live points are each uniformly sampled within a contour of volume X_{i-1}. The volume X_i will be the largest volume out of n_{live} uniform

volume samples in volume X_i. Thus X_i satisfies the recursion relation:

$$X_i = tX_{i-1}, \qquad X_0 = 1, \tag{8.4}$$

$$\mathrm{P}(t) = n_{\mathrm{live}} t^{n_{\mathrm{live}}-1}, \tag{8.5}$$

where the t and X_{i-1} are independent.

It is worth noting that the procedure described below will generate the mean and variance of the distribution, but in fact this is not quite what we want. The evidence is in practice approximately log-normally distributed. Thus, it is better to report the mean and variance of $\log \mathcal{Z}$, defined by:

$$\mathrm{mean}(\log \mathcal{Z}) = 2 \log \overline{\mathcal{Z}} - \frac{1}{2} \log \overline{\mathcal{Z}^2}, \tag{8.6}$$

$$\mathrm{var}(\log \mathcal{Z}) = \log \overline{\mathcal{Z}^2} - 2 \log \overline{\mathcal{Z}}, \tag{8.7}$$

where the above relations are true for any log-normally distributed variable $\mathcal{Z}$.

8.1.2 Computing the mean evidence

While it is possible to take equations (8.1), (8.4) & (8.5) and compute the mean as a general formula (Keeton, 2011), in the case of clustering this is uninformative. In fact, for large-dimensional spaces using the full formula would require storage of a prohibitively large amount of data. The calculation is better accomplished by a set of recursion relations, which update the mean evidence and its error at each step.

For now, assume that we have n live points currently enclosed by some likelihood contour $\mathcal{L}$ of volume X, and $\mathcal{Z}$ is the last value of the evidence calculated from all of the points that have died so far. By considering (8.1), (8.4) & (8.5), when we kill off the outermost point, we may adjust the values of $\mathcal{Z}$ and X using:

$$\mathcal{Z} \to \mathcal{Z} + (1-t)X\mathcal{L}, \tag{8.8}$$

$$X \to tX. \tag{8.9}$$

Taking the mean of these relations, we may use the facts that t and X are independent random variables and that $\mathrm{P}(t) = nt^{n-1}$, to find the recursion relations:

$$\overline{\mathcal{Z}} \to \overline{\mathcal{Z}} + \frac{1}{n+1}\overline{X}\mathcal{L}, \tag{8.10}$$

$$\overline{X} \to \frac{n}{n+1}\overline{X}. \tag{8.11}$$

8.1.3 Computing the evidence error

To estimate $\overline{\mathcal{Z}^2}$, we square (8.8) and (8.9) and multiply both together to obtain:

$$\mathcal{Z}^2 \to \mathcal{Z}^2 + 2(1-t)\mathcal{Z}X\mathcal{L} + (1-t)^2 X^2 \mathcal{L}^2, \tag{8.12}$$

$$\mathcal{Z}X \to t\mathcal{Z}X + t(1-t)X^2\mathcal{L}, \tag{8.13}$$

$$X^2 \to t^2 X^2. \tag{8.14}$$

Note that we now need to keep track of the variable $\mathcal{Z}X$, as these two are not independent. Taking the averages of the above yields:

$$\overline{\mathcal{Z}^2} \to \overline{\mathcal{Z}^2} + \frac{2\overline{\mathcal{Z}X}\mathcal{L}}{n+1} + \frac{2\overline{X^2}\mathcal{L}^2}{(n+1)(n+2)}, \tag{8.15}$$

$$\overline{\mathcal{Z}X} \to \frac{n\overline{\mathcal{Z}X}}{n+1} + \frac{n\overline{X^2}\mathcal{L}}{(n+1)(n+2)}, \tag{8.16}$$

$$\overline{X^2} \to \frac{n}{n+2}\overline{X^2}. \tag{8.17}$$

8.1.4 The full calculation

There are therefore five quantities to keep track of:

$$\overline{\mathcal{Z}}, \quad \overline{\mathcal{Z}^2}, \quad \overline{\mathcal{Z}X}, \quad \overline{X}, \quad \overline{X^2}.$$

These should be initialised at $\{0, 0, 0, 1, 1\}$ respectively, and updated using equations (8.10), (8.12), (8.13), (8.11) & (8.14) in that order. In fact, we keep track of the logarithm of these quantities, in order to avoid machine precision errors.

8.2 Evidence estimates and errors in clusters

Within the context of nested sampling there are many examples where the live points naturally partition into distinct *clusters*. The most obvious example is when the posterior has multiple spatial modes[1]. Other examples include clustering live points into sub-models of a multi-model sampler or clustering into groups that allow for easier posterior exploration.

In all of these cases, it is useful to keep track of the evidence and volume of each of these clusters. This is helpful both for the purposes of posterior analysis, and for maintaining the stability and accuracy of evidence calculation. In this section, we build the analytics that allow one to do compute these quantities. The analysis follows that of Section 8.1.

Throughout the algorithm, there will in general be m identified clusters. In doing so, we wish to keep track of the volume of each cluster $\{X_1, \ldots, X_m\}$, the global evidence and its error $\mathcal{Z}, \mathcal{Z}^2$ and the local evidences and their errors $\{\mathcal{Z}_1, \mathcal{Z}_1^2 \ldots, \mathcal{Z}_m, \mathcal{Z}_m^2\}$. At each iteration, the point with the lowest likelihood $\mathcal{L}$ will be killed from cluster p, $(1 \leq p \leq m)$.

8.2.1 Evidence

We thus need to update the global evidence, the local evidence of cluster p, and the volume of cluster p:

$$\mathcal{Z} \to \mathcal{Z} + (1-t)X_p\mathcal{L}, \tag{8.18}$$

$$\mathcal{Z}_p \to \mathcal{Z}_p + (1-t)X_p\mathcal{L}, \tag{8.19}$$

$$X_p \to tX_p. \tag{8.20}$$

[1]The ability that the live points have to distinguish posterior modes is indeed one of the strengths of nested sampling, and is one of the primary reasons for the successful adoption of the MultiNest algorithm (Feroz and Hobson, 2008, Feroz et al., 2009, 2013) by the scientific community.

Since t will be distributed with $\mathrm{P}(t) = n_p t^{n_p - 1}$, taking the mean of these yields:

$$\overline{\mathcal{Z}} \to \overline{\mathcal{Z}} + \frac{\overline{X}_p \mathcal{L}}{n_p + 1}, \tag{8.21}$$

$$\overline{\mathcal{Z}}_p \to \overline{\mathcal{Z}}_p + \frac{\overline{X}_p \mathcal{L}}{n_p + 1}, \tag{8.22}$$

$$\overline{X}_p \to \frac{n_p \overline{X}_p}{n_p + 1}. \tag{8.23}$$

Keeping track of $\{\overline{\mathcal{Z}}, \overline{\mathcal{Z}_p}, \overline{X_p}, p = 1 \ldots m\}$ and updating them using the recursion relations in the order above will produce a consistent evidence estimate for both the local and global evidence errors.

8.2.2 Evidence errors

We must also keep track of the local and global evidence errors. Taking the square of equations (8.18) & (8.19) yields:

$$\mathcal{Z}^2 \to \mathcal{Z}^2 + 2(1-t)\mathcal{Z} X_p \mathcal{L} + (1-t)^2 X_p^2 \mathcal{L}^2, \tag{8.24}$$

$$\mathcal{Z}_p^2 \to \mathcal{Z}_p^2 + 2(1-t)\mathcal{Z}_p X_p \mathcal{L} + (1-t)^2 X_p^2 \mathcal{L}^2. \tag{8.25}$$

We can see that we are going to need to keep track of $\{\overline{\mathcal{Z}X_p}, \overline{\mathcal{Z}_p X_p}, \overline{X_p^2}\}$ in addition to $\{\overline{\mathcal{Z}^2}, \overline{\mathcal{Z}_p^2}\}$. Taking various multiplications of equations (8.18), (8.19) & (8.20) one finds:

$$\mathcal{Z} X_p \to t\mathcal{Z} X_p + (1-t) t X_p^2 \mathcal{L}, \tag{8.26}$$

$$\mathcal{Z} X_q \to \mathcal{Z} X_q + (1-t) X_p X_q \mathcal{L} \qquad (p \neq q), \tag{8.27}$$

$$\mathcal{Z}_p X_p \to t\mathcal{Z}_p X_p + (1-t) t X_p^2 \mathcal{L}, \tag{8.28}$$

$$X_p^2 \to t^2 X_p^2, \tag{8.29}$$

$$X_p X_q \to t X_p X_q. \tag{8.30}$$

Taking the mean of the above yields the recursion relations:

$$\overline{\mathcal{Z}^2} \to \overline{\mathcal{Z}^2} + \frac{2\overline{\mathcal{Z}X_p}\mathcal{L}_p}{n_p+1} + \frac{2\overline{X_p^2}\mathcal{L}^2}{(n_p+1)(n_p+2)}, \tag{8.31}$$

$$\overline{\mathcal{Z}_p^2} \to \overline{\mathcal{Z}_p^2} + \frac{2\overline{\mathcal{Z}_pX_p}\mathcal{L}}{n_p+1} + \frac{2\overline{X_p^2}\mathcal{L}^2}{(n_p+1)(n_p+2)}, \tag{8.32}$$

$$\overline{\mathcal{Z}X_p} \to \frac{n_p\overline{\mathcal{Z}X_p}}{n_p+1} + \frac{n_p\overline{X_p^2}\mathcal{L}}{(n_p+1)(n_p+2)}, \tag{8.33}$$

$$\overline{\mathcal{Z}X_q} \to \overline{\mathcal{Z}X_p} + \frac{\overline{X_pX_q}\mathcal{L}}{(n_p+1)} \qquad (q \neq p), \tag{8.34}$$

$$\overline{\mathcal{Z}_pX_p} \to \frac{n_p\overline{\mathcal{Z}_pX_p}}{n_p+1} + \frac{n_p\overline{X_p^2}\mathcal{L}}{(n_p+1)(n_p+2)}, \tag{8.35}$$

$$\overline{X_p^2} \to \frac{n_p\overline{X_p^2}}{n_p+2}, \tag{8.36}$$

$$\overline{X_pX_q} \to \frac{n_p\overline{X_pX_q}}{n_p+1}. \tag{8.37}$$

Keeping track of

$$\{\overline{\mathcal{Z}^2}, \overline{\mathcal{Z}_p^2}, \overline{\mathcal{Z}X_p}, \overline{\mathcal{Z}_pX_p}, \overline{X_p^2}, \overline{X_pX_q}, p, q = 1 \ldots m\},$$

and updating them using the recursion relations in the order above will produce a consistent estimate for the local and global evidence errors.

8.2.3 Cluster initialisation

All that remains is to initialise the clusters correctly at the point of creation.

The starting initialisation of the evidence and volume is reasonable, there will be only a single cluster with volume 1, and all evidence related terms 0. At some point (possibly at the beginning, depending on the prior), the live points will split into distinct clusters, and the local volumes and evidences will need to be re-initialised.

At the point of splitting a cluster into sub-clusters, we partition the n live points into a N new clusters, with $\{n_1, \ldots, n_N\}$ live points in each. If the volume of the splitting cluster is X_p initially, we need to know how to partition this volume into $\{X_1, \ldots, X_N\}$. If the points are drawn uniformly from the volume, then the n_i will depend on the volumes via a multinomial probability distribution:

$$\mathrm{P}\left(\{n_i\}|X_p, \{X_i\}\right) \propto {X_1}^{n_1} \ldots X_N^{n_N}. \tag{8.38}$$

We however want to know the probability distributions of the $\{X_i\}$, given the $\{n_i\}$. We can invert the above with Bayes' theorem, using an (improper) logarithmic prior on the volumes subject to the constraint that they sum to X_p:

$$\mathrm{P}\left(\{X_i\}|X_p\right) \propto \frac{\delta(X_1 + \cdots + X_N - X_p)}{X_1 \cdots X_N}. \tag{8.39}$$

Doing this shows the posterior $\mathrm{P}\left(\{X_i\}|X_p, \{n_i\}\right)$ is a Dirichlet distribution with parameters $\{n_i\}$. More importantly, we can use this to compute the means and

correlations for the volumes $\{X_i\}$:

$$\overline{X_i} = \frac{n_i}{n}\overline{X}_p, \tag{8.40}$$

$$\overline{X_i^2} = \frac{n_i(n_i+1)}{n(n+1)}\overline{X_p^2}, \tag{8.41}$$

$$\overline{X_i X_j} = \frac{n_i n_j}{n(n+1)}\overline{X_p^2}, \tag{8.42}$$

$$\overline{X_i Y} = \frac{n_i}{n}\overline{X_p Y} \qquad Y \in \{Z, Z_p, X_q\}. \tag{8.43}$$

The first equation recovers the intuitive result that the volume should split as the fraction of live points. Note, however that this requires a logarithmic prior. The third shows us that since $\overline{X_i X_j} \neq \overline{X_i}\ \overline{X_j}$, the volumes are correlated at the splitting. This is to be expected.

We also need to initialise the local evidences and their errors. A consistent approach is to assume that the evidences also split in proportion to the cluster distribution of live points. Following the same reasoning as above, we find that:

$$\overline{Z_i} = \frac{n_i}{n}\overline{Z}_p, \tag{8.44}$$

$$\overline{Z_i X_i} = \frac{n_i(n_i+1)}{n(n+1)}\overline{Z_p X_p}, \tag{8.45}$$

$$\overline{Z_i^2} = \frac{n_i(n_i+1)}{n(n+1)}\overline{Z_p^2}. \tag{8.46}$$

Thus, at cluster splitting, all of the new local evidences, volumes and cross correlations are initialised according to the above.

This completes the mechanism for keeping track of the local and global evidences, their errors, and the local cluster volumes.

8.3 Further generalisations

It is worth noting that the procedures in the previous sections are amenable to a great many generalisations. We give a few examples in order to demonstrate this.

First, if one desires a higher order of accuracy, one may update the rather simple quadrature in equation (8.1) to:

$$\mathcal{Z} = \sum_i (X_{i-1} - X_i) \times \frac{1}{2}(\mathcal{L}_i + \mathcal{L}_{i-1}). \tag{8.47}$$

In this case, the approach is identical to the method described above, one merely replaces the likelihood value with that of the average value of the last two likelihoods $(\mathcal{L}_i + \mathcal{L}_{i-1})/2$. It is better to phrase the trapezoidal rule as in equation (8.47) as opposed to the more traditionally cited:

$$\mathcal{Z} = \sum_i \frac{1}{2}(X_{i-1} - X_{i+1})\mathcal{L}_i. \tag{8.48}$$

Equations (8.47) and (8.48) are equivalent, but the first formulation can be calculated at run time, whereas the second requires knowledge of future values of X_i, which tends to result in programs with a lot of awkward bookkeeping.

Second, if one wishes to re-phrase the evidence integral:

$$\mathcal{Z} = \int \mathcal{L} \, \mathrm{d}X = \int X\mathcal{L} \, \mathrm{dlog}\, X, \tag{8.49}$$

then the trapezoidal quadrature must be modified:

$$\mathcal{Z} = \sum_i (\log X_{i-1} - \log X_i) \times \frac{1}{2}(\mathcal{L}_i X_i + \mathcal{L}_{i-1} X_{i-1}). \tag{8.50}$$

This therefore amounts to an update of:

$$\mathcal{Z} \to \mathcal{Z} + X \times \frac{1}{2}\left(t \log \frac{1}{t}\mathcal{L}_i + \log \frac{1}{t}\mathcal{L}_{i-1}\right), \tag{8.51}$$

$$X \to tX. \tag{8.52}$$

The analysis then proceeds exactly as before. In general, when playing with various formulations, it suffices to know the averages:

$$\left\langle t^a (1-t)^b \right\rangle = n \frac{b!(a+n-1)!}{(a+b+n)!}, \qquad \left\langle t^a (\log 1/t)^b \right\rangle = \frac{n\, b!}{(n+a)^{b+1}}. \tag{8.53}$$

Chapter 9

PolyChord

> Exploring a hard-edged likelihood-constrained domain should prove to be neither more nor less demanding than exploring a likelihood-weighted space.
>
> John Skilling (2006)

PolyChord is a novel nested sampling algorithm tailored for high-dimensional parameter spaces. This chapter provides an extensive account of the algorithm. PolyChord utilises slice sampling at each iteration to sample within the hard likelihood constraint of nested sampling. It can identify and evolve separate modes of a posterior semi-independently, and is parallelised using openMPI. It is capable of exploiting a hierarchy of parameter speeds such as those present in CosmoMC (Lewis and Bridle, 2002) and CAMB (Lewis et al., 2000), and is now in use in the CosmoChord and ModeChord (Mortonson et al., 2011, Easther and Peiris, 2012, Norena et al., 2012) codes.

9.1 Introduction

Over the past two decades, Bayesian methods have been increasingly adopted as the standard inference procedure for the rapidly increasing volume of astrophysical data.

Bayesian inference consists of *parameter estimation* and *model comparison*. Parameter estimation is generally performed using Markov-Chain Monte-Carlo (MCMC) methods, such as the Metropolis-Hastings (MH) algorithm and its variants (MacKay, 2002). In order to perform model comparison, one must calculate the *evidence*: a high-dimensional integration of the likelihood over the prior density (Sivia and Skilling, 2006). MH methods cannot compute this on a usable timescale, hindering the use of Bayesian model comparison in cosmology and astroparticle physics.

A contemporary methodology for computing evidences and posteriors simultaneously is provided by nested sampling (Skilling, 2006). This has been successfully implemented in the now widely adopted algorithm MultiNest

(Feroz and Hobson, 2008, Feroz et al., 2009, 2013). Modern cosmological likelihoods now involve a large number of parameters, with a hierarchy of speeds. MultiNest struggles with high-dimensional parameter spaces, and is unable to take advantage of this separation of speeds. PolyChord aims to address these issues, providing a means to sample high-dimensional spaces across a hierarchy of parameter speeds.

The layout of the chapter is as follows: We overview the historical implementations of nested sampling in Section 9.2 and provide an account of the slice sampling technique of Neal (2000). We describe the PolyChord algorithm in detail in Section 9.3 and demonstrate its efficacy on toy and cosmological problems in Section 9.4. Section 9.5 concludes the chapter.

This chapter is an extensive overview of our algorithm, which is now in use in several cosmological applications (Planck Collaboration et al., 2016b). It was published as Handley et al. (2015b). A briefer introduction can be found in Handley et al. (2015a).

PolyChord is available for download from the link at the end of the chapter.

9.2 Sampling within an iso-likelihood contour

9.2.1 The unit hypercube

Each iteration of nested sampling requires one to sample from the prior (subject to a hard likelihood constraint). Typically, priors are defined in terms of simple analytic functions such as uniform or Gaussian distributions, and may be sampled using inverse transform sampling (Appendix 7.A.1).

Nested sampling can thus be performed in the unit D-dimensional hypercube. This has numerous advantages, the first being that one only needs to be able to generate uniform random variables in $[0, 1]$. The second is more subtle; it is more natural to define a distance metric in the unit hypercube than in the physical space. Unit hypercube variables all have the same dimensionality: probability.

9.2.2 Previous methods

The most challenging aspect of nested sampling is drawing a new point from the prior subject to the hard likelihood constraint $\mathcal{L} > \mathcal{L}_i$. This may be done in a variety of ways, and distinguishes the various historical implementations.

For some problems, the iso-likelihood contour is known analytically, allowing one to construct a sampling procedure specific to that problem. This is demonstrated by Keeton (2011), and can be useful for testing nested sampling's theoretical behaviour. In most cases, however, the likelihood contour is unknown a-priori, so a more numerical approach must be taken.

Mukherjee et al. (2006) implemented a rejection sampling (Appendix 7.A.3) method, which was extended by Shaw et al. (2007) and later incorporated into the widely-used MultiNest algorithm (Feroz and Hobson, 2008, Feroz et al., 2009, 2013). These algorithms sample by using the live points to construct a set of intersecting ellipsoids which together aim to enclose the likelihood contour, and then performs rejection sampling within the ellipsoids. Whilst being

an excellent algorithm for modest numbers of parameters, any rejection sampling algorithm has an exponential scaling with dimensionality that eventually emerges.

An alternative approach (the one initially envisaged by Skilling) is to sample with the hard likelihood constraint using a Markov-Chain based procedure. One makes several steps according to some proposal distribution until one is satisfied an independent sample is produced. This has significant advantages over a rejection-based approach, the most obvious being that the scaling with dimensionality is polynomial rather than exponential. In rejection sampling, points are drawn until one is found within the likelihood contour (often with extremely low efficiency). Using a Markov-chain approach however, (correlated) points are continually generated within the contour, until one is happy that a sample independent from the initial seed has been generated. These "intra-chain points" which we term *phantom points* have the potential to provide a great deal more information.

A traditional Metropolis-Hastings (MH) or Gibbs sampling approach may be utilised, but in general such algorithms are ill-suited to sampling from a hard likelihood constraint without a significant amount of tuning of a proposal matrix. This is examined in section 6 of Feroz and Hobson (2008).

Galilean (Hamiltonian) sampling (Feroz and Skilling, 2013, Betancourt, 2011) improves upon the traditional MH sampler by using proposal points generated by reflecting off iso-likelihood contours. This however requires gradients to be calculated, and can become inefficient if the step size is chosen incorrectly, or if the contour has a shape which is difficult to 'step back into'

Diffusive nested sampling (Brewer et al., 2009) is an alternative and promising variation on Skilling's traditional nested sampling algorithm. Diffusive nested sampling utilises MCMC to explore a mixture of nested probability distributions. Since it is MCMC based, it scales well with dimensionality. In addition, it can deal with multimodal and degenerate posteriors, unlike traditional MCMC. It does however have multiple tuning parameters.

9.2.3 Slice sampling

We have found that a Markov-Chain based procedure utilising slice sampling (Neal, 2000) at each step is well suited to sampling uniformly within an iso-likelihood contour.

This procedure for sampling within a likelihood bound is ideal for nested sampling. It samples uniformly with minimal information: an initial bound size w, and a point x_0 that is within the contour. In general w must be chosen so that it is roughly the size of the bound, but if one overestimates it then the bounds will contract exponentially. Indeed, one may consider this as being equivalent to a prior space compression (7.43) with $n_{\rm live} = n_{\rm dims} = 1$. As a starting point, one may use one of the live points, which is already uniformly sampled. Since the procedure above satisfies detailed balance, this will produce a point which is also uniformly sampled within the iso-likelihood contour.

In higher dimensions, Neal (2000) suggests a variety of MCMC-like methods. The simplest of these is implemented by sampling each of the parameter directions in turn. Since each one-dimensional slice requires $\sim \mathcal{O}$(a few) likelihood calculations, the number of likelihood calculations required scales linearly with dimensionality, provided that the region is efficiently navigated.

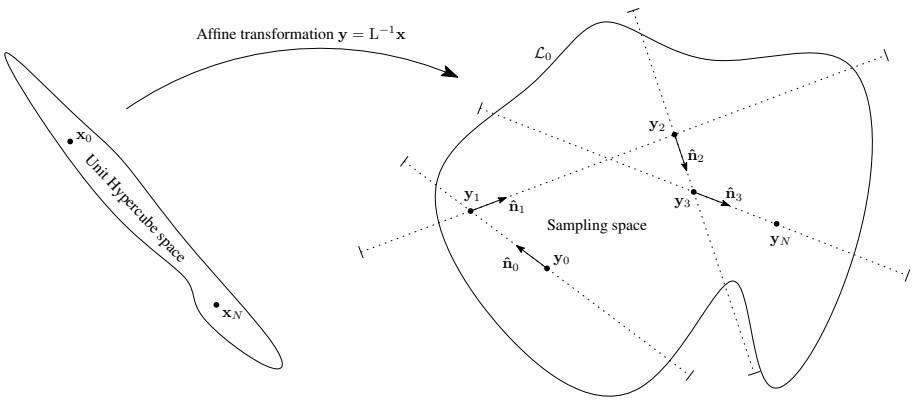


Figure 9.1: Slice sampling in D dimensions. We begin by "whitening" the unit hypercube by making a linear transformation which turns a degenerate contour into one with dimensions $\sim \mathcal{O}(1)$ in all directions. This is a linear skew transformation defined by the inverse of the Cholesky decomposition of the live points' covariance matrix. We term this whitened space the *sampling space*. Starting from a randomly chosen live point $\mathbf{x}_0$, we pick a random direction and perform one-dimensional slice sampling in that direction (Appendix 7.A.5), using $w = 1$ in the sampling space. This generates a new point $\mathbf{x}_1$ in $\sim \mathcal{O}$(a few) likelihood evaluations. This process is repeated $\sim \mathcal{O}(n_{\rm dims})$ times to generate a new uniformly sampled point $\mathbf{x}_N$ which is decorrelated from $\mathbf{x}_0$.

Multi-dimensional slice sampling has many of the benefits of a traditional MH approach, and uses a proposal distribution which is much more efficient at sampling a hard likelihood constraint.

Aitken and Akman (2013) have already applied this procedure to nested sampling. This works exceptionally well for cases in which the parameters are non-degenerate. However, this becomes inefficient in the case of correlated parameters, or curving degeneracies.

9.3 The PolyChord algorithm

PolyChord implements several novel features in comparison with the slice-based nested sampling of Aitken and Akman (2013). It utilises slice sampling in a manner that uses the information present in the live and phantom points to deal with correlated posteriors. PolyChord also uses a general clustering algorithm that identifies and evolves separate modes of the posterior semi-independently, and infers local evidence values. In addition, it has the option of implementing fast-slow parameters, which is extremely effective in its combination with CosmoMC (Lewis and Bridle, 2002). This is termed CosmoChord, which may be downloaded from the link at the end of the chapter.

The algorithm is written in FORTRAN95 and parallelised using open-MPI. It is optimised for the case where the dominant cost is the generation of a new live point. This is frequently the case in astrophysical applications, either due to high dimensionality, or to costly likelihood evaluation.

9.3.1 Multi-dimensional slice sampling

At each iteration i of nested sampling, we generate a new randomly sampled point within the iso-likelihood contour $\mathcal{L}_i$ by our variant of D-dimensional slice sampling. Slice sampling is performed in the unit hypercube with hypercube coordinates denoted in bold ($\mathbf{x}$).

At each iteration i of the nested sampling algorithm, one of the live points is chosen at random as a start point for a new chain with hypercube coordinate $\mathbf{x}_0$. We then make a one-dimensional slice sampling step (Appendix 7.A.5) with initial width w in a random direction $\hat{\mathbf{n}}_0$ chosen from a probability distribution $\mathrm{P}(\hat{\mathbf{n}})$. This generates a new point $\mathbf{x}_1$ which is uniformly sampled in the unit hypercube, but is correlated to $\mathbf{x}_0$. This process is repeated n_{repeats} times, with $\mathbf{x}_{j-1}$ forming the start point for a slice along $\hat{\mathbf{n}}_{j-1}$ to produce $\mathbf{x}_j$. This procedure is illustrated in the right hand half of Figure 9.1.

Since the probability of drawing $\mathbf{x}_j$ from $\mathbf{x}_{j-1}$ is the same as the probability of drawing $\mathbf{x}_{j-1}$ from $\mathbf{x}_j$, this procedure satisfies detailed balance. Thus, the resulting chain will ergodically be uniformly distributed within the iso-likelihood contour. This also applies to multi-modal posteriors, with the chance of jumping out a mode being equal to the chance of jumping back in.

The length of the chain n_{repeats} should be large enough so that the final point of the chain is decorrelated from the start point. This final point may now be considered to be a new uniformly sampled point from the prior distribution subject to the hard likelihood constraint. The intermediate points are saved and stored as phantom points. Whilst phantom points are correlated, they are useful in providing additional information and posterior points.

There are several elements of this which are left undetermined, namely the probability distribution $\mathrm{P}(\hat{\mathbf{n}})$, the initial width w, and the chain length n_{repeats}. These issues are addressed in the next section.

9.3.2 Contour whitening

In order to determine an optimal $\mathrm{P}(\hat{\mathbf{n}})$ and w, an algorithm will need some knowledge of the contour in which the chain is progressing. This information can be supplied by the set of live and phantom points which are already uniformly distributed within the contour. We use the sample covariance matrix of the live and phantom points as a proxy for the size and shape of the contour.

Uniformly sampled points remain uniformly sampled under an affine transformation. The covariance matrix is used to construct an affine transformation which "whitens" the contour. Sampling is then performed in this whitened space, which we term the *sampling space*. In the sampling space, the contour has size $\sim \mathcal{O}(1)$ in every direction. This means that one may choose the initial step size as $w = 1$.

To transform from $\mathbf{x}$ in the unit hypercube to $\mathbf{y}$ in the sampling space we use the relation:

$$\mathrm{L}^{-1}\mathbf{x} = \mathbf{y}, \tag{9.1}$$

where L is the Cholesky decomposition of the covariance matrix $\Sigma = \mathrm{L}\mathrm{L}^T$. This is illustrated further in Figure 9.1.

Working in the sampling space our choice of $\mathrm{P}(\hat{\mathbf{n}})$ is inspired by the default choice of CosmoMC (Lewis, 2013). Here, a randomly oriented orthonormal basis is chosen, and these directions are chosen in a random order. Once a basis

is exhausted, a new basis is chosen. This approach satisfies detailed balance, and mixes rapidly.

The choice of $n_{\rm repeats}$ is slightly harder to justify. We find that for distributions with roughly convex contours $n_{\rm repeats}\sim\mathcal{O}(n_{\rm dims})$ is sufficient, with the constant of proportionality being 2–6. For more complicated contour shapes, one may require much larger values of $n_{\rm repeats}$.

This procedure has the advantage of being dynamically adaptive, and requires no tuning parameters. However, this "whitening" process is ineffective for pronounced curving degeneracies. This will be discussed in detail in Section 9.4.4.

9.3.3 Clustering

Multi-modal posteriors are a challenging problem for any sampling algorithm. "Perfect" nested sampling (i.e. the entire prior volume enclosed by the iso-likelihood contour is sampled uniformly) in theory solves multi-modal problems as easily as uni-modal ones. In practice however, there are two issues.

First, one is limited by the resolution of the live points. If a given mode is not populated by enough live points, it runs the risk of "dying out". Indeed, a mode may be entirely missed if the density of live points is too low. In many cases, this problem can be alleviated by increasing the number of live points.

Second, and more importantly for PolyChord, the sampling procedure may not be appropriate for multi-modal problems. We "whiten" the unit hypercube using the covariance matrix of live points. For far-separated modes, the covariance matrix will not approximate the dimensions of the contours, but instead falsely indicate a high degree of correlation. It is therefore essential for our purposes to have PolyChord recognise and treat modes appropriately.

This methodology splits into two distinct parts: (i) recognising that clusters are there, and (ii) evolving the clusters semi-independently.

Cluster recognition

Any cluster recognition algorithm can be substituted at this point. One must take care that this is not run too often, or one runs the risk of adding a large overhead to the calculation. In practice, checking for clustering every $\sim\mathcal{O}(n_{\rm live})$ iterations is sufficient, since the prior will have only compressed by a factor e. We encourage users of PolyChord to experiment with their own preferred cluster recognition, in addition to that provided and described below.

It should be noted that the live points of nested sampling are amenable to most cluster recognition algorithms for two reasons. First, all clusters should have the same density of live points in the unit hypercube. Second, there is no noise (i.e. outside of the likelihood contour there will be no live points). Many clustering algorithms struggle when either of these two conditions is not satisfied.

We therefore choose a relatively simple variant of the k-nearest neighbours algorithm to perform cluster recognition. If two points are within one another's k-nearest neighbours, then these two points belong to the same cluster. We iterate k from 2 upwards until the clustering becomes stable (the cluster decomposition does not change from one k to the next). If sub-clusters are identified, then this process is repeated on the new sub-clusters.

Cluster evolution

An important novel feature comes from what one does once clusters are identified.

First, when spawning from an existing live point, the whitening procedure is now defined by the covariance matrix of the live points within that cluster. This solves the issue detailed above.

Second, by choosing a random initial live point as a seed, PolyChord would naïvely spawn live points into a mode with a probability proportional to the number of live points in that mode. In fact, what it should be doing is to spawn in proportion to the volume fraction of that mode. In general, these will be approximately the same, but numerical experiments show that the difference between these two ratios exhibits random-walk like behaviour, leading to biases in evidence calculations, or worse, cluster death.

Instead, we keep track of an estimate of the volume in the same manner as equation (7.43), and choose the mode to spawn into in proportion to that estimate. Further, one may track the errors in this estimate, which contribute to the overall evidence error. This methodology is documented fully in Section 8.2.

Thus, the point to be killed off is still the global lowest-likelihood point, but we control the spawning of the new live point into clusters by using our estimates of the volumes of each cluster. We call this 'semi-independent', because it retains global information, whilst still treating the clusters as separate entities.

When spawning within a cluster, we determine the cluster assignment of the new point by which cluster it is nearest to. It does not matter if clusters are identified too soon; the evidence calculation will remain consistent.

In addition to keeping track of local volumes, we may keep track of local evidences. At the moment of splitting, the existing evidence in the initial cluster is partitioned between the new sub-clusters. Upon algorithm completion, one is left with an estimate of the proportion of the evidence contained within each cluster, and thus a measure of the importance of the various modes. By partitioning the local evidences at cluster recognition, the local evidences will sum to give the total evidences, to within the error on our inference.

9.3.4 Parallelisation

PolyChord is parallelised by openMPI using a master-slave structure. One master process takes the job of organising all of the live points, whilst the remaining $n_{\text{procs}} - 1$ "slave" processes take the job of finding new live points. This layout is optimised for the case where the dominant cost is the generation of a new live point due to the calculation of relatively expensive likelihoods.

When a new live point is required, the master process sends a random live point and the Cholesky decomposition to a waiting slave. The slave then, after some work, signals to the master that it is ready and returns a new live point and the intra-chain points to the master.

A point generated from an iso-likelihood contour $\mathcal{L}_i$ is usable as a new live point for an iso-likelihood contour $\mathcal{L}_j > \mathcal{L}_i$, provided it is within both contours. One may keep slaves continuously active, and discard any points returned which are not usable. The probability of discarding a point is proportional to the

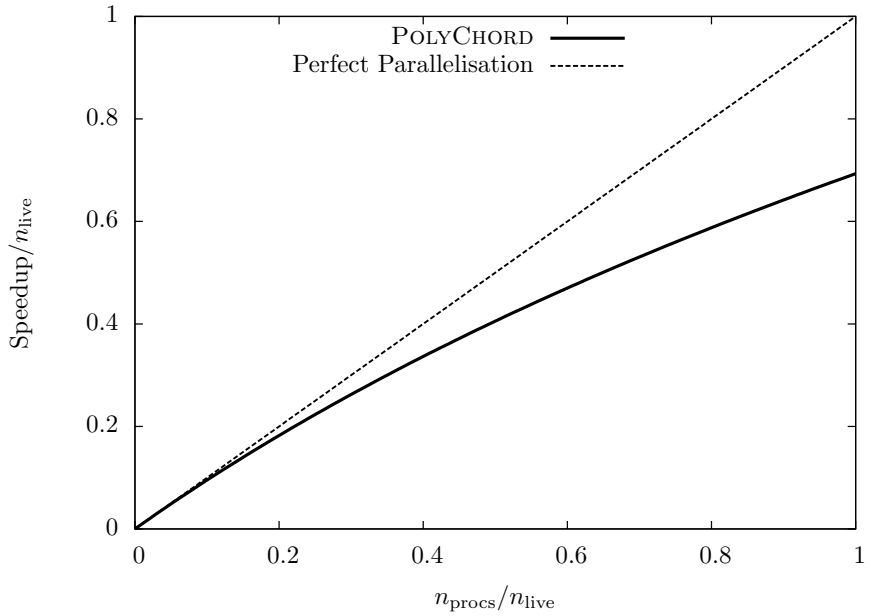


Figure 9.2: Parallelisation of PolyChord. The algorithm parallelises nearly linearly, provided that $n_{\rm procs} < n_{\rm live}$. For most astronomical applications this is more than sufficient.

volume ratio of the two contours, so if too many slaves are used, then most will be discarded. The parallelisation goes as:

$$\text{Speedup}(n_{\rm procs}) = n_{\rm live} \log\left[1 + \frac{n_{\rm procs}}{n_{\rm live}}\right], \tag{9.2}$$

and is illustrated in Figure 9.2. As a rule, PolyChord parallelises well for $n_{\rm procs} < n_{\rm live}$, but from then on exhibits a law of diminishing returns. In practice, $n_{\rm procs} = n_{\rm live}/5$ yields $\sim 90\%$ parallelisation efficiency, and since the number of live points is typically ~ 500, this is more than sufficient for currently available openMPI architectures, and certainly superior to the parallelisation of the standard Metropolis-Hastings algorithm.

9.3.5 Posterior bulking

In addition to lending information on the scale and shape of a contour, phantom points can also be used as posterior samples. Correlations between samples are unimportant for the purposes of parameter estimation, provided one has enough to be well mixed. We may thus use the importance weighting detailed in (7.44) with w_i being set to the volume of the live-point shell which they occupy.

For high-dimensional cosmological applications, this results in a very large number ($\gg$GB) of posterior samples being produced, so PolyChord thins these samples. From a user's perspective, one supplies a parameter which determines the fraction of phantom points to keep.

9.3.6 Fast-slow parameters and CosmoChord

In cosmological applications, likelihoods can exhibit a hierarchy of parameters in terms of calculation speed (Lewis, 2013). Consequently, a likelihood may be quickly recalculated if one changes only a certain subset of the parameters. For PolyChord it is very easy to exploit such a hierarchy. Our transformation to the sampling space is laid out so that if parameters are ordered from slow to fast, then this hierarchy is automatically exploited: a Cholesky decomposition, being a upper-triangular skew transformation, mixes each parameter only with faster parameters.

From a user's perspective, PolyChord does this re-ordering in the hypercube automatically when provided with details of the hierarchy.

Further to this, one may use the fast directions to extend the chain length by many orders of magnitude. This helps to ensure an even mixing of live points. PolyChord automatically times likelihood calculation speeds, so the user just has to provide what fraction of time PolyChord should be spending on each subset of the parameters, and the algorithm will oversample accordingly.

9.3.7 Tuning parameters

From a user's perspective, the PolyChord algorithm has two tuning parameters: $n_{\rm live}$ and $n_{\rm repeats}$, which are detailed below.

We believe that these tuning parameters are fairly straightforward to set in comparison to existing algorithms. More importantly, the number of tuning parameters does not scale with the dimensionality of the problem. This is in contrast to Metropolis-Hastings and Gibbs sampling, which require a proposal matrix to be supplied[1].

There are also several other options controlling run time behaviour, such as the production of equally weighted posterior samples, whether or not to perform clustering and the production and use of files allowing PolyChord to resume from a previous run. These are documented in the input files supplied with the code.

Resolution $n_{\rm live}$

This is a generic nested sampling parameter. $n_{\rm live}$ indicates the number of live points maintained throughout the algorithm. Increasing $n_{\rm live}$ causes nested sampling to contract more slowly in volume (equation 7.43), and consequently sample the space more thoroughly. Thus, it can be thought of as a resolution parameter. Run time scales $\sim \mathcal{O}(n_{\rm live})$

If set too low, posterior modes may be missed. Increasing $n_{\rm live}$ increases the accuracy of the inference of $\mathcal{Z}$, since the evidence error scales $\sim \mathcal{O}\left(n_{\rm live}^{-1/2}\right)$.

Reliability $n_{\rm repeats}$

This is a PolyChord specific parameter. It corresponds to the length of the slice sampling chain used to generate a new live point. Increasing this parameter decreases the correlation between live points, and hence increases

[1] Proposal matrices may be learnt during run-time. However, this learning step can take some time and may reduce the efficacy of these approaches.

the reliability of the evidence inference. Posterior estimations, however, remain accurate even in the event of low n_{repeats}.

Setting this too low can result in correlation between live points, and unreliable evidence estimates. Typically, setting this $\sim \mathcal{O}(3 \times n_{\text{dims}})$ is sufficient, but for curving degeneracies one may need significantly longer chains. Run time scales $\sim \mathcal{O}(n_{\text{repeats}})$.

The total number of live and phantom points $n_{\text{live}} \times n_{\text{repeats}}$ should be large enough that reliable covariance matrices can be calculated. Other than this, the two tuning parameters have independent effects on the algorithm.

In general, n_{repeats} should be scaled linearly with dimensionality D, since one must decorrelate in D independent directions. For typical likelihoods, the logarithmic volume compression from prior to posterior will scale as D. Finally, to keep evidence estimation error constant, the number of live points must be scaled with D. These three effects together mean that PolyChord has a theoretical run time scaling $\sim \mathcal{O}\big(D^3\big)$.

9.4 PolyChord in action

We aim to showcase PolyChord as both a high-dimensional evidence calculator, and multi-modal posterior sampler. We begin by comparing its dimensionality scaling with MultiNest. We then demonstrate its clustering capabilities in high dimensions, and on difficult clustering problems. PolyChord is shown to perform well on moderately pronounced curving degeneracies, and its implementation in CosmoMC is discussed.

9.4.1 High-dimensional evidences

As an example of the strength of PolyChord as a high-dimensional evidence estimator, we compare it to MultiNest on a Gaussian likelihood in D dimensions. In both cases, convergence is defined as when the posterior mass contained in the live points is 10^{-2} of the total calculated evidence. We set $n_{\text{live}} = 25D$, so that the evidence error remains constant with D. MultiNest was run in its default mode with importance nested sampling and expansion factor $e = 0.1$. Whilst constant efficiency mode has the potential to reduce the number of MultiNest evaluations, the low efficiencies required in order to generate accurate evidences negate this effect.

With these settings, PolyChord produces consistent evidence and error estimates with an error ~ 0.4 log units (Figure 9.3). Using importance nested sampling, MultiNest produces estimates that are within this accuracy.

Figure 9.4 shows the number of likelihood evaluations $N_{\mathcal{L}}$ required to achieve convergence as a function of dimensionality D. Even on a simple likelihood such as this, PolyChord shows a significant improvement over MultiNest in scaling with dimensionality. PolyChord at worst scales as $N_{\mathcal{L}} \sim \mathcal{O}\big(D^3\big)$, whereas MultiNest has an exponential scaling which emerges in higher dimensions. However, we must point out that a good rejection algorithm like MultiNest will always win in low dimensions. We therefore recommend using MultiNest for low dimensional problems, although it should be noted that MultiNest's clustering is ineffective in modest dimensionalities.

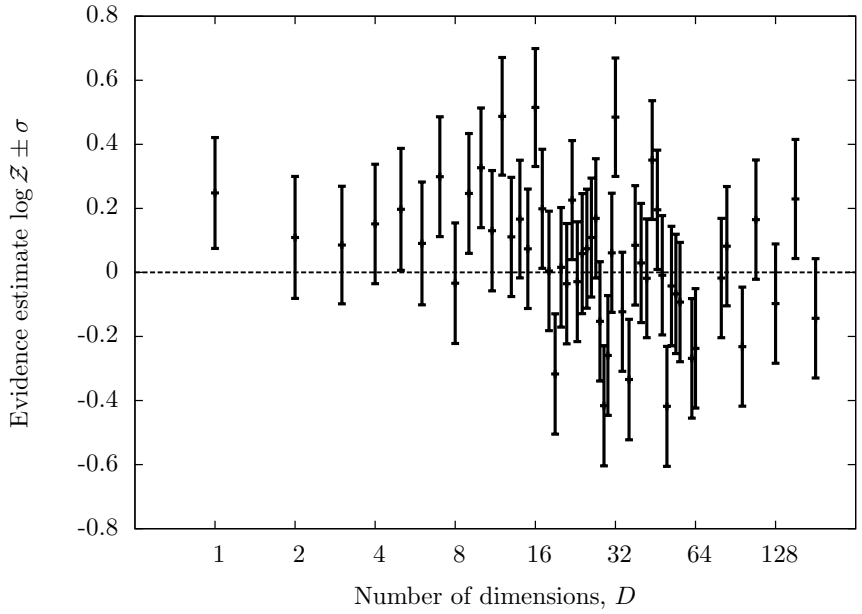


Figure 9.3: Evidence estimates and errors produced by PolyChord for a Gaussian likelihood as a function of dimensionality. The dashed line indicates the correct analytic evidence value.

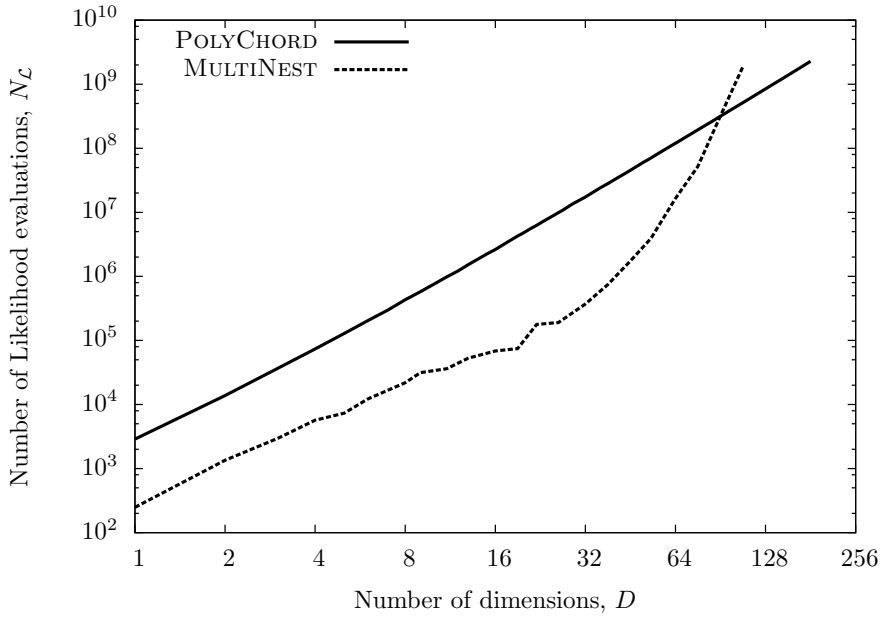


Figure 9.4: Comparing PolyChord with MultiNest using a Gaussian likelihood for different dimensionalities. PolyChord has at worst $N_{\mathcal{L}} \sim \mathcal{O}(D^3)$, whereas MultiNest has an exponential scaling that emerges at high dimensions.

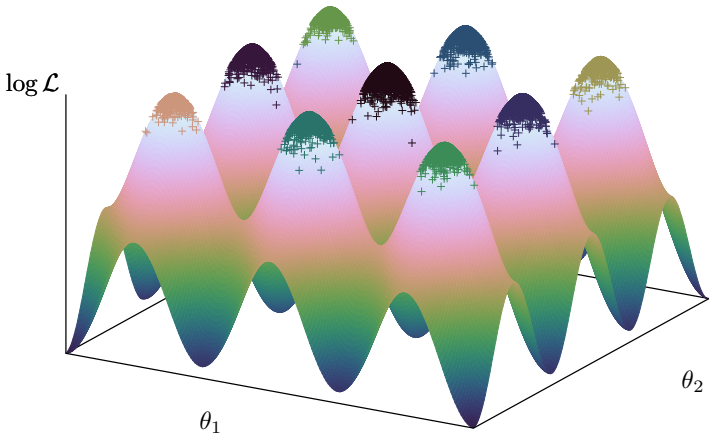


Figure 9.5: The two-dimensional Rastrigin log-likelihood in the range $[-1.5, 1.5]^2$. Within this region there are 8 local maxima, and one global maximum at $(0, 0)$. The clustered samples produced by PolyChord are plotted on the log-likelihood surface, with colours that indicating the separate clusters identified.

9.4.2 Clustering and local evidences

To demonstrate PolyChord's clustering capability we report its performance on a "Twin Peaks" and Rastrigin likelihood.

Twin peaks

PolyChord is capable of clustering posteriors in very high dimensions. We define a twin peaks likelihood as an equal mixture of two spherical Gaussians, separated by a distance of 10σ.

PolyChord correctly identifies these clusters in arbitrary dimensions (tested up to $D = 100$), provided that n_{live} and n_{repeats} are scaled in proportion to D. It calculates a global evidence that agrees with the analytic results. In addition, the local evidences correctly divide the peaks in proportion to their evidence contribution.

The results for a twin peaks likelihood are of an identical character to Figures 9.3 & 9.4, and hence not included.

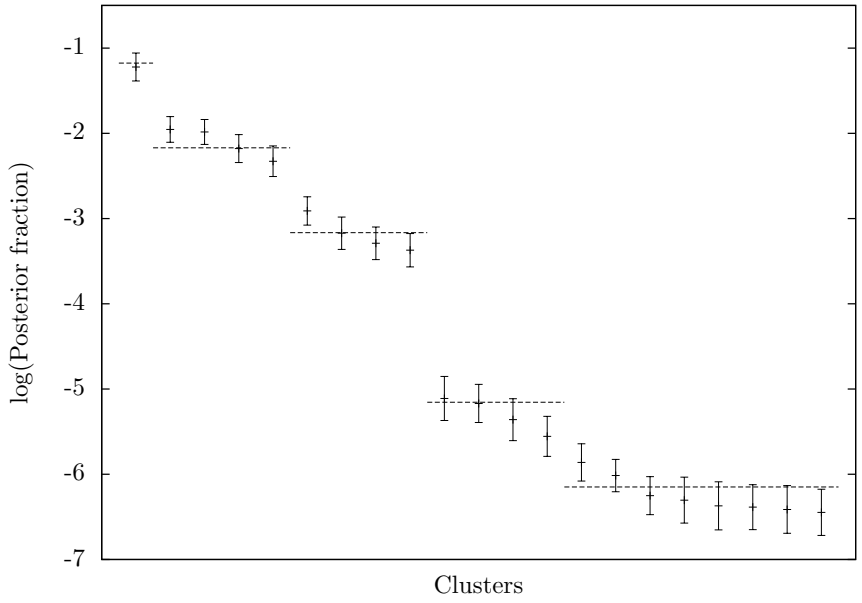


Figure 9.6: PolyChord cluster identification for the Rastrigin function. PolyChord identifies posterior modes and computes their local evidences, expressed here as a logarithmic fraction of the total evidence in the mode. Dashed lines indicate the analytic results computed by a saddle point approximation at each of the peaks. As can be seen, PolyChord reliably identifies the inner 21 modes with increasing accuracy.

Rastrigin function

PolyChord's clustering capacity is very effective on complicated clustering problems as well. The n-dimensional Rastrigin test function is defined by:

$$f(\theta) = An + \sum_{i=1}^{n} \left[\theta_i^2 - A\cos(2\pi\theta_i)\right], \qquad A = 10, \qquad \theta_i \in [-5.12, 5.12]. \tag{9.3}$$

This is the industry standard "bunch of grapes", the two-dimensional version of which is illustrated in Figure 9.5. For our purposes, we will treat (9.3) as the negative log-likelihood so that $\mathcal{L}(\theta) \propto \exp[-f(\theta)]$. This is a stereotypically hard problem to solve, as many algorithms get stuck in local maxima.

We ran PolyChord on a two-dimensional Rastrigin log-likelihood with $n_{\text{live}} = 1000$ and $n_{\text{repeats}} = 6$. With these settings, PolyChord calculates accurate evidence and posterior samples (Figures 9.5 & 9.6), and in addition correctly isolates and computes local evidences for the inner 21 modes. Additional outer modes are also found, but these are combinations of lower modes due to their very low posterior fraction. Increasing the resolution parameter n_{live} further increases the number of modes identified. Examples of clustered posterior samples are indicated in Figure 9.5, coloured using cubehelix (Green, 2011).

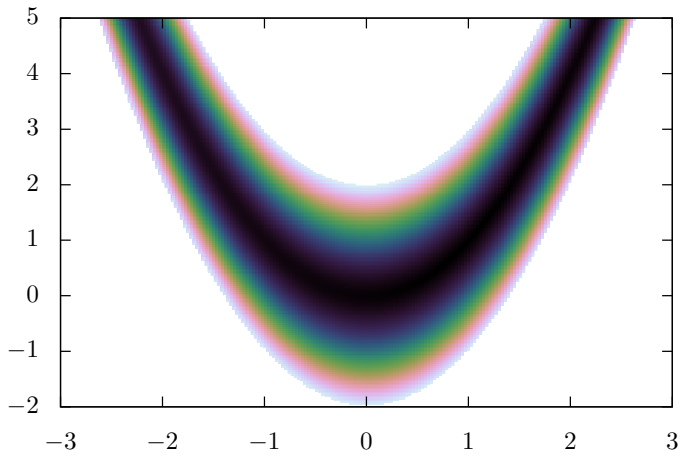


Figure 9.7: Density plot of the two-dimensional Rosenbrock function. The function exhibits a long, thin curving degeneracy, with a global maximum at $(1, 1)$.

9.4.3 Rosenbrock function

PolyChord is also capable of navigating moderate curving degeneracies.

The n-dimensional Rosenbrock function is defined by:

$$f(x) = \sum_{i=1}^{n-1} (a - x_i)^2 + b(x_{i+1} - x_i^2)^2, \tag{9.4}$$

$$a = 1, \quad b = 100, \quad x_i \in [-5, 5], \tag{9.5}$$

the two-dimensional version of which is plotted in Figure 9.7. This is the industry standard "banana", as it exhibits an extremely long and flat curving degeneracy. We consider $n = 4$, in which there is a global maximum at $(1, 1, 1, 1)$ and a local maximum at $(-1, 1, 1, 1)$. The true evidence value is -15.1091, and with $n_{\text{live}} = 1000, n_{\text{repeats}} = 12$, PolyChord reliably finds both peaks (Figure 9.8) and produces a correct evidence estimation.

In higher dimensions, PolyChord reliably finds the local and global maxima. The lack of an analytic evidence value for the Rosenbrock function prevents a verification of the evidence calculation.

9.4.4 Gaussian shells

A "Gaussian shell" with mean μ, radius r and width w is defined as:

$$\log \mathcal{L}_{\text{shell}}(\mathbf{x}|\mu, r, w) = A - \frac{(|\mathbf{x} - \mu| - r)^2}{2w^2}, \tag{9.6}$$

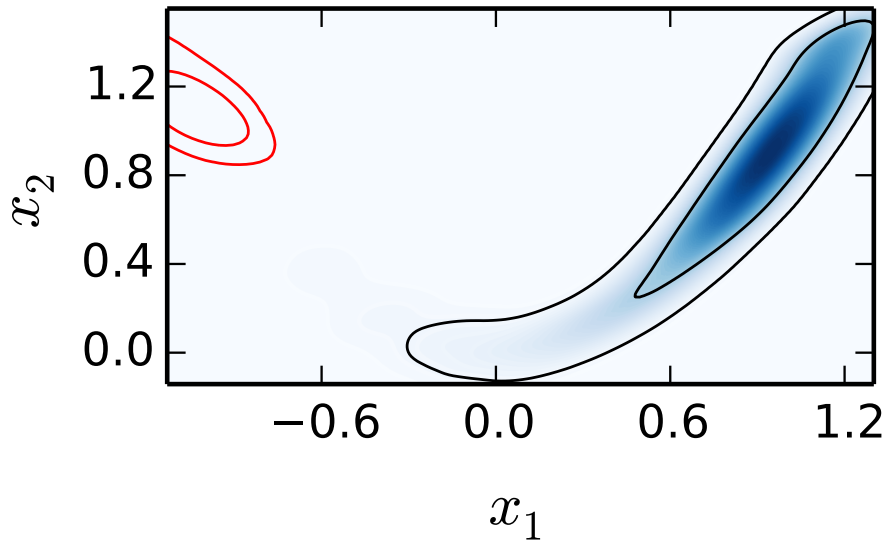


Figure 9.8: The four-dimensional Rosenbrock posterior, with x_3 and x_4 marginalised out. PolyChord correctly identifies both the local (red) and global (blue) maxima.

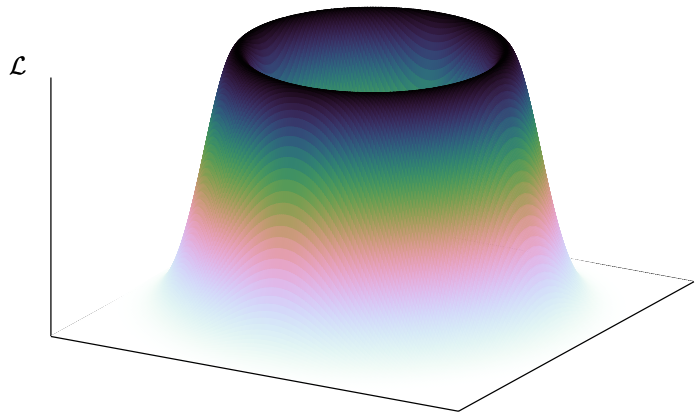


Figure 9.9: The two-dimensional Gaussian shell likelihood.

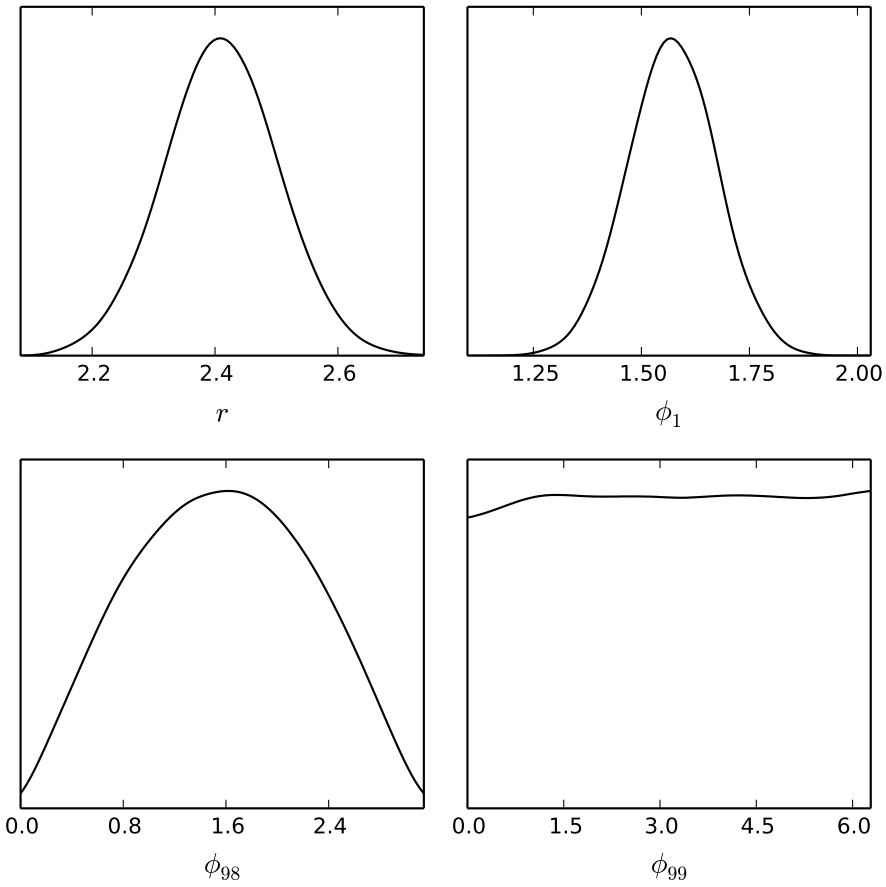


Figure 9.10: Posteriors produced by PolyChord for a $n = 100$-dimensional Gaussian shell, with width $w = 0.1$, radius $r = 2$, and centre $\mu = \mathbf{0}$. Plotting the marginalised posteriors for the Cartesian sampling parameters $\{x_1, \ldots, x_n\}$ yields Gaussian distributions centred on the origin. To see the effectiveness of the sampler it is better to plot the sampling parameters in terms of n-dimensional spherical polar coordinates $\{r, \phi_1, \ldots, \phi_{n-1}\}$. Note that the polar coordinates are *derived parameters*, and that the sampling space still has the strong Gaussian shell degeneracy. In this case we can see that the radial coordinate has a Gaussian profile centred on $r_0 = r \times \frac{1}{2}\left(1 + \sqrt{1 + 4(n-1)(w/r)^2}\right)$ with width $w_0 = w(1 + (n-1)(w/r_0)^2)^{-1/2}$. The azimuthal coordinate ϕ_{n-1} has a uniform posterior, and the other angular coordinates $\{\phi_i\}$ have posteriors defined by $\mathrm{P}(\phi_i) \propto (\sin\phi_i)^{n-i-1}$.

where A is a normalisation constant that may be calculated using a saddle point approximation. This likelihood is centred on some mean vector μ, and has a radial Gaussian profile with width w at distance r from this centre. This radial profile is then revolved around μ to create a spherical shell-like likelihood. A two-dimensional version of this likelihood is indicated in Figure 9.9.

This distribution may be representative of likelihoods that one may encounter in beyond-the-Standard-Model paradigms in particle physics. In such models, the majority of the posterior mass lies in thin sheets or hypersurfaces through the parameter space.

Running PolyChord on a 100-dimensional Gaussian shell with $n_{\text{live}} = 1000$, $n_{\text{repeats}} = 200$ yields consistent evidences and posteriors, shown in Figure 9.10.

Given that this problem is quoted as being "optimally difficult" (Feroz et al., 2009), the ease with which PolyChord tackles this problem in high dimensions is worth explanation. In the two-dimensional case, it is clear that the posterior mass is concentrated in a very thin, curving region of the parameter space. However, as the dimensionality is increased, more and more of the n-sphere's volume is concentrated at the edge, and the thin characteristic of the degeneracy is lost.

This may mean that the Gaussian shell is not a good proxy for a high-dimensional curving degeneracy. However, it could equally suggest that curving degeneracies become easier to navigate in higher dimensions. We can certainly conclude that a particle physics model with a proliferation of phases would be easier to navigate than one with a smaller number of phases.

Twin Gaussian shells

We finish our toy problems by combining the difficulties of multimodality (Section 9.4.2) and degeneracy, by mixing two twin Gaussian shells together:

$$\mathcal{L}(\mathbf{x}) \propto \mathcal{L}_{\text{shell}}(\mathbf{x}|\mu_1, r, w) + \mathcal{L}_{\text{shell}}(\mathbf{x}|\mu_2, r, w). \tag{9.7}$$

We choose $r = 2$, $w = 0.1$, and μ_1 and μ_2 are separated by 7 units. With $n_{\text{live}} = 10n_{\text{dims}}$ and $n_{\text{repeats}} = 2n_{\text{dims}}$, PolyChord successfully computes the local and global posteriors and evidences up to $D = 100$, and reliably identifies the two modes. The comparison of run times with MultiNest recovers a similar pattern to Figure 9.4, although in our experience, the MultiNest parameters require some tuning to ensure that evidences are calculated correctly when $n_{\text{dims}} > 30$.

9.4.5 CosmoChord

An additional strength of PolyChord lies in its ability to exploit a fast-slow hierarchy common in many cosmological applications.

As an example, we consider the likelihoods provided by CAMB (Lewis et al., 2000) and CosmoMC (Lewis and Bridle, 2002). In Boltzmann codes such as CAMB, parameters controlling the primordial power spectrum (such as n_s and A_s) do not require recalculation of transfer functions. These parameters are termed "semi-slow". In addition, modern Planck likelihoods (Planck Collaboration et al., 2014) have nuisance parameters associated with the foregrounds. These may be varied without recalculation of the cosmological back-

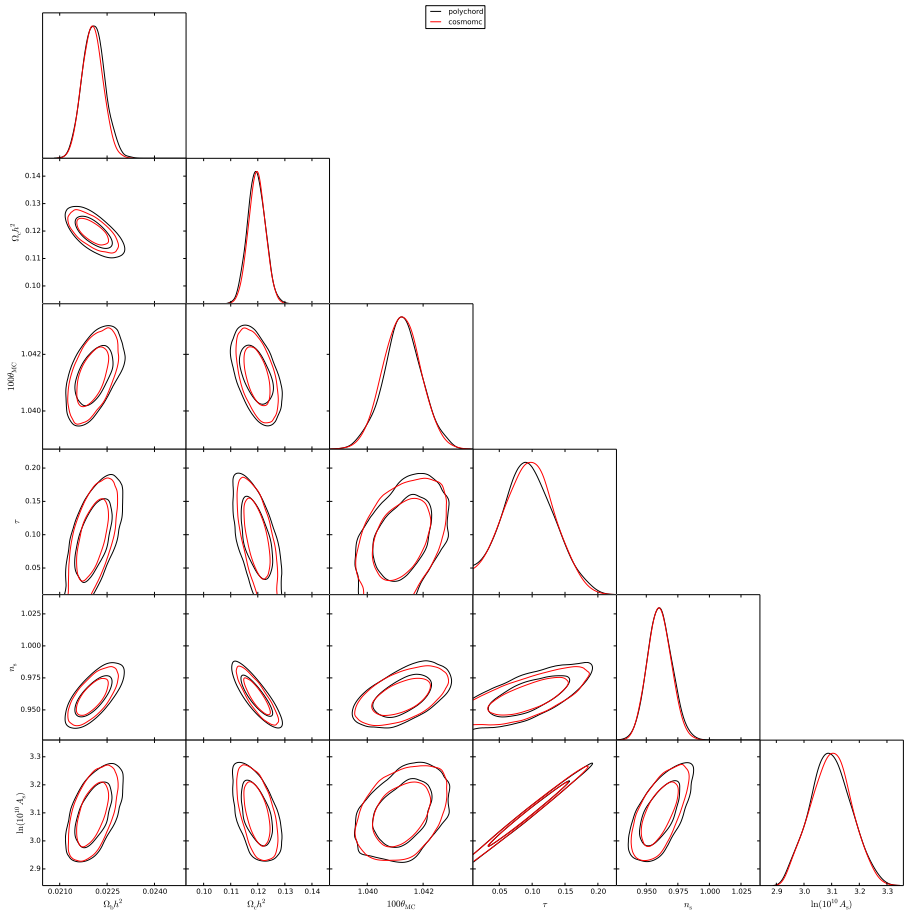


Figure 9.11: CosmoChord (red) vs. CosmoMC (black). We use the 2013 `CAMSPEC+commander` likelihoods with a standard six-parameter ΛCDM cosmology, varying all 14 nuisance parameters. We compare the 1 and 2-dimensional marginalised posteriors of the 6 ΛCDM parameters. CosmoChord is in close agreement with the posteriors produced by CosmoMC, recovering the correct mean values of and degeneracies between the parameters. The slight deviations between the red and black curves are sampling noise.

ground. These parameters are hence termed "fast". CosmoMC (Lewis and Bridle, 2002) implements this hierarchy of speeds in its likelihood calculation.

We have successfully implemented PolyChord within CosmoMC, and term the result CosmoChord. The traditional Metropolis-Hastings algorithm is replaced with nested sampling. This implementation is available to download from the link at the end of the chapter.

The exploitation of fast-slow parameters means that CosmoChord vastly outperforms MultiNest when running with modern Planck likelihoods.

CosmoMC by default uses a Metropolis-Hastings sampler. If this has a well-tuned proposal distribution (e.g. if one is performing importance sampling from an already well-characterised likelihood), then PolyChord is 2–4 times slower than the traditional CosmoMC. If proposal matrices are unavailable (e.g. in the case that one is examining an entirely new model) then CosmoChord's run time is competitive with the native CosmoMC sampler. This is a good example of the self-tuning capacity of PolyChord, since it only requires two tuning parameters, as opposed to $\sim \mathcal{O}(D^2)$.

CosmoChord produces parameter estimations consistent with CosmoMC (Figure 9.11). It has been implemented effectively in multiple cosmological applications in the latest Planck paper describing constraints on inflation (Planck Collaboration et al., 2016b), including application to a 37-parameter reconstruction problem (4 slow, 19 semi-slow, 14 fast). In addition, PolyChord is an integral component of the ModeChord code, a combination of CosmoChord and ModeCode (Mortonson et al., 2011, Easther and Peiris, 2012, Norena et al., 2012), which is available at http://modecode.org/.

9.5 Conclusions

We have introduced PolyChord, a novel nested sampling algorithm tailored for high-dimensional parameter spaces. It is able to fully exploit a hierarchy of parameter speeds such as is found in CosmoMC and CAMB. It utilises slice sampling at each iteration to sample within the hard likelihood constraint of nested sampling. It can identify and evolve separate modes of a posterior semi-independently and is parallelised using openMPI.

Download Link

PolyChord is available for download from:
http://ccpforge.cse.rl.ac.uk/gf/project/polychord/

Chapter 10

The Runge-Kutta-Wentzel-Kramers-Brillouin method

10.1 Introduction

The numerical solution of linear, ordinary differential equations is of critical importance throughout science and mathematics. In this chapter we suggest an efficient approach for navigating highly oscillatory numerical solutions.

Most traditional numerical solvers of differential equations use a generalisation of Runge-Kutta (RK) techniques (Press et al., 2007). These apply Taylor's theorem to create a stepping scheme whereby the value of the solution is updated using derivative information. Good solvers will also incorporate adaptive step-size control. Whilst RK techniques are an excellent workhorse for solving a wide variety of problems, they are known to struggle to solve equations with highly oscillatory solutions.

On the other hand, the Wentzel-Kramers-Brillouin (WKB) method is a well established analytical approach for approximately describing oscillatory solutions (Riley et al., 2006, Bender and Orszag, 1999). Historically this has been used to approximate the global shape and characteristics of an oscillating solution with a "slowly changing" frequency.

We propose that one may combine the two approaches to create a reliable general tool for the numerical solution of oscillatory differential equations, and term the result RKWKB[1].

We note that this approach is similar to the work of Iserles (2002b,a), and explore the similarities and differences in Section 10.6. An extended version of this chapter has been submitted to the Journal of Computational Physics as Handley et al. (2016b).

[1]Readers with experience in the field will note that, as Cambridge authors, we should be insisting on an additional 'J' in WKB (for Jeffreys). Given the length of our proposed acronym, we have opted to use the more efficient nomenclature.

10.2 Background

10.2.1 Oscillating solutions

We seek to create a numerical method which efficiently solves linear differential equations such as:

$$\ddot{x}(t) + \omega(t)^2 x(t) = 0, \qquad \omega(t) \in \mathbb{R}. \tag{10.1}$$

If $\omega(t) = \omega =$ constant, then the solutions are sinusoidal: $x \propto \exp(\pm i\omega t)$. If $\omega(t)$ changes slowly with t, then the solutions are approximately sinusoidal with a slowly varying frequency and amplitude (these ideas will be made more concrete in Section 10.2.3). An example of such a solution can be seen in Figure 10.1.

In general, any second-order linear differential equation may be transformed into the form of (10.1) by either changing the independent variable t or dependent variable x. The method we shall describe can easily be adapted to other linear differential equations, but we shall work with (10.1) for its simplicity of exposition.

Equation (10.1) is ubiquitous in physics, particularly in quantum mechanics. Our particular interest in its efficient solution comes from work in quantum fields in curved spacetime.

Over the next two subsections we will review the traditional techniques available for solving equations such as (10.1).

10.2.2 Runge-Kutta theory

We briefly review the theory of numerically solving ordinary differential equations, before discussing why Runge Kutta techniques are an inefficient tool for solving equations such as (10.1). For a more detailed introduction to the numerical solution of ordinary differential equations we recommend Press et al. (2007).

A general non-linear differential equation in n variables can be written in terms of vectors as:

$$\dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}(t), t). \tag{10.2}$$

Note that any higher order differential equation can be re-written in this form by introducing new variables for each of the higher derivative terms.

Runge-Kutta methods work effectively by generalising the Taylor expansion:

$$\mathbf{y}(t+h) = \mathbf{y}(t) + h\,\mathbf{f}(\mathbf{y}(t), t) + \mathcal{O}(h^2). \tag{10.3}$$

Given the value of a solution $\mathbf{y}_j$ at some time t_j, one may advance to the value of the solution $\mathbf{y}_{j+1}$ at some finite time later $t_{j+1} = t_j + h$ by using the recursion relation:

$$\mathbf{y}_{j+1} = \mathbf{y}_j + h\,\mathbf{f}(\mathbf{y}_j, t_j), \tag{10.4}$$

$$t_{j+1} = t_j + h. \tag{10.5}$$

This is termed *Euler's method*, and for arbitrarily small h will recover the solution to any desired accuracy. It is termed *first order* since each step is accurate to $\sim \mathcal{O}(h)$.

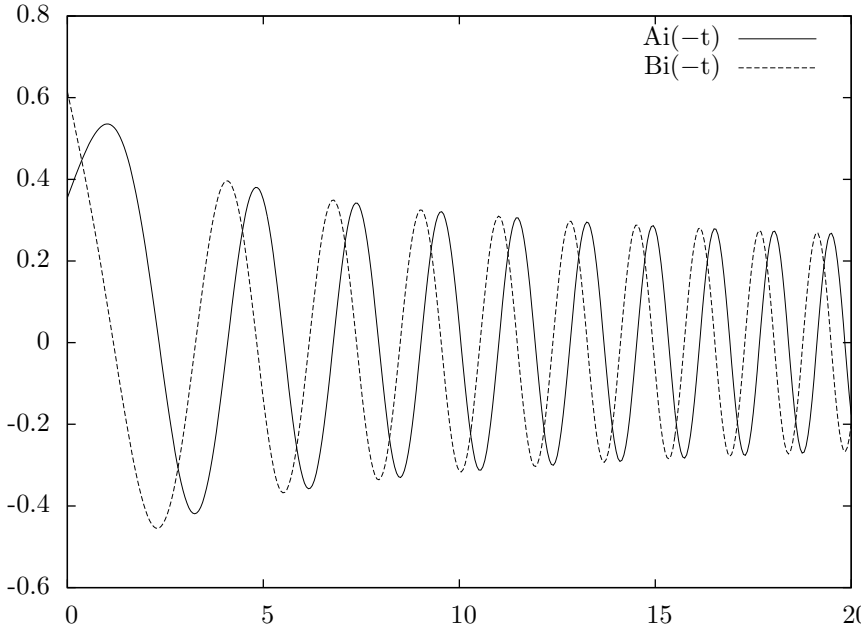


Figure 10.1: The real and imaginary parts of the function $\mathrm{Ai}(-t) + \mathrm{Bi}(-t)i$, where Ai and Bi are the Airy functions of the first and second kind. This is a solution to the equation $\ddot{x}(t) + tx(t) = 0$ (equation 10.33).

Euler's method is normally impractical for real numerical work. Runge-Kutta schemes work by generalising (10.3) & (10.4) by including additional intermediate function evaluations that integrate (10.2) with greater accuracy.

A possibly more important adjustment is to equip the algorithm with the ability to choose the step size h according to the accuracy required. A popular stratagem is to run two steps, one of order p, and another of order $p-1$, and use the difference between the two as an estimate of the error. Particularly smart algorithms use the same function evaluations for both orders. An example of this is the Runge-Kutta-Fehlberg 4(5) method detailed in Appendix 10.A.

All methods based on this principle struggle to solve equations such as (10.1) when the algorithm must scale a very large number of peaks and troughs. Errors accumulate rapidly in these approaches, even if the variation of $\omega(t)$ in t is very simple. Given the regularity of the solution from Figure 10.1, one would imagine that there should be a more efficient method.

10.2.3 WKB theory

WKB approaches are designed to solve linear ordinary differential equations like (10.1) in the limit of a "slowly varying" $\omega(t)$: i.e. the fractional change in frequency $\frac{\Delta\omega}{\omega}$ over several time periods $\Delta t \sim \frac{2\pi}{\omega}$ is relatively small. A systematic way of phrasing this is to rescale the independent variable of (10.1) so $t \to t/T$:

$$\ddot{x}(t) + T^{-2}\omega(t)^2 x(t) = 0, \qquad \omega(t) \in \mathbb{R}. \tag{10.6}$$

If $T \gg 1$ then ω is slowly varying, or equivalently the solutions have very rapid oscillations (large ω). Given this, one can expand the solutions in terms of complex exponential functions:

$$x(t) \sim \exp\left(\frac{1}{T}\sum_{n=0}^{\infty} S_n(t)\, T^n\right). \tag{10.7}$$

Substituting this into (10.6) and setting each coefficient of T equal to zero yields a sequence of solvable equations. One finds the first four solutions are:

$$S_0(t) = \pm i \int^t \omega(\tau)\ \mathrm{d}\tau, \tag{10.8}$$

$$S_1(t) = -\frac{1}{2}\log\omega(t), \tag{10.9}$$

$$S_2(t) = \mp i \int^t \frac{1}{4}\frac{\ddot{\omega}(\tau)}{\omega^2(\tau)} - \frac{3}{8}\frac{\dot{\omega}^2(\tau)}{\omega^3(\tau)}\ \mathrm{d}\tau, \tag{10.10}$$

$$S_3(t) = \frac{1}{8}\frac{\ddot{\omega}(t)}{\omega^3(t)} - \frac{3}{16}\frac{\dot{\omega}^2(t)}{\omega^4(t)}, \tag{10.11}$$

and in general:

$$\dot{S}_0(t) = \pm i\omega, \qquad \dot{S}_n = -\frac{1}{2\dot{S}_0}\left(\ddot{S}_{n-1} + \sum_{j=1}^{n-1}\dot{S}_j\dot{S}_{n-j}\right). \tag{10.12}$$

Note that at 0^{th} order, the solution is $x \propto \exp\left(\pm i\int\omega\,\mathrm{d}t\right)$, which should be compared with the traditional sinusoidal solution. Typically T is considered a power counting parameter, and set equal to 1 at the end of the analysis. For further detail on the intricacies of WKB approaches, the reader should consult Riley et al. (2006), Bender and Orszag (1999).

10.3 Generalised stepping methods

In Section 10.4 we shall combine the power of WKB approaches with RK to form a method specialised to navigating oscillatory solutions. However, to avoid confusion of the details of WKB with the generic nature of our approach, we shall work in a more general case for this section.

In the general case, one aims to solve numerically a linear, homogeneous, second-order differential equation with non-constant coefficients:

$$\alpha(t)\frac{\mathrm{d}^2}{\mathrm{d}t^2}x(t) + \beta(t)\frac{\mathrm{d}}{\mathrm{d}t}x(t) + \gamma(t)x(t) = 0. \tag{10.13}$$

In addition, we shall assume that we have two analytical linearly independent approximate solutions $f_\pm(t)$, as well as access to their derivatives $\dot{f}_\pm$ and $\ddot{f}_\pm$. The aim is to then construct an approach which uses these approximate solutions to create a numerical stepping procedure, analogous to a Runge-Kutta approach. Examples of such functions f could be the WKB solutions (which is described in Section 10.4), or perhaps carefully chosen polynomials.

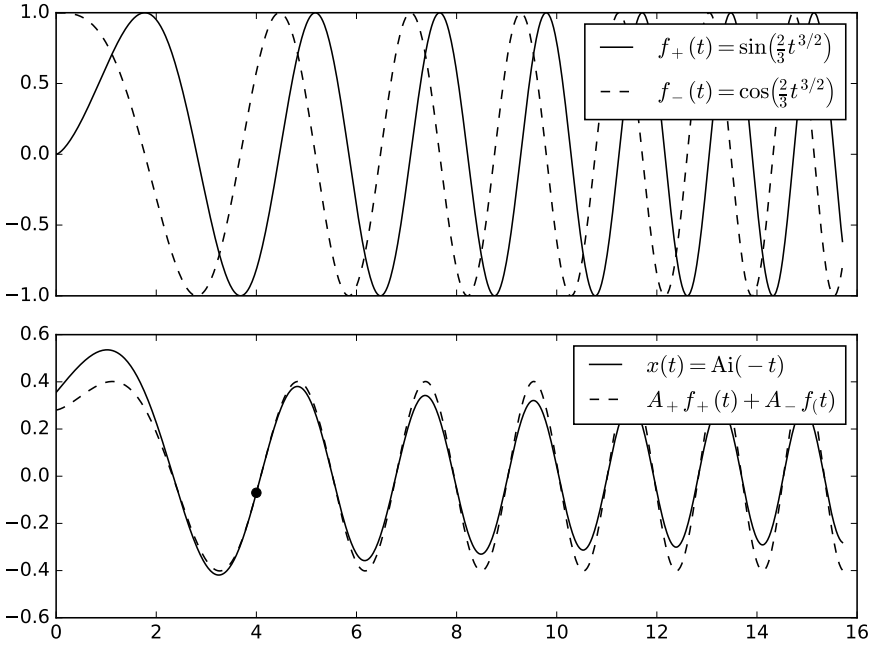


Figure 10.2: Approximating an Airy function with two sinusoids of varying frequency. The lower figure shows the Airy function of the first kind $\mathrm{Ai}(-t)$ (solid line) being approximated in the region around $t_j = 4$ (dashed line). This is done by matching sinusoids of varying frequency onto the Airy function's value and derivative at t_j. The top part of the figure details the sinusoids, which turn out to be equivalent to 0^{th} order WKB solutions.

At a given time t_j with values of the true solution x_j and its derivative $\dot{x}_j$, one may match the approximate solutions onto the correct solution:

$$x(t) \approx A_+ f_+(t) + A_- f_-(t), \tag{10.14}$$

$$A_\pm = \frac{\dot{x}_j f_\mp(t_j) - x_j \dot{f}_\mp(t_j)}{\dot{f}_\pm(t_j) f_\mp(t_j) - \dot{f}_\mp(t_j) f_\pm(t_j)}. \tag{10.15}$$

This provides an approximation to the true solution in the region $t = t_j + h$ where h is small. The degree of its failure is then determined by how well the solutions $f_\pm$ trace the true solutions. This can be seen graphically in Figure 10.2.

A naïve approach would be to use the approximate matched solution (10.14) to create a stepping procedure analogous to the RK method (10.4) & (10.5):

$$x_{j+1} = A_+ f_+(t_j + h) + A_- f_-(t_j + h), \tag{10.16}$$

$$\dot{x}_{j+1} = A_+ \dot{f}_+(t_j + h) + A_- \dot{f}_-(t_j + h), \tag{10.17}$$

$$t_{j+1} = t_j + h. \tag{10.18}$$

Alas, such a method is doomed to failure[2]. Since the approximate solution

[2]Credit here is due to Anthony Challinor for spotting this error in an earlier version of

$A_+ f_+ + A_- f_-$ is defined entirely by the value of x and $\dot{x}$ at any given point, using the values x_j and $\dot{x}_j$ to forecast onto x_{j+1} and $\dot{x}_{j+1}$ merely continues the solution of the previous step. The coefficients $A_\pm$ do not change, and such a method merely follows a single curve ad infinitum.

One can see this more concretely by observing that this method should replicate a simple Runge-Kutta approach in the limit of vanishing step size h (which this method does not). In the limit of small h, one has:

$$f(t_{j+1}) \approx f(t_j) + \dot{f}(t_j)\, h + \mathcal{O}(h^2), \tag{10.19}$$

$$\dot{f}(t_{j+1}) \approx f(t_j) + \ddot{f}(t_j)\, h + \mathcal{O}(h^2). \tag{10.20}$$

Substituting these into equations (10.16) & (10.17) yields:

$$\Rightarrow x_{j+1} = x_j + \dot{x}_j\, h, \tag{10.21}$$

$$\Rightarrow \dot{x}_{j+1} = \dot{x}_j + \frac{(\ddot{f}_+ f_- - \ddot{f}_- f_+)\dot{x}_j + (\ddot{f}_- \dot{f}_+ - \ddot{f}_+ \dot{f}_-)x_j}{\dot{f}_+ f_- - \dot{f}_- f_+}\, h, \tag{10.22}$$

where in the final equation, all f terms are evaluated at t_j. In general, unless f is an exact solution, one cannot expect the coefficient of h in the second equation to be the same as $\ddot{x}_j$, and thus the approach (10.16) & (10.17) fails to recover the Runge-Kutta result. Thus, decreasing the step size does not improve accuracy, and this algorithm is not convergent.

The above issue also suggests a solution. We require an alternative to (10.17) which steps $\dot{x}_j$ such that in the limit of small h the stepping procedure reduces to $\dot{x}_{j+1} = \dot{x}_j + \ddot{x}_j\, h$. We should therefore perform a separate step for $\dot{x}$, with the solution matched onto the values of $\dot{x}_j$ and $\ddot{x}_j$:

$$\dot{x}(t) \approx B_+ \dot{f}_+(t) + B_- \dot{f}_-(t), \tag{10.23}$$

$$B_\pm = \frac{\ddot{x}_j \dot{f}_\mp(t_j) - \dot{x}_j \ddot{f}_\mp(t_j)}{\ddot{f}_\pm(t_j)\dot{f}_\mp(t_j) - \ddot{f}_\mp(t_j)\dot{f}_\pm(t_j)}. \tag{10.24}$$

Most importantly, one may determine $\ddot{x}_j$ from $\dot{x}_j$ and x_j via the original second-order linear differential equation (10.13).

The general stepping procedure is then as follows:

$$x_{j+1} = A_+ f_+(t_j + h) + A_- f_-(t_j + h), \tag{10.25}$$

$$\dot{x}_{j+1} = B_+ \dot{f}_+(t_j + h) + B_- \dot{f}_-(t_j + h), \tag{10.26}$$

$$t_{j+1} = t_j + h, \tag{10.27}$$

and at every iteration, $A_\pm$ and $B_\pm$ are determined by equations (10.15) & (10.24) with $\ddot{x}_j$ calculated from $\dot{x}_j$ and x_j via the original second-order linear differential equation. By definition (or somewhat laborious algebra), in the limit of small h, this recovers the RK result:

$$x_{j+1} = x_j + \dot{x}_j\, h + \mathcal{O}(h^2), \tag{10.28}$$

$$\dot{x}_{j+1} = \dot{x}_j + \ddot{x}_j\, h + \mathcal{O}(h^2). \tag{10.29}$$

---

our approach.

10.4 The RKWKB method

We now specialise the generic technique detailed in Section 10.3 to the case of equations with oscillatory solutions. Our strategy is to combine the versatility of RK methods with the power of WKB in dealing with oscillating solutions, and term the combination RKWKB.

We apply the generalised stepping technique (10.25), (10.26) & (10.27) by choosing $f_\pm$ to be the WKB solutions:

$$f_\pm(t) = \frac{1}{\sqrt{\omega(t)}} \exp(\pm i \int^t \omega(\tau)\, \mathrm{d}\tau + \ldots). \tag{10.30}$$

As can be seen in Figure 10.2, choosing these as $f_\pm$ means that our stepping procedure naturally encodes the oscillatory nature of the solutions, particularly if the frequency is large. Instead of following every peak and trough as a RK scheme must do, it is potentially able to leap over many oscillations at once, greatly increasing the speed of solution.

10.4.1 Step size adjustment

To tune the step size h, we use the same strategy as adaptive Runge-Kutta schemes. We compute both the order n and order $n-1$ WKB solutions, and use the fractional difference between the two:

$$\varepsilon = \left| \frac{x^{(n)} - x^{(n-1)}}{x^{(n)}} \right|, \tag{10.31}$$

as an estimate of the truncation error.

We now assume that the desired accuracy is α. If $\varepsilon < \alpha$ then the solution is within the desired tolerance, and the algorithm makes a step of size h. h is then increased for the next iteration. If $\varepsilon > \alpha$ then the step is unsuccessful, and the step size is reduced. h may therefore be efficiently updated between attempts via:

$$h \to h \times \begin{cases} (\alpha/\varepsilon)^{1/n} & : \varepsilon < \alpha \\ (\alpha/\varepsilon)^{1/(n-1)} & : \varepsilon > \alpha. \end{cases} \tag{10.32}$$

This allows the step size to increase in the regions where the initial step size is unnecessarily small, whilst ensuring that the step size is always small enough to keep forecasts within a given error margin.

10.4.2 Dynamic switching

In general, one cannot expect the WKB expansion to be efficient throughout the solution region. If ω is too small, or too quickly varying, then the step size h will decrease to an inefficiently small size. This problem can be countered by simultaneously attempting a step using a standard adaptive RK method. One chooses between RK and WKB by selecting the method with the smallest error. This provides a natural switching mechanism, without having to delve into the details of whether WKB is valid or not.

We choose the Runge-Kutta-Fehlberg 4(5) method for our alternative solver, which is detailed in Appendix 10.A. However, this may be substituted with any ODE solver according to the user's preference.

10.5 Example: The Airy equation

As an example of the RKWKB approach, we apply it to the Airy equation:

$$0 = \ddot{x}(t) + t\, x(t), \tag{10.33}$$

$$x(0) = \frac{3^{-2/3} + 3^{-1/6} i}{\Gamma(2/3)}, \qquad \dot{x}(0) = \frac{3^{-1/3} - 3^{1/6} i}{\Gamma(1/3)}, \tag{10.34}$$

$$\Rightarrow x(t) = \mathrm{Ai}(-t) + \mathrm{Bi}(-t)\, i, \tag{10.35}$$

whose solution is depicted in Figure 10.1. This is often quoted as being a "maximally hard" problem for RK machinery to solve, since the frequency steadily increases, causing the step size to get smaller as the algorithm goes deeper into the solution.

We set the desired relative error to be 10^{-4}. The algorithm remains in the RK regime until $t \sim 5$. When the WKB solver is activated, instead of following every oscillation of the solution, it rapidly speeds up, skipping many oscillations. This is detailed in Figure 10.3.

The error compared to the true solution is detailed in Figure 10.4. Here we find that initially the error is small, but grows$\sim \mathcal{O}\big(h^{2s} t^{s+5/4}\big)$ where $s = 4$ is the order of the RK method (Iserles, 2002b). After the WKB regime is entered, it begins to make huge strides, and the error levels off.

In contrast to a "pure" RK method, the RKWKB method finds the Airy equation maximally *easy*.

10.6 Comparison with the Iserles approach

Iserles has written extensively on the difficulty of solving problems of the form (10.1). His approach is to turn (10.1) into a Lie-group differential equation (Iserles et al., 2000) by writing:

$$\mathbf{y} = (x, \dot{x})^\top, \mathrm{A}(t) = \begin{pmatrix} 0 & 1 \\ -\omega^2(t) & 0 \end{pmatrix}, \tag{10.36}$$

$$\Rightarrow \quad \dot{\mathbf{y}} = \mathrm{A}(t)\, \mathbf{y}. \tag{10.37}$$

This may then be attacked with a variety of Lie group methods. For example, one may write the full solution as a Magnus expansion:

$$\mathbf{x}(t) = e^{\Omega(t,t_0)} \mathbf{x}_0, \tag{10.38}$$

$$\begin{aligned} \Omega(t, t_0) = & \int_{t_0}^{t} \mathrm{A}(x)\, \mathrm{d}x \\ & - \frac{1}{2} \int_{t_0}^{t} \int_{t_0}^{x_1} [\mathrm{A}(x_2), \mathrm{A}(x_1)]\, \mathrm{d}x_2\, \mathrm{d}x_1 + \ldots, \end{aligned} \tag{10.39}$$

and then use a truncated series to create a stepping algorithm. This approach is much improved by transferring to a fast rotating frame:

$$\mathbf{y}(t_n + \tau) = e^{\tau \mathrm{A}(t_n + h/2)} \mathbf{x}, \tag{10.40}$$

the end product is then termed the *modified Magnus method* (Iserles, 2002a).

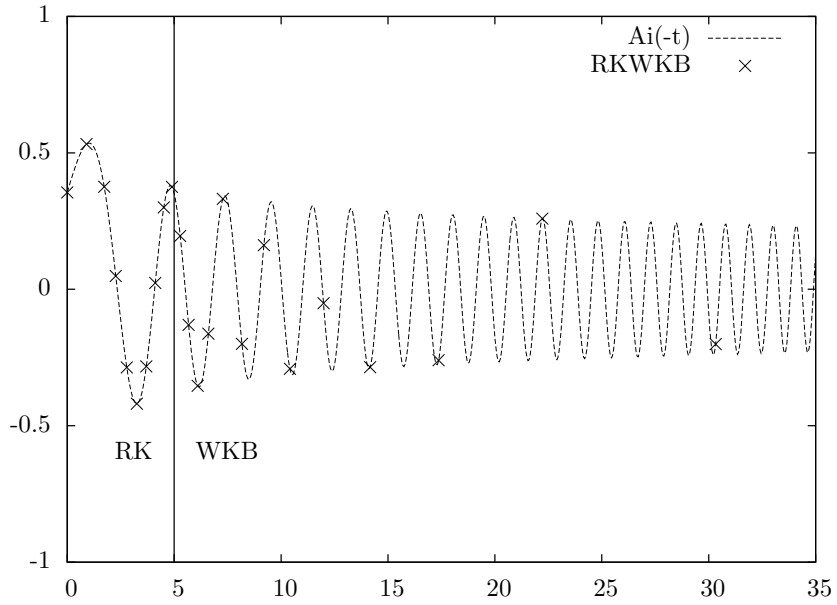


Figure 10.3: The RKWKB method compared with the analytical solution. The algorithm starts at $t = 0$ in the RK regime, since ω is varying quickly relative to the oscillation period. At $t = 5$ it becomes more efficient to use the WKB regime, and the points start to increase in separation. By $t = 15$ the algorithm is skipping multiple periods, and the step size h increases exponentially.

The RKWKB method and the modified Magnus method share some key features. Indeed, the lowest order modified Magnus method is equivalent to a 1^{st} order WKB approach (Iserles, 2002b). However, our approach is distinguished in several ways.

First, the Magnus expansion (10.38) requires multiple integrals for higher order terms. These integrals are tricky to implement, and can introduce considerable computational overhead. The WKB expansion (10.12) on the other hand requires at most single integrals, replacing the double integrals with additional derivative terms of ω, which are typically easier to compute.

Second, our approach uses adaptive step-size control, which is very easy to implement in the WKB framework and crucial for real-world numerical work.

Finally, by using dynamic switching, the algorithm is able to utilise the optimal approach in real time.

However, Iserles' approach has been the inspiration for this work, and we believe that many of the difficulties associated with the implementation of Magnus methods are merely engineering problems. We believe that in the fullness of time Magnus methods could become the de-facto numerical integration tool. In the mean-time, this work provides a simpler, more streamlined methodology.

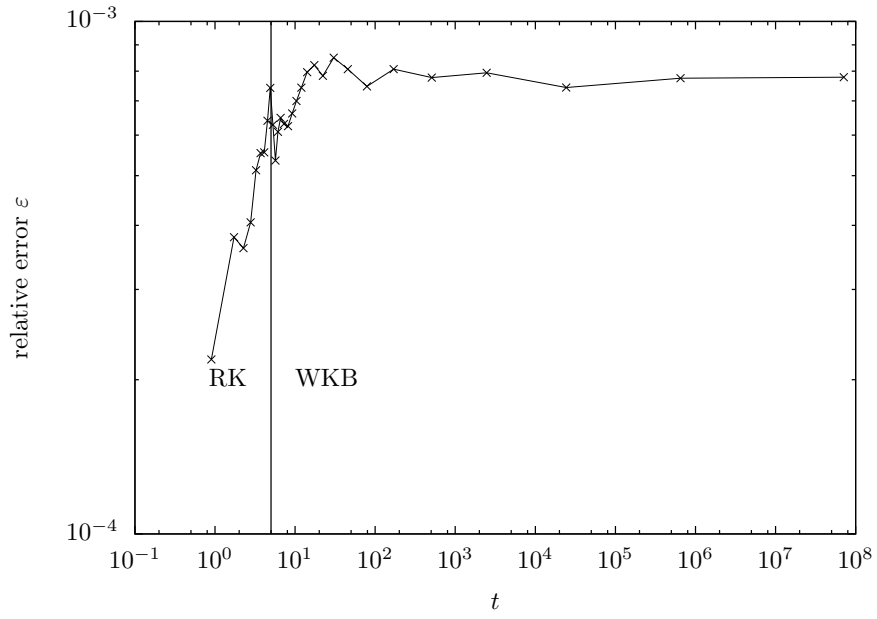


Figure 10.4: Fractional difference between the analytical solution and RKWKB solution from Figure 10.3. The algorithm's fractional error begins at $t = 0$ with an error of $\sim 10^{-4}$, but rises in the RK phase. This is to be expected as RK methods accumulate errors (particularly for oscillatory solutions). Upon entering the WKB region, the fractional error levels off. Note the rapidly increasing step size, and accuracy at extremely late times t. To our knowledge, no numerical scheme to date has demonstrated the ability to solve the Airy equation (10.33) to times as late as this.

10.7 Conclusions

We have presented a novel method for numerically solving linear differential equations with highly oscillatory solutions. We use a Wentzel-Kramers-Brillouin expansion to create an adaptively stepping algorithm in the same manner as a Runge-Kutta scheme. Further, the algorithm will switch back to a normal RK approach when the frequency of oscillation is varying too quickly for WKB to approximate accurately. The method is compared to Iserles existing approaches, and found to be a reasonable alternative without requiring the use of heavy Lie-group machinery. This chapter is not intended to be a complete exposition, but more a proof-of-principle to create a springboard for further investigation.

0					
c_2	a_{21}				
c_3	a_{31}	a_{32}			
$\vdots$	$\vdots$	$\vdots$	$\ddots$		
c_s	a_{s1}	a_{s2}	$\cdots$	$a_{s,s-1}$	
	b_1	b_2	$\cdots$	b_{s-1}	b_s

Table 10.1: Butcher tableau for a general explicit RK method.

0					
$\frac{1}{4}$	$\frac{1}{4}$				
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$			
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$		
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$	
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$
	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0
	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$

Table 10.2: Butcher tableau for the embedded Runge-Kutta-Fehlberg 4(5) method.

Appendix 10.A Runge-Kutta-Fehlberg

A general explicit RK method can be written as:

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i, \tag{10.41}$$

$$k_s = f(t_n + c_s h, y_n + h \sum_{i=1}^{s-1} a_{si} k_i), \tag{10.42}$$

where the coefficients $\{c_i, a_{si}\}$ are determined by the choice of method and are typically written in a Butcher tableau (Table 10.1).

A particularly efficient example is the Runge-Kutta-Fehlberg 4(5) method which uses an embedded approach. It performs a fourth order step and a fifth order step, and uses the difference between these as an estimate of the error. Impressively, both steps are calculated using the same values of $\{k_i\}$ (but different values of $\{b_i\}$), and hence the method only requires five function evaluations of f per step. Its Butcher tableau is detailed in Table 10.2.

Conclusion: Methods

In this second part of the thesis, an account has been given of the methods I have developed in order to further investigate the pre-inflationary universe. Chapter 9 detailed the construction of the next generation of nested sampling algorithms: PolyChord. In Chapter 10 I described a novel method for solving differential equations with oscillating solutions: RKWKB.

Future

Bayesian methods & nested sampling

The nature of nested sampling means that PolyChord is applicable to a much broader range of astrophysical problems, far beyond its use in cosmology for this thesis.

Alongside the content of this thesis, I have been working with other members of the department aiming to test Einstein's general theory of relativity in pulsar timing arrays, and with the AMI team to identify galaxies in clusters (Rumsey et al., 2015). Further afield, work has been planned with teams at Imperial college to analyse FERMI X-rays in the search for dark matter and the Colorado DARE team to analyse the cosmic dark ages using radio wave data. The fast adoption of PolyChord by the community is due to the fact that it is the first nested sampling algorithm capable of navigating complicated, high dimensional problems.

As well as this collaborative work, I am also excited about the scientific possibilities that PolyChord has opened up outside of the field of astrophysics. The kind of problems it attacks are ubiquitous in science, and tend to stand out as long-standing unsolved theoretical issues. Examples range from the training of neural networks in machine learning to computing protein folds in biochemistry.

I have already developed a preliminary PolyChord 2.0, capable of tackling even harder problems, and I anticipate the full exploration of the scientific applications forming a significant part of my future research.

In a more theoretical Bayesian setting, it is my belief that PolyChord is in fact merely the first in a large class of nested sampling algorithms, the true potential of which has only just begun to be tapped.

RKWKB

The RKWKB approach has opened up an equally rich seam of potential research. There are many extensions which require immediate exploration.

First, the multi-dimensional case is of interest. If this method could be extended to multiple coupled linear differential equations with oscillatory solutions, then the applications to cosmology are immediate. Currently, cosmological inference is dominated by the cost of computing transfer functions. The calculation of these involves the integration of coupled linear differential equations with time varying coefficients derived from the background cosmology. Implementing RKWKB in these contexts has the potential to lead to a new class of Boltzmann codes which remove the need for supercomputers to perform cosmological inference.

Second, the method could be improved if it no longer needed Runge-Kutta phases. If the same approach could be generalised so that the WKB stepping procedure was equally powerful regardless of the size of the oscillating term, then the method would be much cleaner and potentially more efficient.

Third, RKWKB has only thus far been applied to the linear case. Similar approaches by Iserles (2003) have been generalised to the non-linear case, and there is little reason to suggest that a similar approach would not also apply here. Along similar lines, it would be interesting to see if RKWKB can also be applied in a Lie group context as in Iserles et al. (2000).

Bibliography

K. N. Abazajian, G. Aslanyan, R. Easther, and L. C. Price. The Knotted Sky II: Does BICEP2 require a nontrivial primordial power spectrum? *ArXiv e-prints*, March 2014.

R.P. Agarwal, R.P. Agarwal, and V. Lakshmikantham. *Uniqueness and Nonuniqueness Criteria for Ordinary Differential Equations*. Series in real analysis. World Scientific, 1993. ISBN 9789810213572. URL http://books.google.co.uk/books?id=q4OkW4H8BCUC.

Stuart Aitken and OzgurE Akman. Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. *BMC Systems Biology*, 7(1):72, 2013. doi: 10.1186/1752-0509-7-72. URL http://dx.doi.org/10.1186/1752-0509-7-72.

A. Albrecht and P. J. Steinhardt. Cosmology for grand unified theories with radiatively induced symmetry breaking. *Phys. Rev. Lett.*, 48:1220–1223, April 1982. doi: 10.1103/PhysRevLett.48.1220.

Andreas Albrecht and Paul J. Steinhardt. Cosmology for grand unified theories with radiatively induced symmetry breaking. *Phys. Rev. Lett.*, 48:1220–1223, Apr 1982. doi: 10.1103/PhysRevLett.48.1220. URL http://link.aps.org/doi/10.1103/PhysRevLett.48.1220.

J.-M. Alimi, A. Blanchard, A. Bouquet, F. Martin de Volnay, and J. Tran Thanh Van, editors. *Particle astrophysics. The early universe and cosmic structures.*, 1990.

C. Armendáriz-Picón and E. A. Lim. Vacuum choices and the predictions of inflation. *J. Cosmology Astropart. Phys.*, 12:006, December 2003. doi: 10.1088/1475-7516/2003/12/006.

G. Aslanyan, L. C. Price, K. N. Abazajian, and R. Easther. The Knotted Sky I: Planck constraints on the primordial power spectrum. *ArXiv e-prints*, March 2014.

J. M. Bardeen. Gauge-invariant cosmological perturbations. *Phys. Rev. D*, 22: 1882–1905, October 1980. doi: 10.1103/PhysRevD.22.1882.

D. Baumann. TASI Lectures on Inflation. *ArXiv e-prints*, July 2009.

V. A. Belinsky, L. P. Grishchuk, I. M. Khalatnikov, and Y. B. Zeldovich. Inflationary stages in cosmological models with a scalar field. *Physics Letters B*, 155:232–236, May 1985. doi: 10.1016/0370-2693(85)90644-6.

C.M. Bender and S.A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. Springer, 1999. ISBN 9780387989310.

A. Berera and L.-Z. Fang. Thermally Induced Density Perturbations in the Inflation Era. *Physical Review Letters*, 74:1912–1915, March 1995. doi: 10.1103/PhysRevLett.74.1912.

A. Berera and A. F. Heavens. Detection limits for super-Hubble suppression of causal fluctuations. *Phys. Rev. D*, 62(12):123513, December 2000. doi: 10.1103/PhysRevD.62.123513.

A. Berera, L.-Z. Fang, and G. Hinshaw. Attempt to determine the largest scale of primordial density perturbations in the universe. *Phys. Rev. D*, 57: 2207–2212, February 1998. doi: 10.1103/PhysRevD.57.2207.

Arjun Berera. Warm inflation. *Phys. Rev. Lett.*, 75:3218–3221, Oct 1995. doi: 10.1103/PhysRevLett.75.3218. URL http://link.aps.org/doi/10.1103/PhysRevLett.75.3218.

M. Betancourt. Nested Sampling with Constrained Hamiltonian Monte Carlo. In A. Mohammad-Djafari, J.-F. Bercher, and P. Bessiére, editors, *American Institute of Physics Conference Series*, volume 1305 of *American Institute of Physics Conference Series*, pages 165–172, March 2011. doi: 10.1063/1.3573613.

N. D Birrell. *Quantum fields in curved space*. Cambridge University Press, Cambridge, [repr. with corrections] edition, 1984. ISBN 0521278589 (pbk.).

D. Boyanovsky, H. J. de Vega, and N. G. Sanchez. CMB quadrupole suppression. II. The early fast roll stage. *Phys. Rev. D*, 74(12):123007, December 2006a. doi: 10.1103/PhysRevD.74.123007.

D. Boyanovsky, H. J. de Vega, and N. G. Sanchez. CMB quadrupole suppression. I. Initial conditions of inflationary perturbations. *Phys. Rev. D*, 74(12): 123006, December 2006b. doi: 10.1103/PhysRevD.74.123006.

B. J. Brewer, L. B. Pártay, and G. Csányi. Diffusive Nested Sampling. *ArXiv e-prints*, December 2009.

C. R. Contaldi, M. Peloso, L. Kofman, and A. Linde. Suppressing the lower multipoles in the CMB anisotropies. *Journal of Cosmology and Astro-Particle Physics*, 7:002, July 2003. doi: 10.1088/1475-7516/2003/07/002.

U. H. Danielsson. Note on inflation and trans-Planckian physics. *Phys. Rev. D*, 66(2):023511, July 2002. doi: 10.1103/PhysRevD.66.023511.

A. de Oliveira-Costa, M. Tegmark, M. Zaldarriaga, and A. Hamilton. Significance of the largest scale CMB fluctuations in WMAP. *Phys. Rev. D*, 69(6): 063516, March 2004. doi: 10.1103/PhysRevD.69.063516.

C. Destri, H. J. de Vega, and N. G. Sanchez. Preinflationary and inflationary fast-roll eras and their signatures in the low CMB multipoles. *Phys. Rev. D*, 81(6):063520, March 2010. doi: 10.1103/PhysRevD.81.063520.

Scott Dodelson. *Modern cosmology*. Beijing World Pub. Corp, Beijing, 2008. ISBN 9787506291996 (pbk.).

R. Easther, B. R. Greene, W. H. Kinney, and G. Shiu. Generic estimate of trans-Planckian modifications to the primordial power spectrum in inflation. *Phys. Rev. D*, 66(2):023518, July 2002. doi: 10.1103/PhysRevD.66.023518.

Richard Easther and Hiranya V. Peiris. Bayesian Analysis of Inflation II: Model Selection and Constraints on Reheating. *Phys.Rev.*, D85:103533, 2012. doi: 10.1103/PhysRevD.85.103533.

F. Feroz and M. P. Hobson. Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *MNRAS*, 384:449–463, February 2008. doi: 10.1111/j.1365-2966.2007.12353.x.

F. Feroz and J. Skilling. Exploring multi-modal distributions with nested sampling. In U. von Toussaint, editor, *American Institute of Physics Conference Series*, volume 1553 of *American Institute of Physics Conference Series*, pages 106–113, August 2013. doi: 10.1063/1.4819989.

F. Feroz, M. P. Hobson, and M. Bridges. MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics. *MNRAS*, 398: 1601–1614, October 2009. doi: 10.1111/j.1365-2966.2009.14548.x.

F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt. Importance Nested Sampling and the MultiNest Algorithm. *ArXiv e-prints*, June 2013.

S.A. Fulling. Remarks on positive frequency and hamiltonians in expanding universes. *General Relativity and Gravitation*, 10(10):807–824, 1979. ISSN 0001-7701. doi: 10.1007/BF00756661. URL http://dx.doi.org/10.1007/BF00756661.

Stephen A Fulling. *Aspects of quantum field theory in curved space-time*. Cambridge University Press, Cambridge, 1989. ISBN 052134400X.

D. A. Green. A colour scheme for the display of astronomical intensity images. *Bulletin of the Astronomical Society of India*, 39:289–295, June 2011.

A. H. Guth. Inflationary universe: A possible solution to the horizon and flatness problems. *Phys. Rev. D*, 23:347–356, January 1981. doi: 10.1103/PhysRevD.23.347.

J. J. Halliwell. Scalar fields in cosmology with an exponential potential. *Physics Letters B*, 185:341–344, February 1987. doi: 10.1016/0370-2693(87)91011-2.

W. J. Handley, S. D. Brechet, A. N. Lasenby, and M. P. Hobson. Kinetic initial conditions for inflation. *Phys. Rev. D*, 89(6):063505, March 2014. doi: 10.1103/PhysRevD.89.063505.

W. J. Handley, M. P. Hobson, and A. N. Lasenby. POLYCHORD: nested sampling for cosmology. *MNRAS*, 450:L61–L65, June 2015a. doi: 10.1093/mnrasl/slv047.

W. J. Handley, M. P. Hobson, and A. N. Lasenby. POLYCHORD: next-generation nested sampling. *MNRAS*, 453:4384–4398, November 2015b. doi: 10.1093/mnras/stv1911.

W. J. Handley, A. N. Lasenby, and M. P. Hobson. Novel quantum initial conditions for inflation. *Phys. Rev. D*, 94(2):024041, July 2016a. doi: 10.1103/PhysRevD.94.024041.

W. J. Handley, A. N. Lasenby, and M. P. Hobson. The Runge-Kutta-Wentzel-Kramers-Brillouin Method. *ArXiv e-prints*, December 2016b.

D. K. Hazra, A. Shafieloo, and G. F. Smoot. Reconstruction of broad features in the primordial spectrum and inflaton potential from Planck. *Journal of Cosmology and Astro-Particle Physics*, 12:035, December 2013. doi: 10.1088/1475-7516/2013/12/035.

G. Hinshaw, D. Larson, E. Komatsu, D. N. Spergel, C. L. Bennett, J. Dunkley, M. R. Nolta, M. Halpern, R. S. Hill, N. Odegard, L. Page, K. M. Smith, J. L. Weiland, B. Gold, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, G. S. Tucker, E. Wollack, and E. L. Wright. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results. *ApJS*, 208:19, October 2013. doi: 10.1088/0067-0049/208/2/19.

R. Hlozek, J. Dunkley, G. Addison, J. W. Appel, J. R. Bond, C. Sofia Carvalho, S. Das, M. J. Devlin, R. Dünner, T. Essinger-Hileman, J. W. Fowler, P. Gallardo, A. Hajian, M. Halpern, M. Hasselfield, M. Hilton, A. D. Hincks, J. P. Hughes, K. D. Irwin, J. Klein, A. Kosowsky, T. A. Marriage, D. Marsden, F. Menanteau, K. Moodley, M. D. Niemack, M. R. Nolta, L. A. Page, L. Parker, B. Partridge, F. Rojas, N. Sehgal, B. Sherwin, J. Sievers, D. N. Spergel, S. T. Staggs, D. S. Swetz, E. R. Switzer, R. Thornton, and E. Wollack. The Atacama Cosmology Telescope: A Measurement of the Primordial Power Spectrum. *ApJ*, 749:90, April 2012. doi: 10.1088/0004-637X/749/1/90.

M.P. Hobson, G. Efstathiou, and A.N. Lasenby. *General relativity : an introduction for physicists*. Cambridge University Press, Cambridge, 2006. ISBN 9780521829519.

M.P Hobson, A.H. Jaffe, A.R Liddle, P.M. Mukherjee, and D Parkinson. *Bayesian Methods in Cosmology*. Cambridge University Press, 2009. ISBN 9780511802461. URL http://dx.doi.org/10.1017/CBO9780511802461. Cambridge Books Online.

Arieh Iserles. Think globally, act locally: Solving highly-oscillatory ordinary differential equations. *Appl. Numer. Math.*, 43(1-2):145–160, 2002a. ISSN 0168-9274. doi: 10.1016/S0168-9274(02)00122-8.

Arieh Iserles. On the global error of discretization methods for highly-oscillatory ordinary differential equations. *BIT Numerical Mathematics*, 42 (3):561–599, 2002b. ISSN 0006-3835. doi: 10.1023/A:1022049814688. URL http://dx.doi.org/10.1023/A%3A1022049814688.

Arieh Iserles. On the numerical analysis of rapid oscillation, 2003.

Arieh Iserles, Hans Z. Munthe-Kaas, Syvert P. Nrsett, and Antonella Zanna. Lie-group methods. *Acta Numerica*, 9:215–365, 1 2000. ISSN 1474-0508. doi: null. URL http://journals.cambridge.org/article_S0962492900002154.

C. R. Keeton. On statistical uncertainty in nested sampling. *MNRAS*, 414: 1418–1426, June 2011. doi: 10.1111/j.1365-2966.2011.18474.x.

A. Lasenby and C. Doran. Closed universes, de Sitter space, and inflation. *Phys. Rev. D*, 71(6):063502, March 2005. doi: 10.1103/PhysRevD.71.063502.

Anthony Lasenby. The cosmic microwave background and fundamental physics. *Space Science Reviews*, 148(1):329–346, 2009. ISSN 1572-9672. doi: 10.1007/s11214-009-9616-4. URL http://dx.doi.org/10.1007/s11214-009-9616-4.

L. Lello and D. Boyanovsky. Tensor to scalar ratio and large scale power suppression from pre-slow roll initial conditions. *ArXiv e-prints*, December 2013.

A. Lewis. Efficient sampling of fast and slow cosmological parameters. *Phys. Rev. D*, 87(10):103529, May 2013. doi: 10.1103/PhysRevD.87.103529.

A. Lewis and S. Bridle. Cosmological parameters from CMB and other data: A Monte Carlo approach. *Phys. Rev. D*, 66(10):103511, November 2002. doi: 10.1103/PhysRevD.66.103511.

Antony Lewis, Anthony Challinor, and Anthony Lasenby. Efficient computation of CMB anisotropies in closed FRW models. *Astrophys. J.*, 538:473–476, 2000.

Andrew R. Liddle and D. H. (David Hilary) Lyth. *Cosmological inflation and large-scale structure*. Cambridge University Press,, Cambridge, 2000. ISBN 9780521660228 (cased).

A. Linde. Toy model for open inflation. *Phys. Rev. D*, 59(2):023503, January 1999. doi: 10.1103/PhysRevD.59.023503.

A. Linde. Fast-Roll Inflation. *Journal of High Energy Physics*, 11:052, November 2001. doi: 10.1088/1126-6708/2001/11/052.

A. Linde. Inflationary Cosmology. In M. Lemoine, J. Martin, and P. Peter, editors, *Inflationary Cosmology*, volume 738 of *Lecture Notes in Physics, Berlin Springer Verlag*, page 1, 2008. doi: 10.1007/978-3-540-74353-8_1.

A. Linde, M. Sasaki, and T. Tanaka. CMB in open inflation. *Phys. Rev. D*, 59 (12):123522, June 1999. doi: 10.1103/PhysRevD.59.123522.

A. D. Linde. A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems. *Phys. Lett. B*, 108:389–393, February 1982. doi: 10.1016/0370-2693(82)91219-9.

A. D. Linde. Initial conditions for inflation. *Physics Letters B*, 162:281–286, November 1985. doi: 10.1016/0370-2693(85)90923-2.

F. Lucchin and S. Matarrese. Power-law inflation. *Phys. Rev. D*, 32:1316–1322, September 1985. doi: 10.1103/PhysRevD.32.1316.

David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002. ISBN 0521642981.

Michael J. Mortonson, Hiranya V. Peiris, and Richard Easther. Bayesian Analysis of Inflation: Parameter Estimation for Single Field Models. *Phys.Rev.*, D83:043505, 2011. doi: 10.1103/PhysRevD.83.043505.

V. F Mukhanov. *Introduction to quantum effects in gravity*. Cambridge University Press, Cambridge, 2007. ISBN 9780521868341. URL http://www.loc.gov/catdir/enhancements/fy0729/2007530526-t.html.

V. F. Mukhanov, H. A. Feldman, and R. H. Brandenberger. Theory of cosmological perturbations. *Physics Reports*, 215:203–333, June 1992. doi: 10.1016/0370-1573(92)90044-Z.

P. Mukherjee, D. Parkinson, and A. R. Liddle. A Nested Sampling Algorithm for Cosmological Model Selection. *ApJ*, 638:L51–L54, February 2006. doi: 10.1086/501068.

R. M. Neal. Slice Sampling. *ArXiv Physics e-prints*, September 2000.

Jorge Norena, Christian Wagner, Licia Verde, Hiranya V. Peiris, and Richard Easther. Bayesian Analysis of Inflation III: Slow Roll Reconstruction Using Model Selection. *Phys.Rev.*, D86:023505, 2012. doi: 10.1103/PhysRevD.86.023505.

Leonard Parker. *Quantum Field Theory in Curved Spacetime : Quantized Fields and Gravity*. Cambridge University Press, Cambridge, 2009. ISBN 9780511813924 (ebook). URL http://ezproxy.lib.cam.ac.uk:2048/login?url=http://dx.doi.org/10.1017/CBO9780511813924.

A. A. Penzias and R. W. Wilson. A Measurement of Excess Antenna Temperature at 4080 Mc/s. *ApJ*, 142:419–421, July 1965. doi: 10.1086/148307.

Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 results. I. Overview of products and scientific results. *ArXiv e-prints*, March 2013a.

Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 results. XXII. Constraints on inflation. *ArXiv e-prints*, March 2013b.

Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 results. XV. CMB power spectra and likelihood. *A&A*, 571:A15, November 2014. doi: 10.1051/0004-6361/201321573.

Planck Collaboration, R. Adam, P. A. R. Ade, N. Aghanim, Y. Akrami, M. I. R. Alves, F. Argüeso, M. Arnaud, F. Arroja, M. Ashdown, and et al. Planck 2015 results. I. Overview of products and scientific results. *A&A*, 594:A1, September 2016a. doi: 10.1051/0004-6361/201527101.

Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, F. Arroja, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, and et al. Planck 2015 results. XX. Constraints on inflation. *A&A*, 594:A20, September 2016b. doi: 10.1051/0004-6361/201525898.

B. A. Powell and W. H. Kinney. Pre-inflationary vacuum in the cosmic microwave background. *Phys. Rev. D*, 76(6):063512, September 2007. doi: 10.1103/PhysRevD.76.063512.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 2007. ISBN 9780521880688.

E. Ramirez. Low power on large scales in just-enough inflation models. *Phys. Rev. D*, 85(10):103517, May 2012. doi: 10.1103/PhysRevD.85.103517.

E. Ramirez and D. J. Schwarz. φ^4 inflation is not excluded. *Phys. Rev. D*, 80 (2):023525, July 2009. doi: 10.1103/PhysRevD.80.023525.

E. Ramirez and D. J. Schwarz. Predictions of just-enough inflation. *Phys. Rev. D*, 85(10):103516, May 2012. doi: 10.1103/PhysRevD.85.103516.

G. N. Remmen and S. M. Carroll. How many e-folds should we expect from high-scale inflation? *Phys. Rev. D*, 90(6):063517, September 2014. doi: 10.1103/PhysRevD.90.063517.

K.F. Riley, M.P. Hobson, and S.J. Bence. *Mathematical Methods for Physics and Engineering: A Comprehensive Guide*. Cambridge University Press, 3rd ed edition, 2006. ISBN 9781139450997.

C. Rumsey, M. Olamaie, Y. C. Perrot, H. R. Russell, F. Feroz, K. J. B. Grainge, W. J. Handley, M. P. Hobson, R. D. E Saunders, and Schammel M. P. AMI observations of ten CLASH galaxy clusters: SZ and X-ray data used together to determine cluster dynamical states. *Submitted to MNRAS*, September 2015.

J. R. Shaw, M. Bridges, and M. P. Hobson. Efficient Bayesian inference for multimodal problems in cosmology. *MNRAS*, 378:1365–1370, July 2007. doi: 10.1111/j.1365-2966.2007.11871.x.

Deviderjit Singh Sivia and John Skilling. *Data analysis : a Bayesian tutorial*. Oxford science publications. Oxford University Press, Oxford, New York, 2006. ISBN 0-19-856831-2. URL http://opac.inria.fr/record=b1133948.

John Skilling. Nested sampling for general bayesian computation. *Bayesian Analysis*, 1(4):833–859, 12 2006. doi: 10.1214/06-BA127. URL http://dx.doi.org/10.1214/06-BA127.

A. A. Starobinsky. Spectrum of relict gravitational radiation and the early state of the universe. *Soviet Journal of Experimental and Theoretical Physics Letters*, 30:682, December 1979.

J. A. Vázquez, M. Bridges, M. P. Hobson, and A. N. Lasenby. Model selection applied to reconstruction of the Primordial Power Spectrum. *J. Cosmology Astropart. Phys.*, 6:006, June 2012a. doi: 10.1088/1475-7516/2012/06/006.

J. A. Vázquez, M. Bridges, M. P. Hobson, and A. N. Lasenby. Model selection applied to reconstruction of the Primordial Power Spectrum. *Journal of Cosmology and Astro-Particle Physics*, 6:006, June 2012b. doi: 10.1088/1475-7516/2012/06/006.

J. A. Vazquez, M. P. Hobson, A. N. Lasenby, M. Ibison, and M. Bridges. Reciprocity invariance of the Friedmann equation, Missing Matter and double Dark Energy. *arXiv:astro-ph/1208.2542*, August 2012.

M. Visser and C. Barceló. Energy Conditions and Their Cosmological Implications. In U. Cotti, R. Jeannerot, G. Senjanovi, and A. Smirnov, editors, *COSMO-99, International Workshop on Particle Physics and the Early Universe*, page 98, 2000.

Robert M Wald. *General relativity*. University of Chicago Press, Chicago ; London, 1984. ISBN 9780226870328. URL http://www.loc.gov/catdir/description/uchi051/83017969.html.

D. Yamauchi, A. Linde, A. Naruko, M. Sasaki, and T. Tanaka. Open inflation in the landscape. *Phys. Rev. D*, 84(4):043513, August 2011. doi: 10.1103/PhysRevD.84.043513.

J. Yokoyama and K.-I. Maeda. On the dynamics of the power law inflation due to an exponential potential. *Physics Letters B*, 207:31–35, June 1988. doi: 10.1016/0370-2693(88)90880-5.